CROSS-CULTURAL VALIDITY OF THE TEST
OF NON-VERBAL INTELLIGENCE

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

Rene S. Parmar, B.A., M.Ed.
Denton, Texas
August, 1988
Parmar, Rene S., Cross-Cultural Validity of the Test of Non-Verbal Intelligence. Doctor of Philosophy (Special Education), August, 1988, 139 pp., 10 tables, references, 173 titles.

The purpose of this study was to investigate the extent to which a non-verbal test of intelligence, the Test of Non-Verbal Intelligence (TONI), may be used for assessing intellectual abilities of children in India. This investigation is considered important since current instruments used in India were developed several years ago and do not adequately reflect present standards of performance. Further, current instruments do not demonstrate adequate validity, as procedures for development and cultural transport were frequently not in adherence to recommended guidelines for such practice.

Data were collected from 91 normally achieving and 18 mentally retarded Indian children, currently enrolled in elementary schools. Data from an American comparison group were procured from the authors of the TONI. Subjects were matched on age, grade, and area of residence. Subjects were also from comparative socio-economic backgrounds.

Literature review of the theoretical framework supporting cross-cultural measurement of intellectual
ability, a summary of major instruments developed for cross-cultural use, non-verbal measures of intellectual ability in India, and issues in cross-cultural research are discussed, with recommended methodology for test transport.

Major findings are: (a) the factor scales derived from the Indian and American normally achieving groups indicate significant differences; (b) items 1, 3, 5, 8, 10, and 22 are biased against the Indian group, though overall item characteristic curves are not significantly different; (c) mean raw scores on the TONI are significantly different between second and third grade Indian subjects; and (d) mean TONI Quotients are significantly different between normally achieving and mentally retarded Indian subjects.

It is evident that deletion of biased items and rescaling would be necessary for the TONI to be valid in the Indian context. However, because it does discriminate between subjects at different levels of ability, adaptation for use in India is justified. It may prove to be a more current and parsimonious method of assessing intellectual abilities in Indian children than instruments presently in use.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>v</td>
</tr>
</tbody>
</table>

## Chapter

### I. INTRODUCTION ........................................ 1

- Purpose of the Study
- Basis for Test Selection
- Significance of the Study
- Review of the Literature
- Considerations in Cross-Cultural Research

### II. METHOD .................................................. 64

- Research Design
- Sample Selection
- Hypothesis Formulation
- Instrument Selection
- Data Collection Procedures
- Statistics for Equivalence
- Statistical Methodology

### III. RESULTS ............................................... 78

### IV. DISCUSSION ............................................ 93

### APPENDICES ................................................ 109

### REFERENCES ................................................ 118
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Mean Age and Age Ranges of Subjects</td>
<td>65</td>
</tr>
<tr>
<td>2.</td>
<td>Percentage of Subjects at Hollingshead Index Socio-Economic Levels</td>
<td>66</td>
</tr>
<tr>
<td>3.</td>
<td>Rotated Factor Pattern Matrix (Varimax Rotation)</td>
<td>79</td>
</tr>
<tr>
<td>4.</td>
<td>Factor Pattern Matrix of Loadings from the American Normative Sample (Principal Components)</td>
<td>82</td>
</tr>
<tr>
<td>5.</td>
<td>Factor Scales Derived Through Principal Components Analysis</td>
<td>83</td>
</tr>
<tr>
<td>6.</td>
<td>Factor Scale Means and Critical Values of F</td>
<td>83</td>
</tr>
<tr>
<td>7.</td>
<td>Point Biserial Correlations of Indian Normally-Achieving and American Normative Sample Subjects</td>
<td>86</td>
</tr>
<tr>
<td>8.</td>
<td>Rasch Scale Values for Items</td>
<td>88</td>
</tr>
<tr>
<td>9.</td>
<td>Items Evidencing Significant Differences Between Indian and American Subjects</td>
<td>98</td>
</tr>
<tr>
<td>10.</td>
<td>Validity Studies on Non-Verbal Tests in India - Methodology</td>
<td>107</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure

1. Item Discrimination Curve .......................115
2. Item Difficulty Curve .........................116
3. Regression of TONI scores on the WISC-R ......117
CHAPTER I

INTRODUCTION

In recent years there has been a growing recognition in India of the need to develop psycho-educational testing measures which demonstrate validity in the Indian culture. Though research exists that is conventional and sound, most of it is directed toward American and Western European needs, rather than problems and trends in India (Sinha, 1973). Thus, there is insufficient data available that could guide policy-making for educational practice in India.

An emerging educational interest in India, special education, has emphasized the need for reliable and efficient instruments to assess cognitive capacity. Psycho-educational assessment devices presently in use were developed and standardized several decades ago, and it is questionable whether the norms established are a relevant basis for comparison today (Sinha, 1981). Further, translation of tests initially developed in the United States (US) or Europe has not always been done with sufficient care and in adherence to guidelines for such procedure; thus some instruments do not demonstrate requisite validity.
Major tests currently used in India for intellectual assessment include (a) Raven’s Progressive Matrices (RPM) (Raven, 1960), (b) Wechsler Intelligence Scale for Children-Revised (WISC-R) (Wechsler, 1974), (c) Non-Language Test of Verbal Intelligence (Chatterji & Mukherjee, 1967), (d) Bhatia’s Battery (Bhatia, 1955), (e) Kerala Non-Verbal Test of Intelligence (Nair, 1968), (f) Cattell Culture-Free Intelligence Test (CCFT) (Cattell & Cattell, 1959), and (g) Goodenough Draw-A-Man Test (Goodenough, 1926). These are adaptations primarily from western psychological thought, and frequently bear little relevance to conditions in India. Further, they do not address the vast differences in socio-economic status, literacy, and language that exist among the various subcultures in India. Most need extensive adaptations, to the extent that the adapted test does not completely reflect the theoretical basis on which it was formed. For example, the Information subtest of the WISC-R is simply deleted when testing Indian subjects, and this scale is not considered when computing IQ scores.

Problems in developing indigenous, functional, standardized instruments have included:

1. The Indian government does not support test development research in special education. Special education services for the disabled are not mandated by
the government. Programs for the mentally handicapped and learning disabled are primarily initiated and supported by private charities, with some government aid through state departments for social welfare. Lack of a government mandate has resulted in (a) low priority given to research from the educational field in the area of intellectual assessment, (b) limited market for standardized measures of intellectual assessment, and (c) absence of widely implemented testing programs to assess cognitive aptitude in school children.

2. Most research in mental measurement has come from the field of psychology and has focused primarily on clinical issues; thus the research has not adequately addressed educational concerns and implications. A close examination of the established validity of instruments used in evaluating educational practices and performance of school children appears absent in the literature.

3. Fourteen major languages and up to six hundred dialects are spoken across India, making any kind of standardization of a test extremely difficult, especially one with verbal components.

4. Numerous cultural differences also exist among the various language groups. All cultural groups would have to be representatively sampled for a test to have nation-wide applicability or group norms for comparison purposes would have to be developed.
5. Inadequate research facilities have hindered the implementation of nation-wide research programs. Poor communication and transportation systems, and lack of adequate facilities in rural areas impede attempts to get a representative sample of subjects. The advent of computers is fairly recent, and sophisticated statistical analyses cannot be conducted easily on a large scale. In addition, dissemination of information is extremely difficult due to inadequate communication networks among professionals in the fields of education and psychology.

Purpose of the Study

The purpose of this study was to investigate the extent to which a non-verbal test of intelligence, the Test of Non-Verbal Intelligence (TONI) (Brown, Sherbenou, & Johnsen, 1982) may be used for assessing intellectual abilities of children in India. This investigation is considered important since current instruments used in India were developed several years ago and do not adequately reflect current standards of performance. Further, current instruments do not demonstrate adequate validity (as procedures for development and cultural transport were frequently not in adherence to recommended guidelines for such practice).
Basis For Test Selection

Problems With Cross Cultural Use of Verbal Tests

1. The hierarchical model of the structure of intelligence, first described by Vernon (1950), subdivides abilities into two main groups: verbal and spatial. Verbal-numerical components are referred to as the v:ed (verbal:educational) factor. The prediction of verbally-based learning has been linked with the v:ed factor in this conception. Research in recent years (Irvine & Berry, 1981) has explored the relationship between linguistic performance and other constructs, for example cognitive style, which refers to an individual's consistent mode of processing information from the environment (Newell, Shaw, & Simon, 1960; Sternberg, 1985). This line of research has opened avenues for use of non-language tests in comparative study of intelligence test usage across cultures.

2. The identification of intelligence tests for cross-cultural use which incorporate verbal items is a difficult task. Verbal items typically measure concepts which are central to a particular culture. Further, concepts which appear to be the same across all cultures under study may take on different meaning in the target culture because of problems of maintaining linguistic equivalence.

Factors which affect translation have been identified
by Brislin, Lonner, and Thorndike (1973). They include (a) the difficulty level of source language materials (Treisman, 1965), (b) specific wording difficulties (Scheuch, 1968), and (c) insufficient context for translating difficult phrases (Campbell, 1968). To overcome these problems and maintain syntactic and semantic equivalence of items (McCaulley & Colberg, 1985), back translation is recommended (Brislin, Lonner, & Thorndike, 1973; Triandis & Brislin, 1984). Back translation is a multi-step process where the original test material is translated into the target language by bilinguals, then translated back into the original language by another group of bilinguals. The original material is then modified. The process is repeated several times, with the expectation that greater convergence is achieved at each comparison.

No verbal intelligence test adapted for use in India can claim linguistic equivalence through stringent back translation. In some instances even conceptual equivalence is questionable, as in the case of the WISC-R, where non-equivalent items (including the entire Information section) are deleted from the test when it is administered to Indian subjects. In addition, verbal tests available are only applicable to one of the fourteen major language groups (generally the Hindi-speaking). In
light of the inadequate translation of available tests, and their limited potential for widespread use, it was considered preferable to examine the validity of a non-verbal measure of intelligence.

3. In examining theories of intelligence, to be discussed later in this chapter, several viewpoints are presented which support the existence of a general factor "intelligence", comprised of behaviors independent of any particular cultural context. Many non-verbal tests in existence purport to measure this general factor. Through componential analysis, problem-solving tasks of analogies, classification, series completion, and syllogisms have been identified that measure complex intellectual behavior (Sternberg, 1977). Such tasks have been utilized in development of test items for Raven's Progressive Matrices, Cattell Culture—Fair Intelligence Test, and the Test of Non-Verbal Intelligence. However, as Irvine and Berry (1983) point out, identical test items do not guarantee equal precepts or processes in humans. Further investigation of tests which claim cross-cultural equivalence must be considered.

**Rationale For Use of The TONI**

The TONI was selected for use in the present study for several reasons. First, it meets criteria specified by Jensen (1980) for selection of culturally-reduced tests, that is, (a) it is a performance measure, (b)
instructions are presented in pantomime to subjects, (c) it provides preliminary practice items, (d) it it untimed, (e) items are comprised of abstract figural content, (f) items require abstract reasoning rather than factual information, and (g) problems are designed in such a way that subjects may not guess at answers from memory of similar items encountered in the past. Due to the nature of the instrument, it can be used in both cultures in its original standardized form which means that the results are accessible to a variety of analysis on equivalence (Malpass & Poortinga, 1986).

Second, the TONI is a relatively recently developed test. Tests commonly used for cross-cultural study have been developed and standardized several decades ago. Norms developed at that time would not be considered representative today. Further, the standardization sample itself, in many cases, has not been representative even of the culture in which the test was originally developed. On the other hand, the standardization sample of the TONI appears to have been carefully chosen to be representative of the relevant population (Clark, 1985).

Third, reviews of the TONI have supported its usefulness as a measure to aid in screening and diagnosis of the educable mentally retarded, deaf, and learning disabled (Clark, 1985). The items are considered well
scaled in difficulty and psychological complexity (Mayo, 1985). The test manual is considered to be, in many ways, "a model for its genre" (Clark, 1985, p. 1580). The TONI appears to deliver reasonably reliable, replicable norm referenced, nonverbal IQ scores which might be useful as part of a comprehensive evaluation (Shelly, 1985).

The authors of the TONI developed the test to fill the need for an instrument which measures intellectual ability of persons for whom traditional tests, which use written or spoken language as part of their content or testing format, are inappropriate. These include "people who are unable to read or write and people who have poor or impaired language skills, such as aphasics, nonEnglish speakers, and individuals who are mentally retarded, learning disabled, deaf, or culturally different" (Brown et al., 1985, p. 2).

Though not considering the TONI 'culture free,' the authors feel that it is significantly 'culturally reduced' as it was developed in adherence to guidelines provided by Jensen (1980) and Duffey, Salvia, Tucker, and Ysseldyke (1981) for culturally reduced tests. Thus the test would address the need for an intelligence test which minimizes bias against subjects from racial, ethnic, SES, and linguistic minorities.

Significance of the Study

As discussed earlier, numerous problems arise when
attempting to develop measurement instruments for use in India. This study is significant in that it addresses those problems and attempts to provide a basis for sound theoretical research that may guide educational planning in India.

**Need For Research In The Field Of Intelligence Testing To Guide Educational Policy**

With the emerging recognition by the Indian government of the need to provide programs for the mentally handicapped and learning disabled, it is imperative that reliable and valid instruments be developed which could be used for purposes of identification and placement of special needs children. Though Sinha (1973) identified as a research priority the need to develop a data bank of factual information which may be used for immediate policy decision in the field of education, to date, widespread attempts to collect relevant data on educational assessment practices are conspicuously absent in the major research journals in education and psychology published in India. To address the research deficit, the present study will provide a basis for current and relevant research utilizing materials and methods based on sound theory to provide information that may be translated into practice in the field of education.
Need To Establish Educational Relevance Of Testing Instruments

Historically, the study of intellectual ability has been integral to the field of special education. A large number of classification and placement practices depend on measures of intelligence (cognitive aptitude) and it thus becomes essential for special educators to closely examine issues surrounding the validity of currently used measuring instruments. Knowledge of intelligence and intelligence testing comes predominantly from the field of psychology. Education, particularly special education, has relied heavily on psychology to guide both theory and method in assessment procedures for classification and placement of children. There is, however, an emerging realization that psychological theories must be tested for educational meaningfulness before they can be widely accepted into practice. By considering the validity of the TONI in light of current classroom performance of the children under study, this study will provide comparative data that relates directly to educational practice in India in addition to comparisons with other psychological measures.

Need For Assessment Instruments Applicable Across Language Groups

A significant barrier to any widespread research in India is the existence of a multitude of languages and
dialects in the various regions that comprise the country. The selection of a non-verbal test for the present research was intended to address this problem, and investigate an alternative to verbal testing of intellectual ability.

**Need For Culturally-Reduced Assessment Instruments**

With cultural differences evident among the various language groups, a nationally-applicable test must be "equally germane" to all cultures for which it is intended. The present study utilizes a test which has potential for being sufficiently culturally-reduced so that validity may be established across various cultural groups.

**Need For Carefully Conducted Research**

Though considerable research on cross-cultural testing exists, most of it is inconclusive, or of limited generalizability because numerous extraneous factors have not been addressed and controlled. It is not clear whether the differential performance of the culturally different group is due to bias inherent in the test, or outside factors (i.e., language differences, socio-economic status of subjects, educational level of subjects, and speed of performance required) (McCauley & Colberg, 1983). To eliminate possible rival hypotheses, the present study controls for the following sample characteristics, based
on guidelines established by Lonner and Berry (1986): (a) age, (b) literacy level, (c) area of residence (urban), and (d) socio-economic status. In addition, the utilization of examiners from the subjects' own culture will control for cultural differences which may bias examiner interpretation of results. The study will thus provide information less ambiguous, and less open to contradiction by rival hypotheses.

Review of the Literature

Standardized tests have been widely used to study intelligence since first developed by Binet in 1904. Efforts to use these tests in comparative studies across cultures, however, have resulted in confusing and often contradictory results and frustration. One explanation for this lack of acceptance of "culture-fair" measures may be that, in studies of intelligence testing and consequent conclusions about cognitive ability, it has rarely been acknowledged that the instruments used in such studies were thoroughly inadequate (Zaidi, 1979). Further, limited attempts at cross-cultural study have not always been guided by sound theoretical orientation taken from main psychological literature.

Theoretical Framework

Diverse opinions have been expressed regarding the possible existence of a general factor, "intelligence," which is present in all cultures and is unique to the
human species. The assumption that intelligent behavior is comprised of the same variables in all cultural contexts has been questioned. Literature on attempts to universally define intelligence and make cross-cultural comparisons to validate the definitions, however, does not provide definitive answers for any school of thought. Seven major conceptualizations, or "themes" (Fry, 1984), regarding concepts, meaning, and definition of intelligence have been identified in the literature.

The Contextualist View

Sternberg (1982a, 1982b, 1984) emphasized the embeddedness of intelligence in its own social-cultural and ecological context. Five major aspects of his definition are:

1. Intelligence is defined in terms of behaviors in real world environments. Thus, experiments or assessments conducted in laboratory or artificial settings have little or no relevance.

2. Intelligence is defined in terms of behaviors in environments that are relevant to one's life. This implies that intelligence has no meaning outside the socio-cultural context and may, in fact, differ for a given individual from one culture to the next.

3. Intelligence involves adaptation to the environment. What is adaptive in one environment may be
maladaptive in another, and hence, actions that are intelligent in one culture may be unintelligent in another.

4. Intelligence involves selection and shaping of environments as well as adapting to them. Where possible, more intelligent people actively seek environments that are more favorable for their adaptive skills, that is, environments in which they perform more intelligently.

5. Intelligence is directed towards a goal (or multiple of goals), that is, it is purposive.

To determine what comprises intelligence in a particular cultural context, Sternberg (1984) recommends asking people from that culture. Based on previous research (Bruner, Shapiro, & Taguiri, 1958; Neisser, 1978; Sternberg, 1982b; Yussen & Kane, 1984), three major behavioral characteristics of intelligent individuals have been identified, (a) practical problem-solving ability, (b) verbal ability, and (c) social competence.

Though the contextualist view argues for the importance of studying performance in the real world, it does not argue in favor of, or in opposition to, any paradigm for studying intelligence. Nor does it specifically exclude some paradigms while preferring others. However, intelligence is considered too complex to be measured by a single assessment device that does not consider interaction of the individual and his or her
culture. Further, no measure of intelligence can exist which is universally applicable across all cultures.

**Piagetian View**

Piagetian stage theory and metatheory has since 1960 had a marked impact on the meaning and assessment of intelligence and cognitive development, particularly in Western psychology. The basic tenet is the individual's ability to adapt to the environment (Fry, 1984).

The procedures devised by Piaget and his co-workers illustrate the progression of individuals through certain organismic stages (changes) that are irreversible, therefore absolute. Most cross-cultural Piagetian research has been devoted to the study of the development of particular concepts in the concrete operational stage, with very few extensions to the formal operational stage. These include the processes of conservation, disassociation, spatial representation, imagining, classification, inclusion, intersection of classes, and elementary and formal logic. There are indications that performance is influenced to a large extent by the task format, namely whether the tasks involve mainly verbal questioning about a display, the manipulation of materials, or drawing (Dasen, 1984).

Construct validation has seldom been attempted with non-Piagetian tasks; when it was, results were variable.
and seldom conclusive. Attempts have been made (cf. Heron & Dowell, 1974) to verify developmental stages as transcultural universals. In general, the stages are robust, except the concrete-operational stage which does not demonstrate sufficient universality in the literature (Pick, 1980). Further, when conservers are contrasted with non-conservers, the classification obtained by Piagetian procedures has not been verified by other mental tests used across cultures (Heron, 1971; Heron & Dowell, 1973, 1974).

**Fluid and Crystallized Intelligence**

Intelligence has been conceptualized in terms of two major interrelated components, fluid and crystallized abilities (Cattell, 1963, 1965; Horn, 1978; Horn & Cattell, 1966, Horn, Donaldson, & Engstrom, 1981). The crystallized factor "loads more highly those cognitive performances in which skilled judgment habits have become crystallized as the result of earlier learning application of some prior, more fundamental general ability" (Cattell, 1963, p. 2-3). The fluid factor is conceptualized as the "more fundamental," biologically rooted adaptational ability. Fluid ability manifests itself in performance on "tests requiring adaptation to many situations, where crystallized skills are of no particular advantage" (Cattell, 1963, p. 3). Thus, fluid ability is considered to be measured by tasks such as those presented in Raven's
Progressive Matrices or on other nonverbal tests involving semantic, figural, or symbolic content such as various versions of the Block Design found on the Weschler scales (Horn, 1979).

The belief in the existence of fluid and crystallized intellectual abilities in human beings as separate classes of abilities supports the measure of fluid intelligence, which should exist in a common form across cultures. Extensive education or acculturation are considered to be of little or no advantage in the development of fluid abilities (Hayslip & Kennelly, 1982). These abilities appear to be utilized when solving figurally presented problems, requiring understanding of abstract and often novel relations, as is required in inductive reasoning tasks (e.g., analogies and series completion) (Cattell, 1971). Further, they appear to be influenced by heredity.

**Hemispheric Lateralization of the Brain**

Vernon (1984) has discussed aspects of the differential contexts and capacities of the two hemispheres of the brain. The sequential and analytic processing of the left brain and the holistic or synthetic processing in the right brain and their interaction are considered.

This paradigm places reduced emphasis on the concept of intelligence, and increases the importance of strategy
utilization as a basic process in the child's development and learning. Most proponents agree that strategies may be modified or developed by appropriate methods of training, whereas intelligence, as traditionally conceived, was essentially untrainable. Lateralization exists in similar form throughout the human species and studies of persons who are unilaterally damaged because of strokes, brain injury, or disease, show similar losses in functioning regardless of the culture of residence (Vernon, 1984).

When assessing cultural differences and similarities in use of strategies, tests using verbal mediums may indicate differences in performance because of differences in the orthographic and linguistic characteristics of the languages used. Non-verbal tests, however, should indicate no such differences. Sex and genetic differences exist in establishing dominance, but no evidence indicates significant cultural differences.

Judgments of Individuals

An "alternative view" of intelligence, proposed by Goodnow (1984) focuses not on what intelligence "is" but rather on the ways in which the judgment is made that someone is more or less intelligent. Thus, intelligence is not considered to be a quality people possess in varying degrees, but rather as a collection of behaviors considered by persons in any given society to be
characteristic of "intelligent people" (Fry, 1984).

This approach challenges traditional modes of assessment which do not take into consideration that intelligence is not a steady state, that is, test scores at one age are not perfect predictors of scores at another age. Further, large differences in test-performance are found between schooled and non-schooled subjects, suggesting that abilities to do well on tests appear related to prior exposure to academic tasks. A discrepancy exists, however, in that work on verbal associations suggests that definitions of intelligence do differ across cultures, whereas analyses of examples of behavior indicate a great deal of agreement on what the term means, at least in "modern" societies (Chen, Braithwaite, & Huang, 1982; Gill & Keats, 1980). Thus universals of "intelligent behavior" may be identified.

Although many behaviors may be agreed upon as "intelligent" some may still be thought "better" than others, may be more valued than other, even though all are equally "correct." For example, in Western thought rationality (a cool, unemotional, logical, controlled style) is valued, while African societies consider behaviors related to social competence to be of great value.
Valser (1984) deals with the concept of intelligence from the point of view of social attributions of causality, made in the process of communication about the phenomena, subsumed under the label "intelligence." Influences on the attribution of causality are (a) the perspective of the explainer of an event determines which aspects of the event are used to project explanations into them, and (b) interpretation of the concept of "causality" by the explaining individual of the given culture determines the search in the semiotic system for suitable explanatory concepts. Measurement of intelligence is through qualitative study of individuals from the culture under study, rather than widespread, standardized assessment.

A number of difficulties in the building of a general theory of intelligence based upon the meaning of intelligence in the common sense of a given culture or cultures are posited. The psychologist's task is to transcend the limitations of the culture-specific common sense meaning and understanding of intelligence (Valser, 1984).

For instance, the separation of the rational from the social, which is so characteristic of "intelligence" in Western cultures need not be shared by people from other cultural backgrounds. Social-interactive characteristics
may be valued highly in other cultures and considered components of intelligent behavior.

Further, cultures cannot be conceptualized as immutable, static entities. People with different cultural backgrounds differ in the sets of characteristics associated with "intelligence" and "intelligent person." The effect of formal schooling as well as changes in economic aspects of the culture can wipe out traditional dissimilarities in the meaning of "intelligence" cross-culturally. However, the homogenization of the meaning of the term across cultures does not prove the universality of our conceptions of intelligence in psychology.

Cultural Relativism

Cultural relativism proposes a framework (Berry, 1984) for exploring the universal features of cognitive competence and functioning. Within this paradigm, generalizations exist when defining intelligence, but may not be made exclusive of local phenomena.

Two bases are proposed for expecting cognitive universals in the human species. The conceptual basis is one, which is provided by the existence of numerous biological, linguistic, and cultural universals which are pan-human in scope (Lonner, 1980) (e.g., we are members of the same species, we all use language, we all form social relationships and institutions, we all process
information, use it, store it, and work it over). The empirical basis exists in that empirical evidence suggests the existence of some common cognitive factors, and inferences from these to common underlying cognitive functions is a conventional step to take.

Cultural relativism was developed in reaction to existing views of general intelligence and specific ability. The concept of general intelligence is the classical view where two assumptions are made: (a) that the cultural life of the test developer and the test taker differ only in one important respect, that of language, and (b) the cognitive abilities characteristic of the cultural life of the test developer and test taker differ only in one respect, that of level of development (Berry, 1984).

On the opposite end of the continuum is the specific abilities approach where "cultural differences" in cognition reside more in situations to which particular cognitive processes are applied than in the existence of a process in one cultural group and its absence in another (Cole, Gay, Glick, & Sharp, 1971, p. 233). Thus there is no universal pattern or structure in ability data and cross-cultural comparisons may not be made with validity.

Cultural relativism searches for systematic relationships among abilities and engages in cross-cultural comparison. This cross-cultural work (Berry,
1976, 1984; Witkin & Berry, 1975) is characterized by an analysis of the local cultural context, by attempts to assess the cognitive performance of individuals in a number of groups, and a search for systematic relationships among performances and between performances and the cultural context.

Ferguson (1954, 1956) argued that differing patterns of ability are developed by persons from different cultural environments, and these patterns become stabilized through over-learning and transfer. Thus, systematic relationships exist in the cognitive styles of persons from a particular culture which may be identified and assessed. A review of cross-cultural studies by Berry (1984) indicates a high degree of cross-cultural agreement on competence domains (mental abilities highest, knowledge second), and within the domain of mental abilities there was also agreement that "understanding" and "thinking" were most important. Attempts to measure intelligence based on assumption of its universality across cultures thus find support in approaches of cultural relativism. Assessment procedures based upon tasks requiring understanding and thinking, the basis for problem solving, appear to have potential for cross-cultural study.

Conclusion

Of the paradigms proposed in the literature, support
for the present study is found in six. First, research based on Piagetian theory indicates the transcultural existence of certain processes of learning and development. The search for further common processes at stages later than the concrete-operational merits study. Second, fluid intelligence is considered to be a universally present ability, which may be measured and compared across cultures. It is independent of knowledge acquired by the individual specific to his/her culture. Third, research on brain lateralization posits transcultural equivalence in right and left hemisphere brain functions. Support is found for the use of non-verbal mediums to assess strategy applications and problem-solving. Fourth, research on conceptualizing intelligence as based on judgments of individuals has demonstrated that in cross-national studies there appears to be considerable agreement as to what behaviors are considered characteristic of intelligent individuals. Fifth, theories on social attribution of causality indicate that schooling and economic advantage appear to eliminate traditional dissimilarities in the interpretation of intelligence. Thus, cross-cultural research applying control for these variables may provide insight into transcultural interpretations of the concept of intelligence. Sixth, cultural relativism provides support for the exploration of possible cognitive
universals which may be measured through tasks which
assess the mental abilities of understanding and thinking. 
Investigating the possibility of transcultural measurement
of intelligence through use of a non-verbal test of
problem-solving abilities thus has validity in light of
current theories on the cross-cultural nature of such
skills.

Cross-Cultural Transportability of Tests

In the 1930s it was generally assumed that racial
differences in intelligence were genetic, and, therefore,
a comparative study of different racial groups with
intelligence tests would reveal the basic genetic
differences between races in intelligence (Manaster &
Havighurst, 1972). This view continues to receive support

Studies of environmental influences on intelligence
have resulted in a decreased importance given to the role
of genetics, and have elicited questions about the meaning
of racial and national differences in test scores on
standardized tests. Further, different social classes
demonstrated different characteristics on psychological
tests, indicating that SES bias may be incorporated in the
items on standardized tests. Such differences could not
be considered genetic in view of the great mobility and
mixture of marriages in the US which tend to make the
genetic composition of various social groups equal.

To address the controversy regarding the role of genetics, an attempt was made to make a variety of comparisons between culturally different groups using many testing procedures. These procedures relied heavily on verbal ability as a means of determining intelligence (Manaster & Havighurst, 1972). The resulting group differences observed were supposedly related to differences in the experience of one group or another with the items and procedures of the tests. Results showed that those tests which depend on reading and those which test knowledge of words and word meanings, were clearly unsatisfactory for study of groups with different levels of literacy. In tests where verbal language is a medium, the researcher faces dual tasks of (a) preserving syntactic variables affecting level of difficulty, and (b) preserving the semantic content of the narrative (McCauley & Colberg, 1983). For literate groups who spoke different languages, the problem of translating the tests from one language to another and retaining the same level of intrinsic difficulty was severe. Though tests have been translated with varying degrees of success (e.g., Peabody Picture Vocabulary Test, WISC-R) and specific guidelines for this procedure exist in the literature (cf. Lonner & Berry, 1986), original validity may be lost since the translated test is frequently substantially different from
As a result, nonverbal performance tests of "intelligence" have been developed that can be used with people of varying degrees of linguistic sophistication. In addition to individual intelligence tests, group tests that required little or no reading ability were developed, such as RPM, Thurstone's Test of Spatial Imagery, the nonverbal tests of the California Mental Maturity Scale, and the Porteus Maze. It soon became clear that nonliterate groups did quite well on nonverbal tests. Further, groups of people with limited English proficiency (such as American Indians and Mexicans) scored better on performance tests of intelligence than on verbal tests. A group of Indian or Mexican children in an American school frequently averaged ten or more IQ points higher on a performance test than on a verbal one.

McCaulley and Colberg (1983) have attempted to construct a theory regarding the transportability of measures based on logic or mathematics, as the criterion of a priori validity was more truly met than when using verbal or non-verbal measures. They felt that measures of reasoning ability were free of cultural idiosyncrasies and stimulus materials would be "equally germane" to all cultural groups. Their results did not substantiate their hypothesis, however. Van de Vijver and Poortinga (1985)
also supported the view that test items based on rules of logic and algebra would reflect universally applicable schemata, but no studies to date appear to support this view.

The conclusion became generally accepted by 1960, that performance tests of intelligence removed much of the difference between social-class and nationality subgroups which had been found in verbal tests, but there remained a set of differences after efforts had been made to control the studies for obvious environmental differences. Substantial evidence is available (e.g., Cole, Gay, Glick, & Sharp, 1971; Luria, 1973) that performance on formal reasoning tasks is dependent on cultural content variables. Residual group differences may be due to undiscovered differences in the experience of the members of the various social groups, or to genetic group differences (Manaster & Havighurst, 1972).

Several tests have been developed and are in use to assess abilities of persons from test-alien cultures (cultures alien to the test in its original standardized form) despite problems connected with functional equivalence of culturally transported tests. The claim has been made that various tests, most notably RPM and CCFT, assess the same abilities in all cultures (Van de' Vijver & Poortinga, 1985). Four of the most commonly used tests are briefly described.
The Culture-Fair Intelligence Test

This is a paper-and-pencil test developed by Cattell and Cattell (1959). It consists of three scales designed for varying levels of ability. Scales 2 and 3, for use with school age children and adults consist of four tests: (a) Series; (b) Classification; (c) Matrices; and (d) Conditions. The purpose of the test is to provide a measure of ability which separates the evaluation of natural intelligence from that contaminated and obscured by education. The authors believe the test can be used for all the purposes to which other intelligence tests are being applied (Cattell & Cattell, 1959).

Deviation IQs can be computed for scores on Scales 2 and 3. Sample representativeness and number of cases at some age levels falls far short of what is desirable for test construction and standardization, therefore generality of findings is limited. The test is highly speeded, although some norms exist for an untimed version. Though verbal directions are extensively required, the author states that these may be translated or pantomimed without seriously compromising results.

Internal consistency and alternate-form reliability coefficients range from .50 to .70, which are considered only marginally acceptable. Validity is demonstrated through correlation with other tests and factor analyses,
and is discussed in terms of saturation with a general intellectual factor \((g)\). Moderate correlations with academic and occupational criteria are reported.

The Cattell tests have been administered in several European countries, America, and a few African and Asian cultures. Most examiners require that a test be demonstrably related to nontest criteria such as actual evidence of ability in the real world. The lack of evidence of this type of validity for the CCFT seriously limits its usefulness. On the other hand, the test has been administered in several countries other than the US and the results have been essentially the same in cultures similar to that of the US. In very dissimilar cultures (e.g., African), however, the results are significantly different from those obtained with the standardization sample (cf. Kidd, 1962; Knapp, 1960; Gonzalez, 1982).

Raven's Progressive Matrices

This nonverbal test was developed in Great Britain by Raven (1938) and was also designed to measure Spearman's \(g\) factor. The matrix tasks are intended to measure ability to discern relations between abstract items.

RPM is composed of 60 matrices or designs from each of which a part has been removed and the examinee chooses the missing insert from 6 or 8 given alternatives. Items are grouped into five series, each different in principle, with earlier series having items relying more on visual
discrimination, and later series testing abilities to perceive analogies, permutations, alternation of pattern, and other logical relations (Anastasi, 1982). The items are arranged throughout each series in order of increasing difficulty.

Percentile norms are provided for each half-year interval between 8 and 14 years. Normative samples consist of British children, men in military service, and civilian adults. Later norms have also been established for children in other European countries. Studies in non-European cultures, however, have raised doubts about validity of the test for groups with dissimilar educational backgrounds.

The test manual presents limited reliability and no validity information. Investigative studies (cf. Mitchell, 1985) reveal moderate (.70) to high (.90) test-retest reliability for older children and adults, but poor reliability (cf. Sigmon, 1983) at younger ages. Correlations with verbal and performance intelligence tests range from .40 to .75. Studies with mentally retarded adults and different educational and occupational groups indicate fair concurrent validity (cf. Irvine, 1965). Factor analytic studies have been conducted which report that PM are heavily loaded with the g factor common to most intelligence tests (cf. Court, 1982; Higgins &

**Goodenough-Harris Drawing Test**

Comparative studies of children from various cultural subgroups with a test that purports to measure intelligence have used the Goodenough Draw-A-Man Test since 1926 when it was first described. In current use is the Goodenough-Harris revision of 1963 (Harris, 1963). The test is usable for children aged six to twelve. It requires only a piece of paper and a pencil and the instruction to "draw a man." The test is scored by counting the number and accuracy of details in the drawing (e.g., inclusion of individual body parts, clothing details, perspective, and proportion), without reference to the aesthetic value of the drawing, except that correct proportions of the body and a profile rather than a front view receives slightly higher scores. The test might be described operationally as one of accuracy of perception, since it is scored for the accuracy with which it represents a man. It might also be called a test of "mental alertness" or attention to the world around the child (Manaster & Havighurst, 1972).

Point scores on each scale are converted to standard scores. Split-half reliability indices of .68 to .89 are reported, and inter-scorer reliability coefficients range from .90 to .98. Construct validity is established by
correlation with other tests of intelligence, and the majority of reported correlations are over .50. The test correlates most highly with tests of reasoning, spatial aptitude, and perceptual accuracy.

The test has been used with American Indian, Mexican, Nigerian, and Turkish children. Cross-cultural studies report that mean scores increase consistently with the SES of subjects, and their exposure to representational art, seriously calling into question the authors' claims of "culture-fairness" (cf. Alzobaie, 1965; Kellmer-Pringle & Pickup, 1963; Ucman, 1978).

The Leiter International Performance Scale

The Leiter International Performance Scale (1948 revision) was designed to measure a wide range of functions similar to those found in verbal scales assessing mental ability (e.g., matching identical colors, shades of gray, forms, or pictures; copying a block design; picture completion; number estimation; analogies; series completion; recognition of age differences; spatial relations; footprint recognition; similarities; memory for series; and classification of animals according to habitat). The test is administered individually, and instructions, either verbal or pantomime, are almost completely eliminated.

The scale is scored in terms of mental age and ratio
IQ, but these do not maintain consistent meaning across age groups. Split-half reliability coefficients of .91 to .94 are reported, but the samples were largely heterogeneous in age, possibly artificially inflating the coefficient. Two methods of establishing validity are used, that is, age differentiation and internal consistency. Correlations with teacher ratings of intelligence, Stanford-Binet, and WISC scores are also reported, and range from .56 to .92, but again, most were obtained from heterogeneous groups.

The Leiter was developed through use with different ethnic groups in Hawaii and with several African groups, and with other nationalities. The 1948 revision also incorporates data based on further testing of American school children and recruits during World War II. Despite these validation efforts, reliability and validity data remain unconvincing.

The fact that the Leiter can be administered without the use of language may make it quite useful in some settings, provided the user is very cautious in the interpretation of scores. The lack of adequate norms and the apparent misplacement of some tests are factors that limit the general usefulness of the scale (cf. Costello & Dickie, 1970; Gonzalez, 1982; Tate, 1952).

Reactions to comparisons resulting from the application of tests in alien cultures have ranged from
criticism and controversy to downright hostility (cf. Irvine & Carroll, 1980). In spite of the existence of such conflicting opinions, so-called "culture-fair" measures have been used for research purposes in India since their inception.

Use of Non-Verbal Measures of Intelligence in India

In spite of the fact that test content influences the nature of the abilities measured by the tests, test constructors and test users in India employ verbal and non-verbal or performance tests interchangeably. This is likely to make the predictions based on the results incorrect. The issue becomes serious when it involves some important decisions concerning the individual in educational or clinical settings (Nair, 1975). Research on use of non-verbal measures of intelligence with Indian subjects as reported in major education and psychology journals are described.

1. Raven's Progressive Matrices. Since its publication in England in 1938, RPM has been extensively used there as well as in other countries, with a wide variety of groups. In India, while some attempts were made to adapt this test, a large number of researchers used it as originally developed, and have expressed widely different views about its validity (Sinha, 1975).

According to Mohsin (1959), RPM scores were not of
much help in screening students with scientific ability.

The tetrachoric correlation (a product moment correlation between continuous variables that have been artificially made dichotomous) between the RPM scores and Intermediate Examination results obtained by Mehrotra (1959) was nearly zero, that is, .02. Pattnaik (1967) obtained low correlations between the RPM and a Verbal Analogies Test. In a study by Sinha and Chandrakala (1972) correlations between the RPM scores and scores on Harris Draw-A-Man were not significant (.32) for boys, but highly significant for girls (.61). In S. N. Rao's (1962) study the correlations between RPM scores and average-grade points were the lowest (.32) for the Arts-Science group, and highest (.37) for the Engineering group studied, but the differences in scores were negligible.

A comparison of four tests of intelligence was conducted by Desai (1980). The tests used for the study were (a) Desai-Bhatt Verbal Tests (Desai & Bhatt, 1979), (b) Bhavsar Non-verbal Test, (c) Raven's Progressive Matrices, and (d) Cattell's Culture Fair Test-3. All tests were administered to a sample of 338 eighth grade students purposively selected from eight different subcultures. Mean scores of students declined progressively when moving from urban to semi-urban, rural, and tribal cultures, though students were all enrolled in school and had achieved the eighth grade level. This
suggests that the RPM scores were sensitive to differences in socio-economic status.

Ghuman (1975) also found that when British and Indian Punjabi boys brought up in Britain were compared as regards their performance on RPM (Colored), no significant differences in scores were found. However, scores of Indian Punjabi boys residing in India differed significantly from boys of the same ethnicity brought up in Britain, with the latter showing higher mean scores.

In investigating effects of socio-cultural disadvantage on intelligence using RPM, Mohanty (1980) found that both SES and grade level contributed significantly to variances in scores of children from various social classes. These findings contradict some previous studies (Rath, 1972; Sahu & Mahanta, 1977). Differences in performance are explained by contrasting high and low social class homes, where the latter are characterized by absence of order, organization, planful regularity, predictable structure, and task orientation.

On the other hand, Dosajh (1958) observed that the scores on the RPM test could safely be taken as a criterion for selection of students for technical and science courses. Rath (1954) concluded that RPM was as effective in measuring intelligence as another verbal Group Test of Intelligence.
A validity study of RPM conducted by Sinha (1975) indicated that RPM scores correlated significantly with school grades for 11 to 15-year-olds (.45, p<.01). Further, reliability analysis indicated that scores discriminated consistently between high-achieving and low-achieving students of the same age levels. Reliability coefficients of between .81 and .91 were reported indicating high internal consistency of the test. In general, means were low compared to the 50th percentile on RPM for British children, but differences were less prominent with increase in age of subjects.

The relationship between RPM and the General Mental Ability test (Jalota, 1952) was investigated by Mohan (1972) on 310 college students. Results indicated significant correlations between subjects' performance on the two measures (.65, p<.001). This lends further support to the use of non-verbal measures such as matrix items for discriminating between students with differing levels of intellectual ability.

Mehrotra (1967) examined the relationship of RPM to WISC scores for 13 to 17 year old high school students from an upper middle class social background, currently enrolled in a private school. Correlations significant at the .01 level were found for WISC Full Scale, Verbal and Performance scores. In addition, the subtests of Information, Similarities, Vocabulary, Block Design, and
Object Assembly also correlated significantly with scores on PM at .01 levels. Comprehension, Arithmetic, and Picture Completion correlated significantly at the .05 level. Thus non-verbal tests using items similar to those in the RPM appear to have validity in identifying students with differing levels of intellectual ability.

2. Cattell Culture-Fair Intelligence Test. A 1971 study by Singh and Hundal indicated that when subtests on the CCFT measuring fluid and crystallized ability were considered separately, the test showed some promise for accurately assessing intellectual ability of Indian children.

In contrast, S. Rao's (1965) attempt to standardize the test for use in one Indian state (Bihar) indicated that mean scores of the Indian sample declined steadily in comparison to scores of the American normative sample as age of subjects increased from 8 to 14 years. Thus a valid standardization could not be obtained.

In a study conducted by Desai (1980), mean scores on CCFT were consistently lower for the entire sample studied than those of the normative group. The differences in average scores on the Desai-Bhatt Verbal Test of Intelligence were more pronounced than those on RPM and Cattell's tests, supporting Cattell's findings that crystallized intelligence assessed by verbal tests shows
greater cultural differences compared to fluid intelligence assessed by culture-fair tests (Cattell, 1963, 1971; Horn & Cattell, 1966).

3. **Kahn Intelligence Test.** Chawla (1969) investigated the relationship between the Kahn Intelligence Test: Experimental Form (KIT) and the WISC. The Kahn Intelligence Test is a non-verbal test designed to measure developmental levels of normal children as well as children and adults who are visually or verbally handicapped. Correlations between scores on the KIT and WISC Full Scale scores were found to be .71 for an 11 to 13-year-old sample of 154 children. Results are encouraging for the use of non-verbal tests to differentiate between ability levels of school children in India.

4. **Kerala Non-Verbal Intelligence Test Battery.** Attempts have been made to develop non-verbal tests of intelligence standardized on Indian populations. One such attempt was the development of the Kerala Non-Verbal Intelligence Test Battery (Nair, 1973). This battery is comprised of four subtests: nonverbal classifications, nonverbal series, nonverbal analogies, and nonverbal matrices. Correlations of subtest scores with verbal IQ scores across 6 age-groups of secondary students indicated significant correlations for all age groups (.39 to .73). Significant correlations were also found with subjects'
A factor analytic validity study using Thurstone's Centroid Method reported high positive loadings of the non-verbal measures on the g factor identified, accounting for nearly 70% of the common variance of the battery.

A validity study of non-verbal intelligence tests by Nair (1975) indicated that subtests of the Kerala Non-Verbal Tests of Intelligence, the Non-Verbal Test of Intelligence, and the General Mental Ability Test (Non-Verbal) all loaded highly on the non-verbal intelligence factor identified. This would suggest that the subtests included in the study (analogies, patterns, series, water reflection, classification, and spatial relations) are consistent and valid measures of non-verbal intelligence.

5. **Draw-A-Man Test.** The Draw-A-Man Test was used by Bevli (1982) to compare children from urban advantaged, urban disadvantaged, and rural disadvantaged schools. Analysis of variance revealed that all subjects showed significant gains with training. Scores of urban advantaged school children were greater than those of both disadvantaged groups, indicating SES bias in Draw-A-Man.

6. **Other tests.** A non-language test comprised of four subtests, namely, classification, opposites, analogy, and picture arrangement, was designed by Chatterji and Mukherjee (1967). Validity studies on this test with 482
subjects aged 11 to 13 years indicated significant
correlations with tests of abstract reasoning and verbal
reasoning (.59 to .65, p<.01). Attempts to correlate
scores with students' class performance, as measured by
grades, however did not yield consistent results.

A study using the R-Test (a non-verbal test of
intelligence developed in Canada) by Warhadpande and Sethi
(1964) indicated that the test may be of some use in
differentiating between ability levels of adult subjects
in different academic fields. Significant differences
were found in the performance of persons majoring in
statistics and engineering and persons with a physical
education major. Further, the test results indicated no
significant differences between various language groups.
However, it did not appear to be of use in differentiating
between high and low performers within the academic fields
under study, which was the original intent of the study.

Use of Alexander's Performance Scale by Sinha (1980)
revealed no significant differences between tribal and
non-tribal subjects, pointing to some usefulness of the
scale with non-literate and culturally backward groups.
The study did not discuss item content of this little-
known scale; thus validity or relevance of items is not
evident. A comparison with older studies (Choudhry, 1955;
Mazumdar, 1958) appear to indicate potential for use of
the test to develop a truly fair measure for comparing
literate and non-literate subjects, but no substantial, incontrovertible evidence is presented.

In comparing performance of school children on Performance and Verbal subtests of Bhatia's Battery, Hundal (1965) found that for seventh graders there was substantial correlation between the two types of measures (.68), indicating overlap in the abilities measured. For children from grade eleven, however, scores from the two groups of subtests clearly separated into two distinct factors when Thurstone's oblique rotation was applied to the data.

**Summary**

As revealed in the literature on non-verbal intelligence testing in India, a need exists to identify instruments that demonstrate appropriate levels of reliability and validity within the cultural context of the country. Further, a need is also evident for validation studies on instruments developed and standardized in the present decade. This line of study will provide insight into the viability of transporting an already established, currently developed instrument to India for use in the educational system.

**Considerations in Cross-Cultural Research**

To control for extraneous variance and eliminate rival hypotheses, cross-cultural studies must employ
controls and procedures that are not generally applied to experimental research. This is because cross-cultural research seldom meets the requirement of the experimental paradigm, such as (a) equating samples "by chance" through randomization (Campbell & Stanley, 1966, p. 2), and (b) exercising control over all treatments under study (Cook & Campbell, 1979, p. 98-99) and, therefore, cannot use the inferential strategy of a proper experiment (Malpass & Poortinga, 1986). Consequently, statistical techniques must be utilized to control for extraneous effects.

Problems in analyzing cross-cultural equivalence of tests arise when only one strategy or technique is used. A multistrategy approach is recommended (Hui & Triandis, 1983, 1985) which combines highly mathematical and statistical techniques with those that place considerable demands on the researcher's conceptualization ability. Research methodology previously used in cross-cultural equivalence studies includes (a) direct comparison and crude translation (t tests and MANOVAS) after administering the same test to both cultures being studied (Gordon & Kikuchi, 1966); (b) regression methods to check if regression parameters of the criteria or constructs to which the score is to be generalized are the same for the population to be studied (Poortinga, 1971, 1975); (c) coscoring methods which presuppose the existence of factors that have conceptual/functional equivalence, and
are equivalently operationalized in the instruments (Cattell, 1957); (d) item response theory approach which measures item characteristic curves (the probability of responding to an item in a certain specified manner at different levels of the latent trait to be measured (Drasgow & Komocar, in press); (e) response pattern method which correlates the order ranking of item difficulty for each pair of cultures concerned (Irvine & Carroll, 1980); (f) translation techniques such as back translation, bilingual and committee approach, decentering and pretests (Brislin, 1970, 1976, 1980); (g) internal structure congruence which includes factor analysis (Cattell, 1969, 1978), multi-dimensional scaling (Hui & Triandis, 1983), maximum likelihood factor analysis (McGaw & Joreskog, 1971), and comparison of correlation matrices (Jennrich, 1970); (h) combined emic-etic approach which describes behaviors primarily in terms of the concepts employed in the culture (Malpass, 1977); and (i) validation by nomological network which tests whether the construct is embedded in the same network of constructs in the same manner as it is in a second culture (Cronbach & Meehl, 1955).

Prior research in cross-cultural transferability of tests has provided evidence that observing certain methodological constraints leads to results that may be
interpreted with greater confidence. A rationale is provided for the methodology used in the present study based on a literature review of methodology. The major areas of differentiation between cross-cultural and traditional experimental research are (a) sample selection, (b) hypothesis formulation, (c) instrument selection, (d) data collection procedures, and (e) statistics for equivalence.

Sample Selection

Rival hypotheses due to sample characteristics arise any time comparisons are made between one culture and another, since the differences found may be attributed to different selection methods or to different qualities of the samples (e.g., age, SES) rather than to cultural differences (Brislin, Lonner, & Thorndike, 1973). Frequently cross-cultural researchers have no sampling plan and tend to interview people who seem intelligent, talkative, cooperative, or who are readily available. Although rare, random samples have been drawn in cross-cultural research but the costs of such samples are great even if the methodological problems can be solved. Further, this procedure still does not ensure that differential results from two cultures are not due to any of the several aspects of the two cultures. Sometimes stratified samples are attempted, but in many cases, especially in non-Western cultures, the necessary
information, such as a list of homes, streets, or even communities, will not exist.

In general, the literature recommends that experimenter control be exercised in selecting subjects for study. The main direct controls (e.g., sex, socio-economic status, age) must be identified so that through the sampling design enough cases will be tested to exercise those controls. Further, equivalence on dimensions not under direct study should be attempted to the extent possible (Brislin et al., 1973). Thus experimental effects will be more closely related to the independent variable.

Based on a review of studies previously conducted in India, the following variables were selected to be controlled: (a) literacy level of subjects, (b) area of residence, and (c) socio-economic level.

1. **Literacy level.** Studies by Sinha (1976, 1978) examining effects of schooling on ability to perform on tasks requiring perceptual discrimination indicated significant effects.

2. **Area of residence.** Nanda, Das, and Misra (1965) compared children between 6 and 10 from urban, rural, and tribal environments on their ability to discriminate pairs of geometrical figures. Ability to discriminate the patterns varied directly with the degree of urbanization.
Being from a supposedly more civilized environment operated as a factor in efficient perception of forms. Sinha (1976; 1978) conducted a number of studies using different perceptual measures to analyze the effects of eco-cultural differences among urban, rural, and tribal children. Urbanization and caste seemed to operate as two of the major significant influences.

3. **Socio-economic level.** In studies conducted, Tripati and Misra (1975, 1976) observed a negative relationship between scores on a multidimensional scale of prolonged deprivation and different measures of cognitive efficiency including an intelligence test, Hudson's depth perception test, concept formation, and perceptual identification of simple drawings embedded in complex figures. Observed variations in cognitive functioning were accounted for in terms of prolonged socio-cultural disadvantage. Panda and Das (1970) observed that children from economically privileged families were better in word reading on the Stroop Test than children from economically deprived families. Further, Jachuck and Mohanti (1974), used the Colored Progressive Matrices and the Stroop Test on mentally retarded children of 8 to 10 and 14 to 16 years of age of varying socioeconomic status. They observed that SES was a significant influence, the higher group performing better. The difference was greater at the higher age level. Das and his associates (Das, 1973;
Das, Jachuck, & Panda, 1970; Das & Singha, 1975) conducted extensive investigations on cultural deprivations and cognitive competence. An array of perceptual and cognitive measures were administered to Brahmin (high caste), economically privileged, and Harijan (untouchable), economically disadvantaged, children. In color naming speed the rich were superior to the poor. Raven's Colored Matrices and the Stroop Test differentiated the groups more than did short-term memory and recognition tests.

Hypotheses Formulation

Guidelines for cross-cultural comparative research procedures do not generally recommend use of the null hypothesis as a basis for testing effects. Since the a priori probability of rejecting the null hypothesis is often very high, a hypothesis of no intercultural difference cannot be considered a meaningful alternative to a research hypothesis.

Within the experimental paradigm, though the null hypothesis cannot be logically proven, an assumption may be made that it is true (Cook & Campbell, 1979, p. 44-45). This is not a reasonable presumption in cross-cultural studies where the allocation of subjects is determined by their membership in a specific cultural group and the observed differences between cultures follow from
antecedent conditions on which the researcher has exercised no influence.

Further, to protect against Type I and Type II errors in experimental research, increase in sample size is a strategy commonly employed. In cross-cultural research increased sample size contributes to an increased confidence in an erroneous rejection of the null hypothesis (i.e., the probability that an observed difference is erroneously attributed to a "real" cultural factor can increase when the body of data is enlarged) [Malpass & Poortinga, 1986].

**Instrument Selection**

Cross-cultural research has been defined by Ecksenberger (1973) as "the explicit, systematic comparison of psychological measures obtained under different cultural conditions, in which cultural conditions...serve as the independent variables" (p. 101). Thus the hypothesis being tested in a typical cross-cultural study is that different cultures result in different forms of behavior. A problem arises regarding the dependent variables: "how can we tell whether or not a test (or other measurement procedure) measures the same psychological construct in different cultures?" (Fredrickson, 1977, p. 14).

Guidelines for development or selection of culturally reduced tests include:
1. The test is well constructed with a representative normative sample and empirical evidence of reliability and validity (Duffey, Salvia, Tucker, & Ysseldyke, 1981).

2. A performance measure is used rather than a paper and pencil task.

3. Instructions are pantomimed to the subjects, not presented orally or in writing.

4. Preliminary practice items are incorporated into the test.

5. The test is untimed.

6. Test items have abstract content instead of pictures and passages to read.

7. Items require reasoning or problem-solving, not specific factual information.

8. Novel problems are presented to avoid the recall of previously learned information (Jensen, 1980).

Data Collection Procedures

Sinha (1981) identified some of the problems which must be addressed when conducting human assessment in India. Types of extraneous effects which may be controlled through application of carefully considered data collection procedures include:

1. Examiner effect. The examiner used to conduct assessment must be from the subjects' culture and possess
an understanding of cultural differences in language style, idiomatic speech and jargon. Further, an indigenous examiner would be able to accurately interpret subject responses (Cronbach & Drenth, 1972; Sinha, 1981).

2. Setting effect. Irvine and Carroll (1980) stipulate that the environment of testing be as convivial and enjoyable as possible. The very presence of the examiner or an unfamiliar setting can lead to variations in subjects' performance that are not related to their true abilities (Sinha, 1981).

3. Effect of unfamiliar materials. If subjects are tested using methods with which they have no prior experience, this may lead to underestimation of ability (Cronbach & Drenth, 1972; Irvine & Carroll, 1980; Sinha, 1981).

4. Response time constraints. Time has different meaning in different cultures, not only the time to complete a test, but the tempo of daily life. Undue time pressure on examinees may affect scores on any measuring instrument being used with them (Anastasi, 1982).

5. Ambiguous directions. Tests which require extensive written directions are not recommended for cross-cultural study. This is because literal translations are sometimes misleading, and properly translated directions do not always maintain the same level of difficulty in the translated language (Cronbach &
Data sets must meet certain criteria for equivalence before relevant comparisons across cultures can be made. When a measuring instrument is constructed on the basis of a representative sample of elements from a cross-culturally identical universe of stimuli and when no inference beyond that inference is to be attempted on the basis of results, data may be said to be equivalent. Stimulus bias exists when there is no precise description of the universe or when only a small number of items are used to represent a heterogenous universe. Method bias may be found when ambient variables such as interviewer-subject interactions or understanding of the task by subjects may have led to cross-cultural differences in results. Universe bias is found when the data sets are obtained with measuring instruments constructed on the basis of non-identical universes (Malpass & Poortinga, 1986).

The considerations associated with several major types of equivalence have been summarized by Hui & Trandis (1985) and are as follows:

1. Conceptual/functional equivalence. A construct that can be meaningfully discussed in the cultures concerned is said to have cross-cultural equivalence.
That is, subjects have an equal understanding of the meaning of behavior or of concepts pertaining to behavior (Malpass & Poortinga, 1986). Further, behaviors are functionally equivalent if they are performed in two cultures to achieve identical grades. "If similar activities have different functions in different societies, their parameters cannot be used for comparative purposes" (Frijda & Jahoda, 1966, p. 116)

2. Equivalence in construct operationalization. If a construct is operationalized in the same procedure in different cultures, the instrument thus derived is equivalent in construct operationalization across cultures. The operationalization should be "equally meaningful" in the cultures being studied.

3. Item equivalence. Assuming that a construct has similar meaning in two cultures, is manifested and operationalized in similar ways, the next consideration is that the construct has to be measured by the same instrument. Only by doing this can cultures be numerically compared.

4. Metric/Scalar equivalence. If other types of equivalence are attained, and if it can be demonstrated that the construct is measured by the same metric, then it has scalar equivalence. This indicates that a numerical value on the scale refers to the same degree, intensity, or magnitude of the construct regardless of the population
of which the respondent is a member. "When the psychometric properties of two (or more) sets of data from two (or more) cultural groups exhibit essentially the same coherence of structure" (Berry, 1980, p. 10) then metric equivalence is established. It is tested by regression and item bias testing.

**Statistical Methodology**

To make comparative statements regarding test validity across cultures, certain statistical procedures have been recommended. These include:

1. **Direct comparison.** Multivariate analysis of variance and t-tests may be used to compare mean scores across cultures after the same test has been administered in both cultures (Gordon & Kikuchi, 1966).

2. **Multiple Regression.** Analysis to check if regression parameters of the constructs to which the score is to be generalized may be conducted. The principal questions that have been raised regarding test bias pertain to validity coefficients (slope bias) and to the relationship between group means on the test and on the criterion (intercept bias) (Anastasi, 1982; Mercer, 1984).

**Slope Bias.**

When both test and criterion scores are expressed as standard scores (SD = 1.00), the slope of the regression line equals the correlation coefficient. For this reason,
if a test yields a significantly different validity
coefficient in the two groups, this difference is
described as slope bias. This type of group difference is
often designated as "differential validity." Several
investigators have also employed the term "single-group
validity" to refer to a test whose validity coefficient
reached statistical significance in one group but failed
to do so in another.

**Intercept Bias.**

A test exhibits intercept bias if it systematically
overpredicts or underpredicts criterion performance for a
particular group. For example, the two groups may have
regression lines with the same slope but different
intercepts. Although the validity coefficients computed
within each group will be equal, any test score will
correspond to different criterion scores in the two
groups. The same test score will thus have different
predictive meaning for the two groups. Intercept bias
discriminates against the group with the higher intercept.
Reilly (1973) has demonstrated mathematically that this
will occur if the two groups differ in a third variable
(e.g., sociocultural background) which correlates
positively with both test and criterion.

3. **Item response theory method.** As one approach to
the investigation of test bias for minority groups, the
analysis of item bias has received considerable attention.
Such analysis is concerned essentially with the relative difficulty of individual test items for groups with dissimilar cultural or experiential backgrounds. A comprehensive survey of studies on this topic can be found in Jensen (1980, p. 552-580). The findings have been largely negative. For widely used, standardized aptitude and achievement tests, few items have emerged as significantly biased. There is also little consistency in the items identified as biased by the different procedures (Ironson & Subkoviak, 1979; Sandoval & Whelan, 1979). Moreover, when significant differences in item difficulty did appear, they were more closely associated with differences in performance level on the test as a whole than with cultural group membership. To meet these methodological difficulties, item bias can be investigated by the methods of item response theory, which permit the measurement of item difficulty independently of other item parameters (Anastasi, 1982, p. 219).

4. **Factor analysis.** "If the factors extracted account for all the extractible variance by being the same in number, if the [same] proportions of variance are extracted by each factor, and if the correlations among all the factors extracted for the control and experimental groups are not significantly different for the groups to be compared, then some confidence can be placed in the
supposition that tests and individuals have interacted, on the average, in the same fashion" (Irvine & Carroll, 1980, p. 219).

Confirmatory factor analysis represents a substantial improvement over the exploratory factor analytic model. In the confirmatory model the researcher imposes "substantively motivated constraints" (Long, 1983). These constraints determine (a) intercorrelations between common factors, (b) the observed variables affected by common factors, (c) unique factors affecting observed variables, and (d) intercorrelations between unique factors. Thus a theorized hypothesis about how a domain is structured may be tested using statistical means (Bentler, 1985).

Maximum Likelihood factor analysis may be conducted using the following procedure. First a variance-covariance matrix of the observed variables is obtained for each sample. These variance-covariance matrices are fitted with a factor analytic model that assumes invariance across the samples. Maximum likelihood is a statistical rather than purely psychometric technique, which gives it an advantage over other techniques (Cattell, 1978), as inferences are made from a given sample to actual population parameters. The maximum likelihood estimator is normally distributed, and has as small a variance as any other estimator (Long, 1983). Parameters of the model (such as factor loading, factor
variance and covariance, etc.) may be assigned a starting value or constrained at a fixed value throughout the process. Thus hypotheses may be tested for different values of loadings, variances, and covariances according to the four combinations of (a) restricted versus nonrestricted, and (b) unique versus non-unique conditions (Cattell, 1978).

A large sample chi-square test of goodness-of-fit may be given at the conclusion of the estimation to indicate whether the data from the several groups are sufficiently similar to be described by the one-factor model specified (Hui & Triandis, 1985). Other goodness-of-fit indices developed include the Bentler-Bonett and Bentler-Bonett non-normed index (Bentler, 1985).

The following constraints must be considered before conducting a factor analysis:

1. Examine data to be certain that linear relationship exists. Little is known about the effects of violation of the assumption of normality on the properties of maximum likelihood estimators (Long, 1983). Inappropriate application is likely to produce incorrect results.

2. Use the common-factor model over principal component analysis. "Principal components analysis summarizes data by means of a linear combination of the
observed data. This does not reveal any underlying causal structure, if such a structure does not exist. The common factor model, on the other hand, represents the covariance structure in terms of a hypothetical causal model" (Kim & Mueller, 1978, p. 14).

Principal components will not necessarily account for the observed correlations, whereas a two-common factor model can. If there is no correlation between the variables under study, there will be no principal components. Every component will be as good or bad as the other, each accounting for only unit variance.

3. Use maximum-likelihood estimation with goodness-of-fit indices to determine number of factors. As discussed above, the maximum likelihood estimator permits inferences to actual population parameters. One consideration, however, is that sample size must be fairly large (at least over 80) (Long, 1983), and preferably between 100 and 200.

4. Specify number of non-zero weights. "Non-zero coefficients for factors indicate that changes in factors do not directly cause changes in the observed values of the variables. Decisions regarding which weights must be constrained is based on theoretical rather than statistical considerations" (Long, 1983, p. 7). A researcher would impose constraints if it made substantive sense to do so based on previous research, and interpretability of
5. Conduct successive higher-order factor analyses and attempt to orthogonalize resulting factor structure. Primary factors may be further factor analyzed to obtain second-order factors. Third-order factors may further be obtained from the second-order factors (Cattell, 1978).

In the present study no attempt will be made to obtain higher-order factors as the intent is not to develop cognitive profiles for subjects. The restrictions in the range of the data obtained do not support such an endeavor.

Factor analytic studies reviewed by Reynolds (1982) have supported the consistency of factors across groups given the same cognitive abilities tests. Though identical factors may not reflect the innateness of abilities being measured, whatever is being measured is measured in the same manner, and is the same construct for both groups (Reynolds & Brown, 1984).

Conclusion

Since a multistrategy approach is recommended by Hui and Triandis (1985) as being most recommended for determining cross-cultural equivalence of tests, the decision was made to utilize (a) regression methods, (b) response pattern methods, and (c) maximum likelihood factor analysis, to investigate the relationship between
performances of the two groups under study (Indian and American).

The following hypotheses were formulated:

1. **Hypothesis 1.** Significant differences will be found in the factor structure of the scores on individual items of the normally achieving Indian subjects and the TONI normative group.

2. **Hypothesis 2.** Significant differences will be found in the item characteristic curves of the individual items for the Indian normally achieving subjects and the TONI normative group.

3. **Hypothesis 3.** The sample of normally achieving subjects from India will show significantly higher mean scores on the TONI than the sample of mentally retarded Indian subjects.

4. **Hypothesis 4.** The sample of normally achieving third grade subjects from India will show significantly higher mean scores on the TONI than the sample of second grade normally achieving subjects.
CHAPTER II

METHOD

Research Design

The research design for the present study was based on the established principles and guidelines discussed in Chapter I. Specific procedures for (a) sample selection, (b) hypotheses formulation, (c) instrument selection, (d) data collection procedures, (e) statistics for equivalence, and (f) statistical methodology are described.

Sample Selection

Based on a review of studies previously conducted in India, the following variables were selected to be controlled: (a) literacy level of subjects, (b) area of residence, and (c) socio-economic level.

1. Literacy level of subjects. Subjects selected for the present study were seven, eight, and nine-year-old school children currently enrolled in second (n = 36) and third (n = 54) grade classes. Ninety-one normally-achieving children were selected from schools in Lucknow, India. Eighteen mentally retarded children selected for the study were also currently enrolled in school. Since no national criteria exist to identify the mentally retarded, children selected were those enrolled in a
school for the mentally retarded, and whose WISC-R scores indicated their cognitive aptitude to be between 69 and 40 IQ points. The comparison group selected from the TONI normative sample consisted of children matched on grade level with no significant difference in mean age from the Indian normally-achieving group (t = 1.08, df = 224). Table 1 presents mean ages of subjects and the age range.

Table 1

Mean Age and Age Ranges of Subjects

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean Age</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indian (Normally-Achieving)</td>
<td>95.9</td>
<td>84-120</td>
</tr>
<tr>
<td>Indian (Mentally Retarded)</td>
<td>108.1</td>
<td>84-130</td>
</tr>
<tr>
<td>American (Normative Sample)</td>
<td>95.0</td>
<td>84-111</td>
</tr>
</tbody>
</table>

2. Area of residence. Since significant differences in performance of Indian children from rural and urban areas have been found, subjects for the present study were selected from schools in a major urban area (Lucknow, India) where they have exposure to a variety of experiences. Subjects selected from the normative sample of the TONI were also from urban residential areas.

3. Socio-economic level. Socio-economic variables were controlled for by selection of subjects from middle
to upper-middle class backgrounds, enrolled in private schools. In India private school education is accessible only to middle- and upper-middle-class children, and is comparable to standards of education in the US. Public school education suffers from a severe lack of funding and is frequently of poorer quality than private school education. Since the children were selected from private schools, an assumption was made of the quality of education as compared to that of the comparison group. In addition, schools where the medium of instruction is English were selected, where children are exposed to western literature, art, and concepts. Hollingshead’s Index was used to verify socio-economic standing. Table 2 presents information on socio-economic background of the selected groups.

Table 2

Percentages of Subjects at Hollingshead Index Socio-Economic Levels

<table>
<thead>
<tr>
<th>Level</th>
<th>Indian (Normally-Achieving)</th>
<th>American (Normative Sample)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>28.1</td>
<td>16.0</td>
</tr>
<tr>
<td>II</td>
<td>40.4</td>
<td>23.2</td>
</tr>
<tr>
<td>III</td>
<td>31.5</td>
<td>48.8</td>
</tr>
<tr>
<td>IV</td>
<td></td>
<td>12.0</td>
</tr>
</tbody>
</table>
Hypotheses Formulation

In keeping with prior research in the area of cross-national testing, it was decided that hypotheses be stated as research hypotheses.

Instrument Selection

The TONI was selected for this investigation as it meets criteria designated by Jensen (1980) and Duffey, Salvia, Tucker, and Ysseldyke (1981) as essential for cross-cultural use of an assessment instrument. These criteria are detailed in Chapter I. Due to the nature of the instrument, it can be used in both cultures in its original standardized form which means that the results are accessible to a variety of analysis on equivalence (Malpass & Poortinga, 1986).

The TONI is a language free test designed to measure cognitive ability, appropriate for use with subjects ranging in age from 5 years, 0 months through 85 years, 11 months. It consists of two equivalent forms, Form A and Form B, of 50 items each arranged in a hierarchy from easy to difficult. Five item types are used in the TONI and, for each of the five types, preliminary training items are included in the test. The test kit consists of a manual with instructions and background information, and a picture book with the test items.

The TONI is untimed, generally requiring
approximately 15 minutes to administer. The examiner
pantomimes the instructions and the subject points to the
appropriate response from a choice of four to six
alternatives. Thus, there is no listening, speaking,
reading, or writing required in the administration of the
test or in the examinees' responses.

The basis of all the TONI items is visual problem-
solving. The authors selected this based on the work of
Sternberg (1980) who considered problem solving to be a
general component of intelligent behavior rather than a
subskill. Further, Resnick and Glaser (1976) have also
cited problem solving as the basis for functional
independence. Problem solving lends itself readily to the
abstract content and the non-verbal testing format used in
the TONI.

The TONI purports to measure problem-solving ability,
rather than general "intelligence," by measuring
respondents' performance on five types of problem-solving
tasks:

1. Simple matching.
2. Analogies: (a) matching, (b) addition, (c)
   subtraction, (d) alteration, and (e) progressions.
3. Classification.
4. Intersections.
5. Progressions.
**Item Selection**

The format of TONI items was developed based on the content and format employed by other non-verbal and performance tests such as Raven's Progressive Matrices, performance subtests on the Wechsler scales, and the Leiter International Performance Scale. Abstract/figural items (items containing designs and shapes that are not symbolic and have no inherent meaning) were considered to be appropriate to meet the goal of developing a culturally and language reduced test.

**Standardization**

The TONI was standardized on 1,929 subjects ranging in age from 5 years, 0 months, to 85 years, 11 months from 28 states in the US. Subjects suspected of having impaired intellectual ability were excluded in the standardization group. In most instances, demographic characteristics approximated the proportions reported in the 1980 census for the United States population as a whole (Statistical Abstract for the United States, 1980). The normative sample consisted of 47% male and 53% female subjects. Seventy-seven percent were Caucasian, 14% Negroid, and 9% Mongolian. Urban and suburban areas accounted for 78% of the sample, and rural areas for 22%.

**Scoring**

TONI scores may be reported as percentile ranks, or TONI Quotients, which are standardized scores with a mean
of 100 and a standard deviation of 15. This reporting format makes scores readily comparable with other measures using the same scale.

The examinee's basal performance is the item just below his/her first error, and the ceiling is considered to be the last of three incorrect items out of five. The raw score is the sum of the number of the basal item and the total correct responses between the basal and ceiling items. Tables are provided to convert raw scores to TONI Quotients for each age group. A sample scoring sheet is presented in Appendix A.

Administration

The TONI may be administered by "classroom teachers, special education personnel, psychologists, psychometrists, educational diagnosticians, or any other professionals who read and follow the directions in this (TONI) manual and adhere rigidly to the specific policies governing intelligence testing and interpretation in their particular administrative unit" (Brown et al., 1982, p. 17).

The test is to be administered individually, presented in a language free format, though the examiner and examinee may communicate prior to testing to establish rapport. Specific administration procedures include:

1. The picture book is to be placed in front of the
examinee so that the stimulus items are on top and response choices on the bottom.

2. Six training items are administered to familiarize the examinee with the testing format and nature of items. The examinee is advised of the correct response for each training item.

3. Testing is initiated at the first item on the appropriate form for very young examinees, those suspected to be intellectually handicapped, or those who did not understand the training items. For older subjects recommended starting items are indicated on the scoring sheet.

4. Examinees' responses are recorded on the scoring sheet corresponding to the form used.

5. If the examinee does not make five correct answers in a row from the starting item, the examiner goes to the first item tested and tests backwards until five correct responses in a row are obtained. Testing is discontinued when the examinee has made three errors in a sequence of five items.

Reliability

Reliability analyses of four types were reported in the TONI manual. They are:

1. Internal Consistency Reliability. This was established using both Coefficient Alpha and Kuder Richardson Formula 21. Coefficient Alphas ranging
from .78 to .91 were reported based on 100 protocols randomly drawn from each of four age groups. Kuder-Richardson 21 coefficients were calculated for each of the age intervals for which norms are reported. K-R 21 coefficients for the 7 year, 0 month, to 8 year, 5 month, and 8 year, 6 month to 10 year, 11 month age groups were .8 and .9 respectively for Form A, and .8 and .8 for Form B.

2. **Standard Error of Measurement.** For TONI raw scores of the 7 year, 0 month, to 8 year, 5 month, and 8 year, 6 month, to 10 year, 11 month age groups SEM was 3 and 2 for Form A, and 3 and 3 for Form B. For the TONI Quotient SEM was 7 and 5 for Form A, and 7 and 7 for Form B.

3. **Alternate Forms Reliability.** Separate coefficients were calculated for Form A and Form B of the TONI. Coefficients of equivalence between the two forms ranged from .78 to .95.

4. **Reliability with Deviant Populations.** This was established through validity studies with educable mentally retarded, deaf, and learning disabled samples. All K-R coefficients computed on these samples exceeded values of .80, and 25% exceeded .90.

**Validity**

Two types of validity are reported in the test manual.

1. **Concurrent Validity.** Scores of subjects on the
TONI were correlated with their performance on (a) Raven's Progressive Matrices (Raven, 1938), (b) the Leiter International Performance Scales (1948), (c) the Wechsler Intelligence Scale for Children-Revised (Wechsler, 1974), (d) the Otis-Lennon Mental Ability Test (Otis & Lennon, 1970), (e) the Iowa Test of Basic Skills (ITBS) (Lindquist & Hieronymus, 1956), (f) the SRA Achievement Series (Naslund, Thorpe, Lefever, 1978), and (g) the Stanford Achievement Test (SAT) (Madden & Gardner, 1972). The coefficients provide strong evidence of the TONI's concurrent validity as all exceeded the .35 criteria, with 41% of them exceeding values of .80. The TONI seems to be related more strongly to intelligence than to achievement and to language-free or composite measures of intelligence than to verbal measures.

2. Construct Validity. This was partially determined by correlational comparison of subjects' performance on the TONI with that on other aptitude measures. High correlations were found (.68 to .86). Further, mean raw scores of subjects followed the developmental pattern reported by other researchers, notably by Wechsler (1949) and Spearman (1923). A comparison of scores of a mentally retarded sample with that of the normative group indicated significant differences in performance ($t = 6.23$, $p < .05$), indicating the TONI's usefulness as a diagnostic tool.
Data Collection Procedures

Procedures which control for extraneous effects as described by Sinha (1981) were utilized. Briefly they are:

1. Examiner effect. To maximize correct interpretation of response, examiners indigenous to the subjects' culture were used. These were persons trained in the fields of education and psychology who were native to the country under study (India).

2. Setting effect. Subjects were tested at their own schools, in as close proximity to their classrooms as possible.

3. Effect of unfamiliar materials. The design of the TONI is in consonance with recommended administration procedures for tests used in alien cultures. The TONI provides practice items, so subjects may familiarize themselves with the format of the test. Correct responses for the training items are presented to clarify to subjects the type of response desired.

4. Response time constraints. The TONI is untimed, allowing subjects sufficient time to make a response without undue pressure for speedy performance.

5. Ambiguous directions. Since pantomime is used for giving subjects directions rather than written or oral language, the risk of noncomprehension by subjects of test
directions due to linguistic complexity are minimized.

Statistics for Equivalence

Concern has been expressed about equivalence of data sets used in cross-cultural studies. The major types of equivalence described as essential by Hui and Triandis (1985) are addressed below.

1. Conceptual/functional equivalence. The selection of subjects with similar levels of literacy in both cultures, and comparative exposure to testing will help maintain functional equivalence. Since the TONI items consist of stimuli that are "not symbolic and have no inherent meaning," conceptual equivalence was maintained, as the items were equally meaningless to subjects from both the cultures compared in the study.

2. Equivalence in construct operationalization. As evidenced in the literature, the concept of intelligence is similarly operationalized in both India and the US. Indian psychology is rooted in American and European psychological thought, and in both problem-solving is considered a primary component of intelligence.

3. Item equivalence. Utilization of identical testing instruments in both cultures (India and the US normative group) ensured item equivalence. This was possible because of the the format of the TONI which does not require the use of language for directions or student response and thus could be used in its original form in
both cultures.

4. Metric/scalar equivalence. Since administration procedures and items were identical for both cultural groups, it was possible to use the same scoring methods and scales. Thus direct comparisons could be made across groups without any adjustment of scores.

**Statistical Methodology**

Methodology for each hypothesis under study is described individually.

1. **Hypothesis 1.** Significant differences will be found in the factor structure of the item scores of the normally-achieving Indian subjects and the TONI normative group.

   Confirmatory factor analysis using maximum likelihood estimation was conducted to test this hypothesis. Goodness of fit indices were examined to evaluate the relationship of the factor structures for the two groups.

2. **Hypothesis 2.** Significant differences will be found in the item characteristic curves of the individual items for the Indian normally achieving subjects and the TONI normative group.

   Item characteristic curves were developed for the two groups under study. The parameter values (discrimination parameter and difficulty parameter) of the curves for both groups were tested to identify possible
differences.

3. **Hypothesis 3.** The sample of normally achieving subjects from India will show significantly higher mean scores on the TONI than the sample of mentally retarded Indian subjects.

   Analysis of variance was conducted to test the relationship between mean scores of the two groups. Regression analysis, using scores of both groups on the WISC-R as the criterion variable, was conducted to determine the relationship between the validity coefficients (slopes) of the regression lines for the two groups and their intercepts.

4. **Hypothesis 4.** The sample of normally achieving third grade subjects from India will show significantly higher mean scores on the TONI than the sample of second grade normally achieving subjects.

   Analysis of variance was conducted to test the relationship between the mean scores of the two groups. Regression analysis, using scores of both groups on the WISC-R as the criterion variable, was conducted to determine the relationship between the validity coefficients (slopes) of the regression lines for the two groups and their intercepts.
CHAPTER III

RESULTS

After appropriate controls were applied to ensure equivalency between samples in terms of literacy level, area of residence, and socio-economic status, and to assure conceptual, construct, metric/scalar, and item equivalence, the data were analyzed to test the proposed hypotheses. Results are presented for each of the four hypotheses under study.

Hypothesis 1

Significant differences will be found in the factor structure of the item scores of the normally-achieving Indian subjects and the TONI normative group.

Confirmatory factor analysis was applied to the data to investigate this hypothesis. An initial exploratory analysis on the data from the normative group revealed a seven-factor structure by the Kaiser eigenvalue criterion. The factor pattern matrix derived from the normative group, and used as a model for the confirmatory analysis, is presented in Table 3.

The seven-factor model obtained by exploratory analysis, using varimax rotation, could not be confirmed as the covariance matrix was not identified.
### Table 3

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor</strong></td>
<td><strong>Pattern Matrix (Varimax Rotation)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>T2</td>
<td>0.001</td>
<td>0.077</td>
<td>0.032</td>
<td>0.721</td>
<td>0.004</td>
<td>0.026</td>
<td>0.026</td>
</tr>
<tr>
<td>T3</td>
<td>0.003</td>
<td>0.030</td>
<td>-0.005</td>
<td>0.546</td>
<td>0.636</td>
<td>0.099</td>
<td>-0.052</td>
</tr>
<tr>
<td>T4</td>
<td>-0.004</td>
<td>0.066</td>
<td>0.059</td>
<td>0.627</td>
<td>0.189</td>
<td>-0.108</td>
<td>-0.066</td>
</tr>
<tr>
<td>T5</td>
<td>0.019</td>
<td>0.201</td>
<td>-0.034</td>
<td>0.757</td>
<td>0.071</td>
<td>0.263</td>
<td>0.147</td>
</tr>
<tr>
<td>T6</td>
<td>0.012</td>
<td>0.146</td>
<td>0.055</td>
<td>0.343</td>
<td>-0.011</td>
<td>0.681</td>
<td>0.011</td>
</tr>
<tr>
<td>T7</td>
<td>0.011</td>
<td>0.059</td>
<td>0.049</td>
<td>0.205</td>
<td>0.272</td>
<td>0.307</td>
<td>0.051</td>
</tr>
<tr>
<td>T8</td>
<td>0.020</td>
<td>0.231</td>
<td>0.066</td>
<td>-0.013</td>
<td>0.737</td>
<td>-0.049</td>
<td>-0.157</td>
</tr>
<tr>
<td>T9</td>
<td>0.030</td>
<td>0.464</td>
<td>0.012</td>
<td>0.554</td>
<td>-0.029</td>
<td>0.010</td>
<td>0.063</td>
</tr>
<tr>
<td>T10</td>
<td>0.031</td>
<td>0.243</td>
<td>-0.032</td>
<td>0.052</td>
<td>0.277</td>
<td>0.792</td>
<td>-0.064</td>
</tr>
<tr>
<td>T11</td>
<td>0.031</td>
<td>0.485</td>
<td>0.057</td>
<td>0.279</td>
<td>0.196</td>
<td>0.030</td>
<td>-0.062</td>
</tr>
<tr>
<td>T12</td>
<td>0.030</td>
<td>0.478</td>
<td>0.173</td>
<td>0.240</td>
<td>0.030</td>
<td>0.228</td>
<td>-0.157</td>
</tr>
<tr>
<td>T13</td>
<td>0.044</td>
<td>0.623</td>
<td>0.193</td>
<td>0.146</td>
<td>0.091</td>
<td>-0.062</td>
<td>-0.134</td>
</tr>
<tr>
<td>T14</td>
<td>0.051</td>
<td>0.465</td>
<td>0.236</td>
<td>0.114</td>
<td>0.067</td>
<td>0.252</td>
<td>0.052</td>
</tr>
<tr>
<td>T15</td>
<td>0.041</td>
<td>0.585</td>
<td>0.303</td>
<td>0.052</td>
<td>0.032</td>
<td>0.248</td>
<td>-0.342</td>
</tr>
<tr>
<td>T16</td>
<td>-0.098</td>
<td>0.604</td>
<td>0.128</td>
<td>0.044</td>
<td>0.064</td>
<td>0.237</td>
<td>0.105</td>
</tr>
<tr>
<td>T17</td>
<td>0.099</td>
<td>0.572</td>
<td>0.346</td>
<td>0.033</td>
<td>0.113</td>
<td>-0.016</td>
<td>0.116</td>
</tr>
<tr>
<td>T18</td>
<td>0.086</td>
<td>0.468</td>
<td>0.588</td>
<td>-0.041</td>
<td>0.099</td>
<td>0.072</td>
<td>-0.208</td>
</tr>
<tr>
<td>T19</td>
<td>-0.035</td>
<td>0.723</td>
<td>-0.142</td>
<td>-0.018</td>
<td>-0.059</td>
<td>-0.026</td>
<td>0.044</td>
</tr>
<tr>
<td>T20</td>
<td>0.127</td>
<td>0.232</td>
<td>0.703</td>
<td>0.005</td>
<td>-0.009</td>
<td>0.144</td>
<td>-0.194</td>
</tr>
<tr>
<td>T21</td>
<td>0.209</td>
<td>0.243</td>
<td>0.690</td>
<td>0.041</td>
<td>0.065</td>
<td>-0.047</td>
<td>0.253</td>
</tr>
<tr>
<td>T22</td>
<td>-0.050</td>
<td>0.368</td>
<td>0.174</td>
<td>-0.126</td>
<td>0.153</td>
<td>0.051</td>
<td>0.715</td>
</tr>
<tr>
<td>T23</td>
<td>0.326</td>
<td>0.154</td>
<td>0.603</td>
<td>0.085</td>
<td>-0.057</td>
<td>-0.102</td>
<td>-0.197</td>
</tr>
<tr>
<td>T24</td>
<td>0.724</td>
<td>0.025</td>
<td>0.252</td>
<td>0.019</td>
<td>0.014</td>
<td>0.007</td>
<td>0.026</td>
</tr>
<tr>
<td>T25</td>
<td>0.375</td>
<td>0.134</td>
<td>0.648</td>
<td>0.029</td>
<td>0.038</td>
<td>-0.084</td>
<td>0.292</td>
</tr>
<tr>
<td>T26</td>
<td>-0.080</td>
<td>-0.044</td>
<td>0.590</td>
<td>0.003</td>
<td>0.055</td>
<td>0.091</td>
<td>0.379</td>
</tr>
<tr>
<td>T27</td>
<td>0.995</td>
<td>0.015</td>
<td>0.072</td>
<td>0.006</td>
<td>0.005</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>T28</td>
<td>0.995</td>
<td>0.015</td>
<td>0.072</td>
<td>0.006</td>
<td>0.005</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>T29</td>
<td>-0.000</td>
<td>0.000</td>
<td>-0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>T30</td>
<td>0.995</td>
<td>0.015</td>
<td>0.072</td>
<td>0.006</td>
<td>0.005</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>T31</td>
<td>0.995</td>
<td>0.015</td>
<td>0.072</td>
<td>0.006</td>
<td>0.005</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>T32</td>
<td>0.995</td>
<td>0.015</td>
<td>0.072</td>
<td>0.006</td>
<td>0.005</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>T33</td>
<td>0.995</td>
<td>0.015</td>
<td>0.072</td>
<td>0.006</td>
<td>0.005</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>T34</td>
<td>0.995</td>
<td>0.015</td>
<td>0.072</td>
<td>0.006</td>
<td>0.005</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>T35</td>
<td>-0.000</td>
<td>0.000</td>
<td>-0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.000</td>
</tr>
<tr>
<td>T36</td>
<td>-0.000</td>
<td>0.000</td>
<td>-0.000</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.000</td>
</tr>
<tr>
<td>T37</td>
<td>0.995</td>
<td>0.015</td>
<td>0.072</td>
<td>0.006</td>
<td>0.005</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>T38</td>
<td>0.995</td>
<td>0.015</td>
<td>0.072</td>
<td>0.006</td>
<td>0.005</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>T39</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>T40</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T50</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Identification indicates whether the parameters of a given model are uniquely determined. Since the model was not identified, it would be theoretically possible to find an infinite number of values for the parameters, each of which would be consistent with the covariance equation. Attempts to estimate this unidentified model would result in arbitrary estimates of the parameters and meaningless interpretation (Long, 1983).

It is apparent from the data in Table 3 that the model in this case could not be identified because:

1. Variables loaded on more than one factor. Variables 3, 6, 7, 15, 17, 18, 22, 24, and 25 showed loadings of over .40 on two factors. In the case of variable 15, loadings were significant on three factors.

2. Loadings of over 0.9 were found on Factor 1, even after rotation.

3. Factors 6 and 7 comprised of less than three variables. That is, less than three variables had loadings of over .40.

Based on procedures recommended by Tatsuoka (1971), it was decided to convert the data into factor scales and test differences between these scales using discriminant analysis. Conversion to factor scales is considered an appropriate approach when minor factors, unspecified and conceptually unrelated to the domain of interest, may
account for some correlations observed, and in turn affect the weights obtained (Kim & Mueller, 1978), or when artificial linear constraints are imposed on the data. The variance-covariance matrix, in this case, becomes singular and cannot be inverted to obtain appropriate regression weights. In the present study, minor Factors 5, 6, and 7 were affecting the correlations in the model. Further, loadings of variables on more than one factor created a linear dependency.

When variables are subject to such an artifical linear constraint, \((p - 1)\) principal components of the \(p\) variables may be used as predictors in lieu of the original data set. Discriminant analysis may then be conducted on the principal components (Tatsuoka, 1971). The factor pattern matrix obtained through Principal Components Analysis is presented in Table 4.

The factor scales derived from the data, through principal components analysis, are presented in Table 5. Factor scales were obtained by multiplying the score on each item with it's factor loading, and summing up the resulting values for all items loading on each principal component.

Overall, significant differences were found between the two groups \((F = 18.31, df = 6, 219, p < .01)\). Wilk’s criterion was used to determine overall significance. In an investigation of group differences for individual
Table 4

Factor Pattern Matrix of Loadings from the American Normative Sample (Principal Components)

<table>
<thead>
<tr>
<th>Item</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
<th>Factor 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>T2</td>
<td>0.070</td>
<td>0.429*</td>
<td>0.476*</td>
<td>0.182</td>
<td>-0.365</td>
<td>-0.016</td>
<td>-0.121</td>
</tr>
<tr>
<td>T3</td>
<td>0.066</td>
<td>0.437*</td>
<td>0.547*</td>
<td>0.348</td>
<td>0.209</td>
<td>-0.173</td>
<td>0.161</td>
</tr>
<tr>
<td>T4</td>
<td>0.059</td>
<td>0.335</td>
<td>0.358</td>
<td>0.235</td>
<td>-0.202</td>
<td>-0.322</td>
<td>0.091</td>
</tr>
<tr>
<td>T5</td>
<td>0.095</td>
<td>0.518*</td>
<td>0.144</td>
<td>-0.377</td>
<td>0.051</td>
<td>0.035</td>
<td></td>
</tr>
<tr>
<td>T6</td>
<td>0.088</td>
<td>0.454*</td>
<td>0.348</td>
<td>-0.001</td>
<td>-0.101</td>
<td>0.367</td>
<td>-0.354</td>
</tr>
<tr>
<td>T7</td>
<td>0.082</td>
<td>0.427*</td>
<td>0.393</td>
<td>0.273</td>
<td>0.495*</td>
<td>0.053</td>
<td>0.071</td>
</tr>
<tr>
<td>T8</td>
<td>0.094</td>
<td>0.361</td>
<td>0.128</td>
<td>0.101</td>
<td>0.594*</td>
<td>-0.203</td>
<td>0.259</td>
</tr>
<tr>
<td>T9</td>
<td>0.127</td>
<td>0.523*</td>
<td>0.216</td>
<td>-0.100</td>
<td>-0.367</td>
<td>-0.135</td>
<td>0.165</td>
</tr>
<tr>
<td>T10</td>
<td>0.098</td>
<td>0.479*</td>
<td>0.338</td>
<td>-0.126</td>
<td>0.314</td>
<td>0.465*</td>
<td>-0.293</td>
</tr>
<tr>
<td>T11</td>
<td>0.136</td>
<td>0.520*</td>
<td>0.114</td>
<td>-0.137</td>
<td>-0.001</td>
<td>-0.132</td>
<td>0.146</td>
</tr>
<tr>
<td>T12</td>
<td>0.158</td>
<td>0.559*</td>
<td>0.047</td>
<td>-0.186</td>
<td>-0.045</td>
<td>-0.047</td>
<td>-0.122</td>
</tr>
<tr>
<td>T13</td>
<td>0.184</td>
<td>0.555*</td>
<td>-0.122</td>
<td>-0.247</td>
<td>-0.005</td>
<td>-0.215</td>
<td>0.113</td>
</tr>
<tr>
<td>T14</td>
<td>0.189</td>
<td>0.536*</td>
<td>-0.074</td>
<td>-0.104</td>
<td>0.012</td>
<td>0.131</td>
<td>-0.023</td>
</tr>
<tr>
<td>T15</td>
<td>0.204</td>
<td>0.612*</td>
<td>-0.131</td>
<td>-0.304</td>
<td>0.116</td>
<td>-0.117</td>
<td>-0.252</td>
</tr>
<tr>
<td>T16</td>
<td>0.035</td>
<td>0.593*</td>
<td>-0.109</td>
<td>-0.240</td>
<td>-0.001</td>
<td>0.181</td>
<td>0.086</td>
</tr>
<tr>
<td>T17</td>
<td>0.264</td>
<td>0.341*</td>
<td>-0.285</td>
<td>-0.085</td>
<td>-0.001</td>
<td>-0.007</td>
<td>0.179</td>
</tr>
<tr>
<td>T18</td>
<td>0.291</td>
<td>0.580*</td>
<td>-0.389</td>
<td>-0.032</td>
<td>0.148</td>
<td>-0.161</td>
<td>-0.176</td>
</tr>
<tr>
<td>T19</td>
<td>0.029</td>
<td>0.419*</td>
<td>-0.102</td>
<td>-0.527*</td>
<td>-0.080</td>
<td>0.020</td>
<td>0.279</td>
</tr>
<tr>
<td>T20</td>
<td>0.326</td>
<td>0.443*</td>
<td>-0.392</td>
<td>0.145</td>
<td>0.056</td>
<td>-0.115</td>
<td>-0.357</td>
</tr>
<tr>
<td>T21</td>
<td>0.404*</td>
<td>0.402*</td>
<td>-0.458*</td>
<td>0.323</td>
<td>-0.081</td>
<td>0.030</td>
<td>0.050</td>
</tr>
<tr>
<td>T22</td>
<td>0.052</td>
<td>0.338</td>
<td>-0.309</td>
<td>0.119</td>
<td>-0.056</td>
<td>0.492*</td>
<td>0.498*</td>
</tr>
<tr>
<td>T23</td>
<td>0.475*</td>
<td>0.247</td>
<td>-0.330</td>
<td>0.148</td>
<td>-0.055</td>
<td>-0.287</td>
<td>-0.221</td>
</tr>
<tr>
<td>T24</td>
<td>0.759*</td>
<td>-0.047</td>
<td>-0.056</td>
<td>0.081</td>
<td>-0.006</td>
<td>0.000</td>
<td>-0.031</td>
</tr>
<tr>
<td>T25</td>
<td>0.535*</td>
<td>0.247</td>
<td>-0.422*</td>
<td>0.357</td>
<td>-0.107</td>
<td>0.047</td>
<td>0.072</td>
</tr>
<tr>
<td>T26</td>
<td>0.067</td>
<td>0.260</td>
<td>-0.378</td>
<td>0.486*</td>
<td>-0.076</td>
<td>0.230</td>
<td>-0.020</td>
</tr>
<tr>
<td>T27</td>
<td>0.972*</td>
<td>-0.198</td>
<td>0.084</td>
<td>-0.043</td>
<td>0.005</td>
<td>0.020</td>
<td>0.021</td>
</tr>
<tr>
<td>T28</td>
<td>0.972*</td>
<td>-0.198</td>
<td>0.084</td>
<td>-0.043</td>
<td>0.005</td>
<td>0.020</td>
<td>0.021</td>
</tr>
<tr>
<td>T29</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>T30</td>
<td>0.972*</td>
<td>-0.198</td>
<td>0.084</td>
<td>-0.043</td>
<td>0.005</td>
<td>0.020</td>
<td>0.021</td>
</tr>
<tr>
<td>T31</td>
<td>0.972*</td>
<td>-0.198</td>
<td>0.084</td>
<td>-0.043</td>
<td>0.005</td>
<td>0.020</td>
<td>0.021</td>
</tr>
<tr>
<td>T32</td>
<td>0.972*</td>
<td>-0.198</td>
<td>0.084</td>
<td>-0.043</td>
<td>0.005</td>
<td>0.020</td>
<td>0.021</td>
</tr>
<tr>
<td>T33</td>
<td>0.972*</td>
<td>-0.198</td>
<td>0.084</td>
<td>-0.043</td>
<td>0.005</td>
<td>0.020</td>
<td>0.021</td>
</tr>
<tr>
<td>T34</td>
<td>0.972*</td>
<td>-0.198</td>
<td>0.084</td>
<td>-0.043</td>
<td>0.005</td>
<td>0.020</td>
<td>0.021</td>
</tr>
<tr>
<td>T35</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>T36</td>
<td>-0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>T37</td>
<td>0.972*</td>
<td>-0.198</td>
<td>0.084</td>
<td>-0.043</td>
<td>0.005</td>
<td>0.020</td>
<td>0.021</td>
</tr>
<tr>
<td>T38</td>
<td>0.972*</td>
<td>-0.198</td>
<td>0.084</td>
<td>-0.043</td>
<td>0.005</td>
<td>0.020</td>
<td>0.021</td>
</tr>
<tr>
<td>T39</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>T40</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>T50</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Note: Asterisks (*) indicate variables comprising Factor Scales.
Table 5

**Factor Scales Derived Through Principal Components Analysis**

<table>
<thead>
<tr>
<th>Scale</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale 1</td>
<td>.4xT21 + .4xT23 + .7xT24 + .5xT25 + .9xT27 + .9xT28 + .9xT30 + .9xT31 + .9xT32 + .9xT33 + .9xT34 + .9xT37 + .9xT38</td>
</tr>
<tr>
<td>Scale 2</td>
<td>.4xT2 + .4xT3 + .5xT5 + .4xT6 + .4xT7 + .5xT9 + .4xT10 + .5xT11 + .5xT12 + .5xT13 + .5xT14 + .6xT15 + .5xT16 + .5xT17 + .5xT18 + .4xT19 + .4xT20 + .4xT21</td>
</tr>
<tr>
<td>Scale 3</td>
<td>.4xT2 + .5xT3 + .5xT5 + .4xT21 + .4xT25</td>
</tr>
<tr>
<td>Scale 4</td>
<td>.5xT19 + .4xT26</td>
</tr>
<tr>
<td>Scale 5</td>
<td>.4xT7 + .5xT8</td>
</tr>
<tr>
<td>Scale 6</td>
<td>.4xT10 + .4xT22</td>
</tr>
</tbody>
</table>

Table 6

**Factor Scale Means and Critical Values of F**

<table>
<thead>
<tr>
<th>Group</th>
<th>Scale 1</th>
<th>Scale 2</th>
<th>Scale 3</th>
<th>Scale 4</th>
<th>Scale 5</th>
<th>Scale 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indian</td>
<td>0.01</td>
<td>2.88</td>
<td>1.21</td>
<td>0.02</td>
<td>0.52</td>
<td>0.21</td>
</tr>
<tr>
<td>American</td>
<td>0.15</td>
<td>5.10</td>
<td>1.41</td>
<td>0.09</td>
<td>0.82</td>
<td>0.36</td>
</tr>
<tr>
<td>F</td>
<td>2.37</td>
<td>98.42*</td>
<td>22.72*</td>
<td>11.95*</td>
<td>64.17*</td>
<td>41.53*</td>
</tr>
<tr>
<td>p</td>
<td>.01</td>
<td>.01</td>
<td>.01</td>
<td>.01</td>
<td>.01</td>
<td>.01</td>
</tr>
</tbody>
</table>

Note: An asterisk (*) indicates significant differences between the group means.
scales (Table 6), it was found that Scale 1 indicated no significant differences between the two groups at the .01. However, significant differences were found for Scales 2, 3, 4, 5, and 6. Hypothesis 1 was accepted.

Hypothesis 2

Significant differences will be found in the item characteristic curves of the individual items for the Indian normally achieving subjects and the TONI normative group.

Hypothesis 2 investigates the shape of the item characteristic curves for Indian normally achieving subjects and the TONI normative group. Individual items may be considered as inappropriate for the Indian group if they show too little ability to discriminate between levels of ability on the latent attribute as compared to discrimination ability in the normative sample. An individual item may be too difficult or too easy for one of the two groups, which would also indicate its inappropriateness for a given group.

1. Discrimination Indices. The average discrimination index, the average point biserial correlation between item scores and total test scores, was computed for the Indian normally achieving group and the TONI normative group. For the Indian subjects the average point biserial score was .2135, and for the American normative group the average was .2918. The difference of
0.079 was not significant at \( p < .01 \), for 324 degrees of freedom. The first part of Hypothesis 2 was rejected.

For further analysis, individual point biserial correlations for each of the fifty test items were examined. Point biserial correlations are presented in Table 7. For the Indian group, any point biserial correlations of less than .30 would indicate that the item should probably be deleted for that group, in order to improve the discrimination ability of the test (Thorndike, 1982). Items 1, 3, and 5 indicated point biserial correlations of less than .30, and should not be included in a test developed for an Indian population, based on information from this sample.

Further, differences between point biserial correlations for the two groups were compared. Significant differences at the .01 level were found for items 4, 8, 10, and 22. Thus, items 4, 8, 10, and 22 should be deleted from the test, if it is to be used with the Indian group, if the intent is to maximize discrimination ability of the test for that group. The Indian group did not score beyond item 23, and therefore point biserials for items 23 onward were, by definition, also significantly different between the two groups.
Table 7

Point Biserial Correlations of Indian Normally-Achieving and American Normative Sample Subjects

<table>
<thead>
<tr>
<th>Item</th>
<th>Indians</th>
<th>Americans</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.187*</td>
<td>.000</td>
<td>.187</td>
</tr>
<tr>
<td>2</td>
<td>.358</td>
<td>.339</td>
<td>.019</td>
</tr>
<tr>
<td>3</td>
<td>.214*</td>
<td>.338</td>
<td>.124</td>
</tr>
<tr>
<td>4</td>
<td>.601</td>
<td>.268</td>
<td>.333+</td>
</tr>
<tr>
<td>5</td>
<td>.268*</td>
<td>.434</td>
<td>.166</td>
</tr>
<tr>
<td>6</td>
<td>.373</td>
<td>.397</td>
<td>.024</td>
</tr>
<tr>
<td>7</td>
<td>.616</td>
<td>.366</td>
<td>.250</td>
</tr>
<tr>
<td>8</td>
<td>.681</td>
<td>.358</td>
<td>.323+</td>
</tr>
<tr>
<td>9</td>
<td>.637</td>
<td>.504</td>
<td>.133</td>
</tr>
<tr>
<td>10</td>
<td>.739</td>
<td>.447</td>
<td>.292+</td>
</tr>
<tr>
<td>11</td>
<td>.764</td>
<td>.517</td>
<td>.247</td>
</tr>
<tr>
<td>12</td>
<td>.753</td>
<td>.566</td>
<td>.187</td>
</tr>
<tr>
<td>13</td>
<td>.683</td>
<td>.594</td>
<td>.089</td>
</tr>
<tr>
<td>14</td>
<td>.464</td>
<td>.575</td>
<td>.111</td>
</tr>
<tr>
<td>15</td>
<td>.460</td>
<td>.652</td>
<td>.192</td>
</tr>
<tr>
<td>16</td>
<td>.510</td>
<td>.549</td>
<td>.039</td>
</tr>
<tr>
<td>17</td>
<td>.384</td>
<td>.621</td>
<td>.237</td>
</tr>
<tr>
<td>18</td>
<td>.507</td>
<td>.649</td>
<td>.142</td>
</tr>
<tr>
<td>19</td>
<td>.449</td>
<td>.410</td>
<td>.039</td>
</tr>
<tr>
<td>20</td>
<td>.404</td>
<td>.557</td>
<td>.153</td>
</tr>
<tr>
<td>21</td>
<td>.312</td>
<td>.546</td>
<td>.234</td>
</tr>
<tr>
<td>22</td>
<td>.000</td>
<td>.317</td>
<td>.317</td>
</tr>
<tr>
<td>23</td>
<td>.312</td>
<td>.458</td>
<td>.146</td>
</tr>
<tr>
<td>24</td>
<td>.000</td>
<td>.353</td>
<td>.353+</td>
</tr>
<tr>
<td>25</td>
<td>.000</td>
<td>.478</td>
<td>.478+</td>
</tr>
<tr>
<td>26</td>
<td>.000</td>
<td>.238</td>
<td>.238</td>
</tr>
<tr>
<td>27</td>
<td>.000</td>
<td>.340</td>
<td>.340+</td>
</tr>
<tr>
<td>28</td>
<td>.000</td>
<td>.340</td>
<td>.340+</td>
</tr>
<tr>
<td>29</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>30</td>
<td>.000</td>
<td>.340</td>
<td>.340+</td>
</tr>
<tr>
<td>31</td>
<td>.000</td>
<td>.340</td>
<td>.340+</td>
</tr>
<tr>
<td>32</td>
<td>.000</td>
<td>.340</td>
<td>.340+</td>
</tr>
<tr>
<td>33</td>
<td>.000</td>
<td>.340</td>
<td>.340+</td>
</tr>
<tr>
<td>34</td>
<td>.000</td>
<td>.340</td>
<td>.340+</td>
</tr>
<tr>
<td>35</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>36</td>
<td>.000</td>
<td>.340</td>
<td>.340+</td>
</tr>
<tr>
<td>37</td>
<td>.000</td>
<td>.340</td>
<td>.340+</td>
</tr>
<tr>
<td>38</td>
<td>.000</td>
<td>.340</td>
<td>.340+</td>
</tr>
<tr>
<td>39</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>40</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>41</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>42</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>43</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>44</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>45</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>46</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>47</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>48</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>49</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>50</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
</tbody>
</table>

Note: An asterisk (*) indicates point biserial correlations below .30. A plus (+) indicates differences significant at the .01 level.
2. **Difficulty Indices.** Simple differences in average item difficulty does not indicate pervasive bias, as it may be reflective of real differences between the performances of the two groups. To compare the difficulty indices for the two groups, procedures outlined by Thorndike (1982) were applied.

The raw percentage of correct answers for each item were converted to a Rasch scale. Steps for conversion are described in Appendix B. Thus the mean scale value of the items was set at zero for each group. Differences in scale values were compared. A .01 significance level would indicate that differences greater than 2.58 are significant. Data in Table 8 presents scale values for the Indian normally achieving sample and the American normative sample. Item 4 is the only one that evidenced a significant difference. Thus, Item 4 would be inappropriate in a test used with both groups. This item, however, represents only .05% of the total items for the test, and the difference may be attributed to error in random sampling. Deletion of this item may have little effect on the overall fairness of the test. The second part of Hypothesis 2 was also rejected, indicating that the scores of the two groups produce item characteristic curves that are not significantly different in shape.
<table>
<thead>
<tr>
<th>Item</th>
<th>Indian Normally-Achieving</th>
<th>American Normative Sample</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-12.890</td>
<td>-10.441</td>
<td>2.450</td>
</tr>
<tr>
<td>2</td>
<td>-11.406</td>
<td>-8.856</td>
<td>2.550</td>
</tr>
<tr>
<td>3</td>
<td>-11.917</td>
<td>-11.022</td>
<td>0.895</td>
</tr>
<tr>
<td>4</td>
<td>-11.406</td>
<td>-8.283</td>
<td>3.123*</td>
</tr>
<tr>
<td>5</td>
<td>-9.886</td>
<td>-8.142</td>
<td>1.744</td>
</tr>
<tr>
<td>6</td>
<td>-10.166</td>
<td>-8.184</td>
<td>1.982</td>
</tr>
<tr>
<td>7</td>
<td>-10.348</td>
<td>-8.051</td>
<td>2.297</td>
</tr>
<tr>
<td>8</td>
<td>-9.339</td>
<td>-7.163</td>
<td>2.174</td>
</tr>
<tr>
<td>9</td>
<td>-8.670</td>
<td>-7.636</td>
<td>1.034</td>
</tr>
<tr>
<td>10</td>
<td>-9.084</td>
<td>-7.345</td>
<td>1.739</td>
</tr>
<tr>
<td>11</td>
<td>-8.390</td>
<td>-7.047</td>
<td>1.343</td>
</tr>
<tr>
<td>12</td>
<td>-7.704</td>
<td>-6.698</td>
<td>1.206</td>
</tr>
<tr>
<td>13</td>
<td>-7.345</td>
<td>-6.391</td>
<td>0.954</td>
</tr>
<tr>
<td>14</td>
<td>-7.114</td>
<td>-5.229</td>
<td>1.885</td>
</tr>
<tr>
<td>15</td>
<td>-7.077</td>
<td>-5.088</td>
<td>1.989</td>
</tr>
<tr>
<td>16</td>
<td>-6.360</td>
<td>-5.229</td>
<td>1.131</td>
</tr>
<tr>
<td>17</td>
<td>-5.849</td>
<td>-5.088</td>
<td>0.761</td>
</tr>
<tr>
<td>18</td>
<td>-5.630</td>
<td>-4.930</td>
<td>0.700</td>
</tr>
<tr>
<td>19</td>
<td>-4.888</td>
<td>-4.482</td>
<td>0.406</td>
</tr>
<tr>
<td>20</td>
<td>-4.949</td>
<td>-4.731</td>
<td>0.218</td>
</tr>
<tr>
<td>21</td>
<td>-4.183</td>
<td>-4.482</td>
<td>0.299</td>
</tr>
<tr>
<td>22</td>
<td>-3.076</td>
<td>-4.133</td>
<td>1.057</td>
</tr>
<tr>
<td>23</td>
<td>-3.283</td>
<td>-3.552</td>
<td>0.269</td>
</tr>
<tr>
<td>24</td>
<td>-1.702</td>
<td>0</td>
<td>1.702</td>
</tr>
<tr>
<td>25</td>
<td>-3.076</td>
<td>-3.552</td>
<td>0.476</td>
</tr>
<tr>
<td>26</td>
<td>-1.702</td>
<td>0</td>
<td>1.702</td>
</tr>
<tr>
<td>27</td>
<td>-0.851</td>
<td>0</td>
<td>0.851</td>
</tr>
<tr>
<td>28</td>
<td>-0.851</td>
<td>0</td>
<td>0.851</td>
</tr>
<tr>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>30</td>
<td>-0.851</td>
<td>0</td>
<td>0.851</td>
</tr>
<tr>
<td>31</td>
<td>-0.851</td>
<td>0</td>
<td>0.851</td>
</tr>
<tr>
<td>32</td>
<td>-0.851</td>
<td>0</td>
<td>0.851</td>
</tr>
<tr>
<td>33</td>
<td>-0.851</td>
<td>0</td>
<td>0.851</td>
</tr>
<tr>
<td>34</td>
<td>-0.851</td>
<td>0</td>
<td>0.851</td>
</tr>
<tr>
<td>35</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>36</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>37</td>
<td>-0.851</td>
<td>0</td>
<td>0.851</td>
</tr>
<tr>
<td>38</td>
<td>-0.851</td>
<td>0</td>
<td>0.851</td>
</tr>
<tr>
<td>39</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>41</td>
<td>0</td>
<td>0</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Note: Signs have been omitted in 'Difference' column. An asterisk (*) indicates significant difference at p < .01.
Item curves representing difficulty and discrimination parameters are presented in Figures 1 and 2, Appendix C. Items 1, 2, 7, and 8 also indicate differences of over two standard scale units, with the Indian group scoring consistently lower. While not statistically significant, the differences are substantial. Deletion of these items from the test for the Indian group may result in an improved difficulty scale.

Hypothesis 3

The sample of normally achieving subjects from India will show significantly higher mean scores on the TONI than the sample of mentally retarded Indian subjects.

Results indicate that significant differences do exist between the two groups ($F = 107.14$, $df = 1,107$, $p < .05$). Mean scores for the normally achieving sample were 82.28 and mean scores for the mentally retarded group were 60.30. Thus a twelve-point difference was evident, which was highly significant. These results support the ability of the TONI to discriminate between normally achieving and mentally retarded children of the same age group, which is supportive of the test's construct validity. Hypothesis 3 was accepted.

Hypothesis 4

The sample of normally achieving third grade subjects from India will show significantly higher mean scores on
the TONI than the sample of second grade normally achieving subjects from India.

This hypothesis is supported by the results of an analysis of variance of mean raw scores of the two groups. Significant differences were found between the mean raw scores of the second and third graders ($F = 13.91$, $df = 1$, 89, $p < .01$). The mean raw scores of second graders was 7.03, and mean raw scores of third graders was 9.87. Thus, a 2.84 point difference was evident.

Further investigation of this hypothesis was conducted by comparing the TONI Quotients for the two groups. TONI Quotients were not significantly different ($F = 1.32$, $df = 1$, 89). This indicates that when scores are converted to standard scores, the quotients derived accurately reflect the students' performance relative to their grade level. The TONI thus demonstrates validity in trans-grade comparisons for these elementary level grades. Hypothesis 4 was accepted on the basis of the foregoing analysis.

Scores of the Indian normally achieving second and third graders on the Indian WISC were used as criterion measures to investigate possible slope or intercept bias between the groups. Thus, predictive validity of the TONI could be established.

**Slope Bias**

Multiple regression analysis yielded a slope (beta)
of 0.048 for the second graders, and -0.152 for the third graders. No significant differences were found between the two groups ($t = -0.105, \text{df} = 89$). Since both sets of scores were identical standard scores, the slope values were equal to the validity coefficients between the two tests. Thus, it may also be said that since the slopes were not significantly different, the TONI does not demonstrate differential validity between the two groups.

The extremely small values of the slope coefficients indicates little or no relationship between performance of these students on the TONI and their performance on the WISC (Figure 3, Appendix C).

**Intercept Bias**

Multiple regression analysis yielded an intercept ($\alpha$) of 110.25 for the second graders and 130.78 for the third graders. The TONI score, thus, has different predictive meaning for the two groups in relationship to the WISC. This situation indicates that the two groups differ in a third variable which correlates positively with both the test and the criterion. In comparisons between children at different grade levels, literacy level may be considered a possible extraneous variable. Age may also have accounted for this effect. Correlation analysis of the WISC scores and the TONI scores with grade revealed significant positive correlations ($\geq (\text{WISC, Grade}) = 0.50$,}
In summary, Hypothesis 1 indicates that factor scales derived from observed variables through principal components analysis differ considerably in their composition for the Indian normally-achieving and American comparison groups. The item characteristic curves, however, do not differ significantly in average discrimination parameters, though discrimination ability of some items are below acceptable levels, or differ substantially between the two groups. Only one item differs significantly in a comparison of the difficulty parameters. The TONI shows ability to significantly differentiate between normally achieving Indian third grade and second grade students. Further, it differentiates consistently between Indian normally achieving and Indian mentally retarded children of the same age level. Implications of these findings are discussed further in the following chapter.
CHAPTER IV

DISCUSSION

"The question of the extent to which basic psychological processes are common to mankind is perhaps the major one being pursued by cross-cultural psychology" (Jahoda, 1980, p. 111). There appears to be no a priori criterion by which to distinguish the culture-specific from the universal (Triandis, 1980). Thus, the investigation of psychological processes in each culture of interest warrants research. The present study adds to the literature on the investigation of the above question by examining the cross-cultural validity of the TONI, a test of abstract problem-solving ability viewed as a component of intelligence.

The purpose of the present study was to determine the validity of the TONI when administered to a sample of Indian normally-achieving and mentally retarded elementary grade children. In addition to contributing to the broad area of cross-cultural test transferability and the search for cognitive universals, this area of research was felt to be of interest to special educators in India, as it would provide information on the feasibility of utilizing a non-verbal test to identify mentally handicapped children.
Implications of Findings

Based on the definition of a scientific model presented by Reynolds and Brown (1984), a hierarchical procedure was used to determine the validity of the TONI. Content validity issues were first addressed by selection of instrumentation that meets criteria defined by Jensen (1980), Salvia and Ysseldyke (1981), Berk (1982), and Reynolds (1982b) for a "culturally reduced" test. However, since such attempts in the past have failed to produce a test that was not culturally loaded in some way, further analysis was conducted.

Factor Scale Differences

Bias exists in regard to construct validity when a test is shown to measure different hypothetical traits (psychological constructs) for one group than another, or to measure the same trait, but with a differing degree of accuracy. (Reynolds, 1982a, p. 194)

Factor analysis has been a commonly used method of assessing the construct validity of testing instruments. Similar factor structures indicate that the trait being measured by the instrument is measured in the same manner for two cultural groups, and is, in fact, the same construct in both groups (Reynolds & Brown, 1984).

In testing Hypothesis 1, overall significant
differences were evident between the factor scales derived from scores of Indian and American normally-achieving children. Statements about the construct validity of the instrument in measuring the underlying hypothetical trait, intelligence, can not be made as the appropriate confirmatory factor analysis could not be conducted.

The difference in factor scales indicates differences in the underlying response pattern of the two groups on the hypothetical trait being measured (intelligence). The scientific value of this difference is made more interpretable by the fact that many "ambient" variables were accounted for. Ambient variables are defined as antecedents that are not considered within the context of the focal theory, but which can reasonably provide alternative explanations of observed differences (Poortinga & Malpass, 1986). In the present study, grade, area of residence, and SES were accounted for by experimental control in subject selection.

The Indian children did not approach the task of abstract problem-solving in the same manner as the American subjects, which reflects a difference in cognitive strategy utilization of subjects. This conclusion can be drawn because rival hypotheses were eliminated by careful control of ambient variables, that is (a) variance in subject characteristics, and (b) variance in testing procedures.
Differences in cognitive strategy application can be attributed to differences in educational training, or a different approach to the task of problem-solving. Further information on this dimension may be obtained by interviewing subjects and determining the reasons for their chosen responses. Such a qualitative approach would determine the strategies applied by subjects to the task, however, this procedure was beyond the scope of the present study. This would be an avenue for future research.

Possible rival hypotheses may have been:

1. Age, literacy, and socioeconomic status. These were eliminated by selection of subjects of the same age group as the normative sample. Further, Indian subjects were currently enrolled in second- and third-grade classes, and were verified by Hollingshead's Index as not belonging to low SES families.

2. Lack of familiarity with intelligence testing. Differences could not be attributed to lack of familiarity, or lack of understanding of the task, because preliminary practice items were administered, and it was established that subjects comprehended the task. Indian subjects were able to correctly answer items at a higher degree of complexity than the ones on which significant differences were evident. Further, subjects were
administered the WISC-R, and scored at an average, or above average, level. Thus, their below average performance on the TONI may not be considered the result of lack of familiarity with the process of standardized testing, in general.

The implications of factor scale differences are considerable. The assumption may not be made that since the TONI is devoid of language and utilizes non-representational symbols, it is 'culture-fair'. This supports the findings of Irvine (1969), Jensen (1968), Ortar (1972), Vernon (1969), Gonzalez (1982), and Court (1982) on other non-verbal measures of abstract problem-solving such as the CCFT, RPM, and the Leiter. It also supports the view proposed by Berk (1982), Reynolds (1982b), and Brown and Reynolds (1984) that attempts to reduce cultural loading of aptitude test items have thus far resulted in failure.

Implications for testing minority children in the US are also evident. The findings support other research (Cole, Gay, Blick, & Sharp, 1971) which states that performance on formal reasoning tasks is dependent on cultural context variables. Since specific test content influences test scores in ways that are unrelated to the ability the test is designed to assess, immigrant and other culturally different children may demonstrate performance that is not reflective of their true ability.
level. Aptitude tests should not be given to such children unless procedures to validate use for each particular minority have been implemented. The cultural neutrality of non-verbal tests may be more apparent than real (Pick, 1980).

**Item Characteristic Curve Differences**

In examining items found to be biased in terms of their discrimination ability and difficulty level, it was found that differing items were predominantly those which required simple matching, or analogous matching strategies.

This finding indicates that there appears to be some fundamental difference in the way young Indian and American literate children approach the task of matching to a given sample. It is not evident from the results whether this difference is due to the way the problem is presented, or whether the task of matching itself is approached differently, with the Indian group looking for a non-matching figure, rather than a matching one. Deletion of items requiring matching skills may provide a test that is more valid for use with Indian children.

It is not possible to infer that Indian children do not possess matching abilities, as evidence of their matching ability exists in that they have successfully acquired language and mathematical skills, both of which
Table 9

Items Evidencing Significant Differences Between Indian and American Subjects

<table>
<thead>
<tr>
<th>Item</th>
<th>Form A</th>
<th>Form B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>simple matching</td>
<td>simple matching</td>
</tr>
<tr>
<td>3</td>
<td>simple matching</td>
<td>matching analogy</td>
</tr>
<tr>
<td>4</td>
<td>simple matching</td>
<td>matching analogy</td>
</tr>
<tr>
<td>5</td>
<td>matching analogy</td>
<td>classification</td>
</tr>
<tr>
<td>8</td>
<td>matching analogy</td>
<td>matching analogy</td>
</tr>
<tr>
<td>10</td>
<td>matching analogy</td>
<td>matching analogy</td>
</tr>
<tr>
<td>22</td>
<td>intersection</td>
<td>classification</td>
</tr>
</tbody>
</table>

depend to some degree on the ability to match. The discovery of difference in a particular skill supports the findings of numerous other studies reviewed by Pick (1980). In the research discussed, differences in cognitive strategies of subjects are evident between two groups on a standardized test. However, both groups demonstrate the same level of ability when tested in a format more closely related to the demands of their cultural environment. Thus, claims of apparent cultural differences in ability to abstract or think in generalities may not be made based on the results of a standardized test alone.

No indication of different approaches to tasks
requiring matching have been reported in the literature on abstract problem-solving ability of Indian children. Further research with a larger sample would be necessary before conclusive statements may be made regarding matching skills of Indian children. No literature was found which examined item characteristic curves of Indian children on non-verbal tests which had been adapted from Western Europe and the US, so supportive statements could not be made.

Reynolds and Brown (1984) state that panels of expert judges in both minority and majority cultures have not, thus far, been able to predict any better than chance which items in any given aptitude test will prove to be biased against one of the cultures. The findings support their contention. There is no existing premise on which any judge may have considered the task of matching to reflect bias in the TONI.

**Ability Level Differences**

In a further examination of construct validity, the relationship between increase in chronological age and raw scores were examined (Reynolds & Brown, 1984), as also the relation between differing levels of intellectual ability and scaled scores.

Hypotheses 3 and 4 examine test validity from the point of view of the test’s ability to discriminate between
groups of differing levels of ability. The TONI distinguished between Indian normally-achieving second and third grade children, and Indian normally-achieving and mentally retarded children, with the lower ability group consistently demonstrating a lower level of performance.

A difference of 22 points was found between the TONI Quotients for the normally-achieving and mentally retarded samples, far greater than one standard deviation (15 points). This demonstrates that the TONI may be a useful tool for identifying mentally retarded children. The difference was less than two standard deviations, however, which is the conventional discrepancy required for a diagnosis of mental retardation in the US (AAMD, 1973). Further study with children at different levels of mental retardation would be necessary before the usefulness of the TONI to determine if children at different levels of mental retardation (mild, moderate, severe) could be identified. If so, utilization of the TONI rather than the WISC-R may be of value in providing a more parsimonious test in terms of time taken to administer, and interpretability of results.

Indian third graders' raw scores were, on average, three points higher than those of second graders. This difference was significant at the .01 level. Further, when scores were converted to TONI Quotients, no significant differences were found, indicating that
standard scores computed for the two groups reflect their different levels of ability. Further study on children at different grade levels would be necessary to determine the discrimination ability of the TONI at various age and grade levels.

Bias in criterion-related validity was examined in the context of the ability of the TONI to predict WISC-R scores of Indian children at two different grade levels.

A test is considered biased with respect to predictive validity when the inference drawn from the test score is not made with the smallest feasible random error, or if there is constant error in an inference or prediction as a function of membership in a particular group. (Reynolds, 1982a, p. 20)

When regressing TONI scores with those on the WISC-R, it was found that for third graders the slope had a minor negative trend. Thus, children demonstrating a higher level of ability on the TONI appear to actually show comparatively lower WISC-R scores. Differences between the slopes and intercepts of the two groups were not statistically significant, however. Thus homogeneity of regression was evident across the two groups (Reynolds & Brown, 1984) indicating fairness in prediction.

Further, normally-achieving Indian children showed, generally, average to above average performance on the
WISC-R, while their performance on the TONI was at below average levels. Possible explanations which may be considered are:

1. While WISC Performance scores indicated above average performance by Indian children (Mean scores = 123.4), when presented with abstract problem-solving tasks, Indian subjects are actually hampered by the non-verbal format.

2. The dimension of intelligence measured by the TONI is not placed in as high priority by persons in India as the dimensions measured by the WISC-R.

3. Since the Indian WISC-R was standardized on a wide cross-section of population, including literate and non-literate subjects, this may have artificially inflated the scores for literate subjects.

If similar negative regression were to be found for children at higher grade levels, this would imply that educational approaches used in India actually deter children from engaging in the kind of mental activity required to answer TONI items. Research with illiterate children would also be of interest, to determine the relationship of the TONI with schooling effects.

Limitations of the Study

Due to difficulties in conducting cross-cultural study, the present research is not as comprehensive as would be desirable. The following limitations are
present:

1. Nature of the sample. In the interest of developing a carefully controlled study, a very select sample was used. This procedure is recommended for cross-cultural study by Brislin et al. (1973). Extensive control, however, limits generalizability of findings (Reynolds & Brown, 1984).

2. Methodology. Confirmatory factor analysis was not applicable in the present situation as the data did not support any theoretical model, and derived matrices were not identified. Thus, statements relating the findings to the theoretical model can only be made with a great deal of caution. The utilization of a multi-method approach addresses this limitation to some degree.

3. Quantitative nature of the study. While significant differences were evident in the approach of Indian subjects to the task of abstract problem-solving, interviews or dynamic assessment was not possible to qualitatively determine reasons for these differences.

Significance for India

Based on the research needs identified in Chapter I, an attempt was made to design the study in such a way that the literature review and findings would be relevant in the context of special education in India. The following needs were addressed:
1. The present study provides a summary of major research on utilization of non-verbal intelligence tests in India. Such synthesis of research has, to date, not been developed for access by educators and psychologists, though leading persons in both fields have recognized the need (eg. Sinha, 1973).

2. In the field of special education in India very little research has been conducted to identify test instruments which have relevance to conditions in India. The assumption has been that instruments derived from western psychological thought are applicable directly in a translated form. Efforts to identify possible cognitive style differences which may bias the test against the culturally different group have been absent. The result has been a large number of instruments measuring many overlapping skills, none of which however, can claim to include only items which reflect unique cognitive styles of a particular group. The conflicting statements regarding the validity of non-verbal instruments, reviewed in Chapter I, are evidence of the inconclusiveness of much of the research to date.

The present study provides evidence that an assumption may not be made that tests developed in a different culture will be directly applicable to Indian conditions simply because they are non-verbal and abstract in format.
3. Many theories postulate that intelligent behavior can be assessed across cultures with instruments that measure problem-solving ability in a context free from specific culturally-bound knowledge. Several existing theories are discussed in Chapter I, in the section 'Theoretical Framework'. This study provides evidence that even in abstract problem-solving tasks certain types of problems are approached very differently by the two cultural groups under study. Items measuring specific types of abstract problem-solving ability must be identified which indicate significant cultural differences and eliminated from tests intended for cross-cultural use. The content of each test item must be equally germane or equally unfamiliar (Frijda & Jahoda, 1966) to all cultures utilizing the test.

4. Many studies conducted on non-verbal intelligence testing in India have indiscriminately mixed literate and illiterate, urban and rural, high and low socioeconomic status children. The results obtained have thus not been clearly interpretable, as several rival hypotheses may have accounted for findings. Validity studies examined in this paper are listed in Table 10. Several of the studies published are merely descriptive. No study has been reported which utilizes multiple methods, or exercises the degree of control necessary for unequivocal statements
about test validity.

Table 10

Validity Studies on Non-Verbal Tests In India - Methodology

<table>
<thead>
<tr>
<th>Author</th>
<th>Method</th>
<th>Type of Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bevli (1982)</td>
<td>ANOVA</td>
<td>High &amp; Low SES</td>
</tr>
<tr>
<td>Chawla (1969)</td>
<td>Correlations</td>
<td>Reliability</td>
</tr>
<tr>
<td>Deb (1966)</td>
<td>Correlations</td>
<td>Reliability</td>
</tr>
<tr>
<td>Desai (1980)</td>
<td>Factor Analysis</td>
<td>Verbal &amp; Non-Verbal Tests</td>
</tr>
<tr>
<td>Ghuman (1975)</td>
<td>ANOVA</td>
<td>British &amp; Indian Students</td>
</tr>
<tr>
<td>Hundal (1965)</td>
<td>Factor Analysis</td>
<td>Verbal &amp; Non-Verbal Tests</td>
</tr>
<tr>
<td>Kakkar (1975)</td>
<td>t tests</td>
<td>Language &amp; Non-Verbal Tests</td>
</tr>
<tr>
<td>Mehrotra (1967)</td>
<td>t tests</td>
<td>Verbal &amp; Non-Verbal Tests</td>
</tr>
<tr>
<td>Mohan (1970)</td>
<td>Correlations</td>
<td>Males &amp; Females</td>
</tr>
<tr>
<td>Mohanty (1980)</td>
<td>ANOVA</td>
<td>High &amp; Low SES</td>
</tr>
<tr>
<td>A. S. Nair (1973)</td>
<td>Factor Analysis</td>
<td>Non-Verbal Tests</td>
</tr>
<tr>
<td>S. Nair (1975)</td>
<td>Factor Analysis</td>
<td>Verbal &amp; Non-Verbal Tests</td>
</tr>
<tr>
<td>Singh &amp; Hundal (1971)</td>
<td>Correlations</td>
<td>Verbal &amp; Non-Verbal Tests</td>
</tr>
<tr>
<td>M. Sinha (1975)</td>
<td>Means, Std. Devs.</td>
<td>Age levels</td>
</tr>
<tr>
<td>R. R. P. Sinha (1980)</td>
<td>t tests</td>
<td>Tribal &amp; Non-Tribal</td>
</tr>
<tr>
<td>Warhadpande &amp; Sethi (1964)</td>
<td>t tests</td>
<td>Language &amp; School</td>
</tr>
</tbody>
</table>

In order to identify an instrument that can consistently discriminate between children of high and low intellectual ability, more controlled research is necessary. The present study provides a workable
methodology for determining possible bias in a measuring instrument used cross-culturally.

**Conclusion**

Despite control on variables which might influence scores, significant differences were found in performance of Indian and American second and third graders on the TONI. One specific item type evidenced significant bias. Thus, the TONI is not transferable for use in India in its present form. Evidence that it reliably discriminates between children at different levels of ability, however, implies, that with appropriate revision, the test may have utility in identifying children of lower intellectual ability, and may be a more parsimonious method than the use of cumbersome intelligence tests such as the WISC-R.
APPENDIX A

TONI SCORE SHEETS
**TONI**

**TEST OF NONVERBAL INTELLIGENCE**

A Language-Free Measure of Cognitive Ability

**FORM A ANSWER SHEET**

Linda Brown, Rita J. Sherbenou, Susan K. Johnson

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
</tr>
</thead>
</table>

| School (Grade) | ( ) |
| Examiners Name & Title | |

| Date Tested | year | month |
| Date of Birth | year | month |

| Age | |

**PROFILE**

<table>
<thead>
<tr>
<th>TONI QUOTIENT</th>
<th>OTHER MEASURES OF INTELLIGENCE, LANGUAGE, OR ACHIEVEMENT</th>
</tr>
</thead>
</table>

| Basal Item (just below first error) | | |
| Number of correct responses between Basal Item and last error | + | |

**SCORE SUMMARY**

| Total Raw Score | | |
| TONI Quotient | | |
| %ile Rank | | |

**ADMINISTRATIVE CONDITIONS**

Was the subject tested ________ individually or ________ in a group (group size ________)?

Was the examiner experienced? ________

Were there any departures from the administration procedures specified in the TONI manual? ________

If yes, explain: ________

Briefly describe the physical conditions under which the TONI was administered (presence of other people, noise, distraction): ________

Briefly describe the condition of the subject when the TONI was administered (understanding of the task, energy level, attention, cooperation): ________

**OTHER RELEVANT TEST DATA**

---

© Copyright 1982 Linda Brown, Rita J. Sherbenou, Susan K. Johnson

Reprinted with permission.
TONI
TEST OF NONVERBAL INTELLIGENCE
A Language-Free Measure of Cognitive Ability
FORM A ANSWER SHEET
Linda Brown, Rita J. Shergold, Susan K. Johnson

INSTRUCTIONS
Administer the training items. Begin testing with item 1 only if the subject is very young or is suspected of having significant intellectual impairment. In all other cases begin at a level where you believe the subject will be successful. If the subject does not make five correct responses in a row, return to the first item administered and move to successively easier (lower numbered) items until a basal of five consecutive correct responses is achieved. Resume testing at the item following the most difficult (highest numbered) item previously administered and continue until the subject has missed three out of five items. Discontinue testing at this point.
Rasch Scale Conversion Procedure

1. For each item on the Test of Non-Verbal Intelligence, the ratio of persons getting the item wrong to persons getting it right (W/R) was calculated for the Indian normally achieving group and the American normative sample.

2. The natural log of each W/R was calculated.

3. The average log value and the deviation log values for each group were calculated.

4. The variance of log values for each group was calculated.

5. An expansion factor was calculated to compensate for variability in person ability or item difficulty.

The expansion factor for items is:

\[ Y = \frac{1 + V/2.89}{1 - UV/8.35} \]

where: 
- \( V \) = variance in log values of scores
- \( U \) = variance in log values of items

6. The deviation log value of each item was multiplied by the expansion factor, \( Y \), to get an estimated scale value of the item's difficulty.
Figure 1. Item Discrimination Curve.
Figure 2. Item Difficulty Curve.
Figure 3. Regression of TONI scores on the WISC-R.
REFERENCES


Conclave of Social Dynamics.


Plenum.


Ironson, G. H. & Subkoviak, M. J. (1979). A comparison of


Lindquist, E. F. & Hieronymus, A. N. (1956). *Iowa Test of*


Nair, A. S. (1968). *The Kerala University Nonverbal Group Test of Intelligence*. Trivandrum, India: Kerala University Department of Education.


Intelligence, 3, 217-227.


Rath, R. (1954). Correlation of Progressive Matrices Test with verbal intelligence and written tests on some other subjects. *Education and Psychology, 1*, 4-5.


Cultural Psychology, 1, 131-138.


Publishers.


of Cross-Cultural Psychology, 6, 4-87.
