A COMPARISON OF PRESCHOOLERS' PERFORMANCE ON CONVENTIONAL AND COMPUTERIZED VERSIONS OF THE PPVT-R

BY

ANDREW T. ROBSON

A thesis

Submitted to the Faculty of Graduate Studies

in Partial Fulfillment of the Requirements

for the degree of

MASTER OF SCIENCE

Department of Family Studies

University of Manitoba

Winnipeg, Manitoba

(c) September, 1997

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-23478-9

Canada

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION PAGE

A COMPARISON OF PRESCHOOLERS' PERFORMANCE ON CONVENTIONAL AND COMPUTERIZED VERSIONS OF THE PPVT-R

BY

ANDREW T. ROBSON

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Andrew T. Robson 1997 (c)

Permission has been granted to the Library of The University of Manitoba to lend or sell copies of this thesis/practicum, to the National Library of Canada to microfilm this thesis and to lend or sell copies of the film, and to Dissertations Abstracts International to publish an abstract of this thesis/practicum.

The author reserves other publication rights, and neither this thesis/practicum nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

I hereby declare that I am the sole author of this thesis.

I authorise the University of Manitoba to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Andrew T. Robson

I further authorise the University of Manitoba to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Andrew T. Robson

The University of Manitoba requires the signature of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

The validity and practicality of a computerized version of the Peabody Picture Vocabulary Test-Revised (PPVT-R) was investigated. Specific guidelines for developing computerized versions of conventional tests were formulated and followed in the creation of a version of the PPVT-R that runs on a Macintosh computer. The conventional and computerized versions of the test were administered to 53 preschoolers. A within-subjects design was used and order of administration was counterbalanced. Pearson product-moment correlations were calculated to assess the concurrent validity of the computerized PPVT-R. Significant positive correlations were found between the two versions of the tests in the entire sample (r = .88), as well as within gender and age subgroups (r = .88 and .82 respectively). Results of t-tests revealed no significant difference between scores on the two versions of the test among four-year-olds, but three-year-olds performed at a significantly higher level on the conventional version. Possible explanations for this age difference are put forth and the utility of computerized testing is discussed.

Acknowledgments

Working on this thesis has been one of the most rewarding experiences of my life. The long hours, hard work, and dedication were at times difficult and I would like to thank the people who provided direction, advice and support throughout the process. I would like to thank Dr. Beverly Zakaluk my external committee member for her constructive feedback and suggestions. I would also like to thank Dr. Dale Berg for providing general advice and support, and for his insights and comments about computer technology.

Most importantly I would like to acknowledge my indebtedness to Dr. Joan Durrant, my mentor for academia and for life. From the initial idea to the defense, Dr. Durrant provided the sincerest encouragement and support. The many hours she spent reviewing and editing my writing deserves my highest appreciation and recognition. Thank you Joan for your expertise and your friendship.

Finally, I would like to thank my family and friends for there encouragement and support. Specifically I would like to extend profound gratitude to my wife Alma and my daughter Elise for letting me take the time to pursue a dream. Thank you all.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER I

Introduction

The practice of testing individuals for specific abilities began around 1100 BC when the Chinese began to use formal tests to choose civil servants (Dubois, 1970). Since this time, wisemen, doctors, educators and psychologists have searched for ways to improve the efficiency and accuracy of their tests. Originally, tests were presented orally, and then in written form, and were administered individually by trained professionals who recorded the results with pencil and paper. By the early 20th century, however, scientists began to look at mechanical machines as a way to improve the efficiency and accuracy of presenting and scoring tests (Pressey, 1926). By the 1960s, assessment professionals began investigating the computer as a means of improving the accuracy and efficiency of testing (Pearson, Swensen, Rome, Mataya, & Brannick, 1964; Finney, 1966). In the last 15 years, computers have become widely used in the field of educational and clinical testing (Maddux & Johnson, 1993). Meire and Geiger (1986) note that, "human services professionals are witnessing an unprecedented growth in the automation of instruments for psychological and career assessment" (p. 29). Predictions have also been made that by the next century all testing will be done via computers (Johnson, 1979).

In recent years, however, some assessment professionals have questioned the wide-spread acceptance of computer-based tests (Wise & Plake, 1989). Maddux (1984), for example, states that the computer is being used simply because it is available; Ebery and Cech (1986) point to a lack of critical

research with computer implementation. Following their review of computer testing, Maddux & Johnson (1993) concluded that computerization is inappropriate in many cases and that only certain tests or testing tasks benefit from being computerized. Continued research is needed in this area to determine which components of testing can be improved with the use of a computer and for which components computerization is inappropriate. There is also growing evidence that some computerized versions of conventional tests may not be equivalent due to the use of nonstandardized equipment (Maddux & Johnson, 1993; Moe & Johnson, 1988; Watkins & Kush, 1988). Which may compromise the validity of computerized tests when the normative, reliability and validity data of their conventional counterpart is assumed to be generalizable to the computer format. The purpose of the present study is to assess the concurrent validity of a computerized conventional test among a sample of preschool children, using conservative criteria for its selection, development, and administration.

In the following section, the history of the use of computers in educational and clinical assessment will be reviewed. This review will provide a framework to demonstrate how the historical use of the computer in testing has led to its current use and misuse today.

A Brief History of the Use of Computers in Testing

The notion of using machines to administer tests was first conceived early in this century. The first attempts were delivered through slow cumbersome mechanical devices. Today, the devices are quick, ergonomic and electronic.

Advances in technology through the 20th century have prompted this transition.

The 1920s to 1950s: The teaching machine era. Attempts to computerize assessment tools began as early as the 1920s when Sidney Pressey (1926,1927) designed a machine to help teachers with the routine tasks of administering and scoring objective tests. The machine administered multiple-choice items to students by presenting a question followed by several alternative answers in slots on a machine. Although the machine tests were practical, such assessment formats never became popular. It was not until the 1950s that a resurgence in research on teaching and testing machines occurred, heavily influenced by the work of B.F. Skinner (1954, 1958) who developed machines that presented students with multiple-choice as well as essay-type questions. These machines permitted individual self-administration and automatic scoring. However, like Pressey's machine, Skinner's technology never gained popularity. The mechanical age faded as old technologies were replaced by electronics and, by 1960, electronics were used in the development of assessment tools.

The early 1960s: The mainframe era. By the 1960s, many universities had acquired large mainframe computers. It was at the University of Illinois that one of the first programs for use in education and assessment was developed. The system was called Programmed Logic for Automatic Teaching Operations (PLATO) and is reputed to be the largest computer-based education and testing system ever developed (Burke, 1982). The system was initially very specialized with one PLATO terminal connected to a mainframe computer. The system quickly evolved, however, to include a great number of PLATO terminals

connected to a variety of mainframes on various college campuses. A variety of educational and testing programs were developed, including topics such as fourth-grade mathematics, vocational aptitude testing, and university chemistry. In all, more than one thousand PLATO programs were developed (Maddux & Johnson, 1993). The project became widely known and documented but the system was very large, costly, and unreliable. This technology was eventually deemed to be impractical, particularly given the small number of educational users.

The late 1960s and early 1970s: The mini-computer era. Between the mid 1960s and mid 1970s computer engineers were continually developing smaller computers. When they reached the size of refrigerators, computers became less costly to produce and purchase, more manageable and more widely available. For the first time, researchers had a tool that could be used to develop, administer, score and interpret psychological tests with a sufficiently large population of end users to make it viable.

One of the first documented practical uses of this type of system was carried out within the Mayo Clinical Program (Pearson, Swenson, Rome, Mataya & Brannick, 1964); a computer system was developed for the scoring and interpretation of the Minnesota Multiphasic Personality Inventory (MMPI). Shortly thereafter, Finney (1966) computerized the California Psychological Inventory and, by 1968, Kleinmuntz and McLean (1968) had developed a computerized MMPI that not only scored and interpreted, but also administered the MMPI in an intuitive manner by branching through the item base according to the user's

responses to previous items. A number of additional assessment tools were computerized during the late 1960s and early 1970s, including projective measures, such as the Rorschach (Piotrowski, 1964) and the Holzman Ink Blot Test (Gorham, 1967); intellectual measures such as the Wechsler Adult Intelligence Scale (WAIS: Elwood & Griffin, 1972) and the Raven Progressive Matrices (Paitich, 1973); and psychiatric measures, such as the Psychiatric Evaluation Form (Herz, Endicott, Spitzer & Mesnikoff, 1971) and the Current and Past Psychopathology Scales (Endicott & Spitzes, 1972). A thorough review of the wide variety of computer assessment tools developed during this era can be found in Bunderson, Inouye & Olsen (1989).

While today computer users interact with computers through a fairly standardized set of keyboards, mice, monitors and printers, in the mini-computer era this was not the case. Rather, a wide variety of "user stations" were created that used a variety of input and output devices ranging from push buttons to slide projections to microphones. For example, in an early attempt to computerize the Peabody Picture Vocabulary Test (PPVT: Dunn, 1959), Knight, Richardo & McNarry (1973) used a special-purpose student terminal developed by the Radio and Electrical Engineering Division of the National Research Council of Canada (Brahan & Brown, 1972). The special "user station" included a special self-contained slide projection device, touch sensitive tablets, a random access audio disc unit, a typewriter-like keyboard printer unit, loudspeakers, and a computer unit for overall control. A similar project was carried out by Overton and Scott (1972). They used a variety of hardware such as slide projectors, touch response

panels and a main computer in automating the PPVT. Both of these research teams found that correlations between scores on the computer and pencil and paper versions of the PPVT fell within the range of .90 to .95. These efforts represent the first attempts to computerize the PPVT. More generally, they demonstrated the utility of the mini-computer and other technologies for the development of viable and practical assessment instruments. The hardware, however, was elaborate and expensive. In addition, each "user station" was unique and a variety of different types of computers and software were in use. This limited the development and wide-spread use of standardized computerized tests.

The late 1970s and early 1980s: The microcomputer era. In the late 1970s, the computer industry began to standardize computer peripherals with the development of the smaller and faster microcomputer. In 1978, the Apple Corporation marketed one of the first microcomputers that was intended for use by the general public. This computer had standardized input and output peripherals, was desktop size, and had a television-like computer monitor, a keyboard, tape drives, and a printer.

Researchers quickly realized that the lower price of and greater similarity among the new microcomputers made them a potential platform for the development of assessment measures that could be administered in a standardized way on a large number of computers around the world. The practicality and viability of computerized assessment now appeared to be a reality. Beaumont (1981), for example, used an Apple II+ microcomputer in the

development of psychological assessments. The instruments used the standard keyboard instead of specialized response panels. Beaumont realized that, in order to provide a common standardized computer test that could be widely used, one must use a computer with standardized input and output devices. The introduction of the Apple II+ with its common input and output devices provided this standardization.

The late 1980s and early 1990s: Rapid test development. During the 1980s, researchers developed a wide variety of computer assessment instruments for the standard microcomputer, such as the Wechsler Adult Intelligence Scale-Revised (Martin & Wilcox, 1989), the Ravens Colored Progressive Matrices (Collins & Odell, 1986), The Harris-Goodenough Draw-A-Man Test (Levy & Barowsky, 1986), the Matching Familiar Figures Test (Van Merrienboer, Jeroen, & Jelsma, 1989), and the Gollin Incomplete Figures Test (Foreman & Hemmings, 1987). Professionals rationalized the development of these computerized tests by citing a variety of advantages that can be grouped into four categories: (a) enhanced motivation of test-takers, (b) increased accessibility, (c) increased accuracy and efficiency, and (d) improved standardization (Eberly & Cech, 1986; Madsen, 1986). Each of these advantages will be discussed in the following section.

Advantages of Computerized Testing

Researchers have explored the advantages of computer testing since the 1920s. A majority of researchers have found that the computer's primary advantage is the elimination of errors in the calculation of test scores (Maddux &

Johnson, 1993). Although scoring advantages are the most cited, the modern microcomputer provides other significant benefits.

Enhanced motivation. One of the main challenges that confronts educators and clinicians is the need to provide assessments that are motivating to the test-taker. Educators discovered in the 1980s that computers seemed to increase students' achievement motivation in the classroom (Seymour, Sullivan, Story, & Mosley, 1987). Since that time, researchers have been trying to uncover the features of computerized interaction that lead to increased levels of achievement.

Most studies of motivation have focused on three questions: (a) What initiates interest and participation? (b) What causes an individual to persevere at a task? and (c) What causes an individual to strive for a goal? (McMillian & Forsyth, 1991). In their attempts to answer these questions, researchers have developed a variety of methodologies and theories that have contributed to an understanding of the relationship between motivation and performance and have targeted three primary areas of inquiry.

The first area concerns mastery motivation, achievement motivation, challenges, and competence seeking (Harter, 1981; Kagan, 1972; McClelland, 1965; White, 1959). This research has focused on the intrinsic human need to overcome or master the environment. One's ability to meet this need is facilitated when problems presented are of moderate difficulty (Leeper, 1985). This area of investigation has yielded findings supporting the hierarchical organization of student or client goals to provide an optimal level of difficulty.

The second major area of motivational research concerns the importance of curiosity, incongruity, discrepancy, and complexity (Berlyne, 1966; Hunt, 1965). These studies have demonstrated that humans value experiences that provide a moderate level of surprise, or cognitive disequilibrium, which is typically caused by factors such as novelty, variability, figurality and problem solving and provide an opportunity to enhance self-efficacy (Leeper, 1985).

The third group of motivational studies has investigated locus of control, or perceived control, and self determination (Condry, 1977; deCharmes, 1968). These studies have been based on the assumption that humans have a basic need to believe that they can exert control over their environments. Perceived control is enhanced by high levels of choice and by a responsive or reinforcing environment.

On the basis of the demonstrated importance of gearing task difficulty to individual skill levels, provoking cognitive disequilibrium (i.e., a disparity between previous knowledge and new information), and providing a sense of control, as well as the features of computers, researchers have begun to study the utility of computers in optimizing children's motivational levels (Butzin, 1990; Leeper, 1985; McClendon, 1989; Seymour, Sullivan, Story & Mosley, 1987). It has been reasoned that software packages can provide the user with hierarchically arranged levels of difficulty to optimize achievement motivation and can provide novelty, variability and, for many first time users, a sense of uncertainty (Malone, 1981). It has also been suggested that the user's sense of control may be enhanced through the array of choices offered as well as the high level of

response and reinforcement provided by many programs (Malone, 1981).

This theoretical connection between motivation and computerized administration of tasks is being confirmed by empirical research. A number of studies have demonstrated that students of all ages prefer using computers over other media (McClendon, 1989; Saccardi, 1991; Seymour et al., 1987). For example, Seymour et al. (1987) found that when 69 fifth-and-sixth-graders were given a choice to return to a task that could be performed on a computer or by paper and pencil, 97% chose the computer. When interviewed, these students stated that they found the computerized presentation to be more interesting and easier than the conventional format.

Butzin (1990) found that early elementary school children who could choose from different learning stations situated around the classroom preferred the computer station to other stations, such as the book station. These children identified the computer station as their best-liked station 206 times, but their least liked station only 9 times. Their second choice was the book station, which was liked best 71 times, and liked least 42 times.

Strong student preferences for instruction via computers were also noted among older students in a study by Kinzie, Sullivan & Berdel (1992). In this study, students were given a choice of different curriculum areas, some of which were taught on a computer, and others by a teacher. Regardless of the curriculum area that was offered, computer instruction was preferred consistently over teacher instruction. For example, when a science curriculum was offered on the computer, there was a 44% increase in participation over the level seen

when it was offered by a teacher.

Other studies have demonstrated that persistence is increased through computerized instruction. For example, McClendon (1989) found that grade one students stayed on task longer in a spelling exercise when a computer rather than a paper and pencil task was used. Another recent study showed that quantity of reading increased when a computer reading program was introduced to junior high school students (Saccardi, 1991).

The empirical studies described above support the hypothesis that motivation to perform a task may be enhanced by the use of a computer. However, it should be emphasized that the personal computer was still relatively novel during the time period in which these early studies were conducted. Although computers will always be capable of gearing task difficulty to individual skill levels and providing a sense of control to the user, the component of motivation that is affected by curiosity and incongruity factors may wane in the future as computers become more common place in everyday human activity.

In any case, the findings of these studies indicate that computers appear to have the potential to increase children's motivation to achieve. However, the motivation enhancing capabilities of the computer are not likely restricted to an educational setting. The increased levels of intrinsic and extrinsic motivation that are fostered by the computer may also be seen in clinical settings, particularly those involving assessment. As in the case of educational software, assessment software can provide novelty, which may enhance interest and attention as may the examinee's ability to control the testing situation. Assessment software can

present material to users in a hierarchical fashion that will provide an optimal level of difficulty. In fact, computerized assessment tools have the potential to analyze each response and immediately provide an item of optimal difficulty.

Increased accessibility. Psychological testing of clients with special needs is often difficult due to the limited response repertoire of some physically challenged individuals (Wilson, Thompson, & Wylie, 1982). A large number of psychological assessment instruments require the examinee to respond to test questions verbally, in writing, or by some other physical means. As a result, some physically handicapped individuals may be deemed untestable or, worse, cognitively impaired (Wilson, Thompson, & Wylie, 1982). Some assessment professionals have attempted to modify tests to accommodate specialized responses but these, at best, have been makeshift (Maguire, Knobel, Knobel, Sedlacek, & Piersel, 1991). Application of the standardized norms to the modified test may also be inappropriate.

One of the most common types of modification has been the development of computerized adaptive devices. The keyboard, for example, has been adapted in a variety of ways, such as the use of key guards that expose only particular keys. Other options include enlarged keyboards with enlarged keys or the use of a stick-like pointer device that can be held, attached to the hand or to a headband, or held in the mouth (Hagan, 1984). Another common device is the microswitch, which provides binary input to a computer when it is activated by a stick, a foot pedal, an eye blink, a breath, or any part of the body that has muscle control. Touch-sensitive screens allow users to interact with the computer by

touching the monitor screen in software specific areas, eliminating the use of a keyboard or a pointing device, such as a mouse. Speech synthesizers allow the computer to enunciate speech in a human-like voice through a speaker to enable any text or commands on the screen to be heard by the user. Through voice recognition, the user can interact with and control the computer through voice commands transmitted through a microphone.

Although some of these adaptive devices are being used with great success in education and special education classes (Hagan,1984), there are few examples of their use in assessment. The few studies that have been conducted, however, have yielded promising findings. For example, Wilson and her colleagues (1982) adapted three different tests for computer use with physically handicapped individuals. They discovered that scores on the computer tests and the conventional tests were found to be positively and significantly correlated ($r = .77$ to .91).

The PPVT also has been adapted for physically and cognitively challenged individuals. Knights, Richardson, & McNarry (1973) and Overton and Scott (1972) presented the test on an automated visual display apparatus where subjects responded by pressing a large panel button. In both cases, the researchers found no significant differences between scores on the standard and automated versions of the test. In a more recent study Maguire, Knobel, Knoble, Sedlacek, & Peirsel (1991) adapted a microcomputer to administer the Peabody Picture Vocabulary Test-Revised (PPVT-R: Dunn & Dunn, 1981). The investigators concluded that the standard and adapted versions correlated to a

degree that was "positive, substantial, and acceptable for clinical use" (r = .91) (Maguire, Knobel, Knoble, Sedlacek, & Peirsel, 1991: p. 199).

These studies highlight the potential of computerized adaptive devices to increase the accessibility of standard assessment tools. However, while promising, the results of these studies are based on older technologies. Modern computers, capable of higher resolution graphics, video, three-dimensional renderings, better audio presentations and a wider variety of adaptive devices have an even greater potential for use with special populations.

Increased accuracy and efficiency. The reliability and validity of any psychological measurement can be greatly affected by the accuracy of the professional administering the test, who may make administration or scoring errors by determining basals and ceilings incorrectly, missing or assigning incorrect point values to examinee responses, or miscalculating sub-scores or final scores. Examiners may also spend valuable time inefficiently calculating these scores. Whitten, Slate, Jones, Shine, and Raggio (1994), examined 57 administrations of the Wechsler Preschool and Primary Scales of Intelligence (WPPSI-R) and discovered a total of 4,177 errors in administration and scoring.

The errors that examiners make are often related to time constraints due to their excessive caseloads (Miller, Witt, & Finley, 1981; Reiner & Hartshorne, 1982; Slate & Hunnicutt, 1988; Whitten, Slate, Jones, Shine, & Raggio, 1994). Reschly and Grimes (1990) examined the errors made by assessment professionals in using intelligence tests and discovered that many of these errors shortened the time needed to administer and score the test. Similarly, Whitten et

al. (1994) found that most examiner errors were due to carelessness resulting from attempts to save time.

In the late 1960s and early 1970s, assessment professionals began to examine the practicality and viability of automated testing as a time saving measure and as a method to minimize examiner error (Elwood, 1972; Knights, Richardson & McNarry, 1973; Overton & Scott,1972). These early researchers discovered that computers could make the assessment process more efficient. Since that time, researchers have refined this initial research to understand the use of computers better in providing increased accuracy and efficiency (Elwood, 1969; Maddux & Johnson, 1993; Overton & Scott, 1972; Schuerholz, 1984-1985; Weizenbaum, 1976; Wise & Plake, 1989). Researchers have computerized assessment measures in an effort to increase examiner accuracy and efficiency in: (a) administration, (b) scoring and arithmetic manipulation or transformation of test or sub-test scores, (c) interpretation of test results, and (d) production of test or assessment reports.

As early as the 1960s, a variety of assessment instruments were automated, such as the MMPI (Finney, 1966), the WAIS (Elwood, 1969), and the Rorschach (Piotrowski, 1964), but the time saved in the administration of the tests was minimal. These early computerized assessments are now classified as computer-based tests (CB). Computer-based testing refers to conventional linear paper and pencil based tests that have merely been adapted for presentation on a computer. The administration of these tests in a linear fashion is often not any faster on a computer than it would be in conventional form. Computerized

Adaptive tests (CA), on the other hand, are presented in a format whereby the administration of an item is determined by the examinee's previous responses. This type of computerized testing can save significant time in administration, as the calculations necessary to determine the sequence of presentation of test items are made instantly.

Calculating a subject's chronological age, determining basals and ceilings, summing categorical responses, calculating raw scores, transferring raw scores to standard scores, and plotting scores on a graph are error prone, time-consuming tasks for examiners. The increase in accuracy and efficiency a computer can provide in completing these tasks is substantial. In fact, these are the most commonly cited advantages of using computerized tests (Maddux & Johnson, 1993).

Interpretation of test results is one of the fastest growing areas in the use of computers to improve efficiency (Maddux & Johnson, 1993). For example, there are more than a dozen commercial programs for interpreting the Wechsler Intelligence Scale for Children-Revised (WISC-R: Kramer, 1988). Although computers are widely used to interpret tests and are reported to be substantial time savers, the computer interpretations should only be used to supplement an examiners interpretation. Otherwise, results may be too general or ambiguous.

Improved standardization. Standardized administration is one of the defining characteristics of any psychological test (Walsh & Betz, 1995) and is crucial to ensuring the validity of the test results. For example, for each administration, instructions provided to examinees should not vary, sample

questions should be identical, and time limits should not be truncated or extended.

A brief review of the literature, however, indicates that standardized test administration procedures may often be compromised. Standardized school testing has been reported to be affected by teacher attitudes toward testing. Teachers have been found to teach to the test, coach during the test, provide inaccurate timing and alter answer sheets (Monsaas & Engelhard, 1990). In the clinical field, problems with standardized administration have included errors in selecting appropriate sub-tests, inaccuracy in timing, failure to read directions verbatim, and inappropriate manipulation of test materials (Choi & Proctor, 1994).

Examiner attitudes and biases have also been documented as a problem in test standardization in clinical testing. For example, Lasky, Felice, Moyer, Buddington and Elliot (1973) gave examiners inflated or true reports on examinees' previous PPVT scores and then asked the examiners to administer the PPVT a second time to the same examinees. The findings demonstrated that there were significant examiner effects.

Examinee bias has been demonstrated to be affected by the format of the test. For example, Johnson and Mihal (1973) in an investigation of the differences between human and computer testing on black and white children, found no differences between results on the two types of administration among white children, but found that black children performed better on the computer version. They hypothesized that the performance of the black children was

enhanced by the computer version because the use of the computer reduced anxiety which is often induced when examiners represent a more advantaged background. Evan and Miller (1969) also have found that subjects responded to a computer administration of a questionnaire with greater honesty and candor than they did to a human paper and pencil presentation. These two studies emphasize the fact that examinee bias can be reduced in a computer testing situation in which there is no human interaction present. However, this advantage would need to be weighed against the inherent disadvantages of eliminating human presence in the testing situation.

While standardization of test administration may never be absolute, it can be improved with the use of a computer. For example, the same sample questions could be administered each time the test is given. Sub-tests could be automatically selected and the correct items presented in the correct order. The test instructions could be digitally recorded and presented in an identical way to all examinees, with a standardized voice, inflection, and rhythm. Basals and ceilings could be calculated instantaneously and the order of item presentation could be determined automatically so that the timing of item presentation could be precisely controlled, as could the recording of response latencies in examinees' answers to questions.

Summary. Computerized testing has been shown to posses important advantages over standard testing formats. The primary advantages are increased motivation, accessibility, accuracy and efficiency, and standardized administration. These advantages provide strong justification for the

development of computerized tests. However, it is important that the psychometric properties of such tests are thoroughly assessed to ensure that they provide useful forms of measurement. It is also important that the process of computerization is carried out in a rational and structured manner in terms of test selection, program design, administration, scoring, and interpretation. In the following section, a critique of computerized testing is presented. Then a set of criteria is put forth to guide this process and to provide a foundation for the approach taken to computerized testing in the present study.

Critiques of Computerized Testing

An important weakness of much computerized test development was recognized during the 1980s; little research was conducted on the validity or reliability of the measures that were being created. Maddux (1984) characterized the situation as the "Everest Syndrome;" researchers were implementing computer testing simply because computers were there. Eberly and Cech (1986) stated that "computer technology has been almost uncritically integrated into the counseling process" (p.24).

Recognizing both the advantages and the dangers of computerized assessments, the American Psychological Association has published Guidelines for Computer Based Tests and Interpretations. (APA,1986). The guidelines were designed to address the "rapid increase in the availability and use of these applications of computer technology" (p. 5), and to "assist professionals in applying computer-based assessment competently and in the best interests of their clients" (p. 5). The specific purpose of the Guidelines was to help

developers and administrators interpret the Standards for Educational and Psychological Testing (APA, 1985) as they relate to computer-based testing and test interpretation.

Nine of the guidelines are directed to test administrators and cover areas of administration and interpretation. Administrators are warned about using faulty computer or adaptive equipment and, if non-standard equipment is used, the need for appropriate calibration. Test takers should be properly trained in the use of computer equipment and the environment should be appropriate for all populations. Finally, test-takers should be monitored and assisted if needed, and professional judgment should be used with computer-generated interpretive reports.

Guidelines for test developers covered such human factors as using appropriate response devices and giving examinees feedback and information on performance. Another recommendation stipulates that the testing procedures developed should permit replication and also provide for confidentiality. Guidelines for assessing the psychometric properties of computerized tests included methods for determining the equivalency of the computer version to the conventional test, and appropriate ways for establishing and documenting validity and reliability. The remaining guidelines relate to the validation of computer generated test interpretations.

In response to these guidelines, a number of researchers began to evaluate computerized assessment instruments. Their critiques have focused on five major issues. First, many tests do not lend themselves to computerization

(Schuerholz, 1984-1985), such as those that require examinees to: manipulate objects or materials, read material to the examiner, or respond in ways that are not compatible with current computer technology. Developers of computerized versions of tests were warned that alteration of response modes to suit the features of a computer may be inappropriate.

Second, computerized tests may not allow the examinee to skip items, review items, or change responses (Ronau & Battista, 1988), as would be allowed on a conventionally administered test. There is evidence that this limitation can lower scores (Wise & Plake, 1989).

Third, computer tests may not be administered in a completely standardized way due to differences among individual computers (Madsen, 1986). Some researchers have voiced concern about the use of computers to standardize test presentations because examiners may use different computers with different background colors, resolutions, fonts, audio capacities, and/or speeds (Madsen, 1986).

Fourth, there is evidence that computerized instruments may not be equivalent to their convention versions (Maddux & Johnson, 1993). Watkins and Kush (1988) have suggested that equivalency is confounded when examinees must deal with keyboards or mice instead of paper and pencils, and in that case, using normative data from conventional versions is inappropriate.

Fifth, interpretation of test results is still highly suspect (Maddux & Johnson, 1993). Interpretation of test results can be, at least partially, a subjective process in which the examiner relies on clinical judgment or

professional experience. At the present time, computers are not able to emulate human intelligence and affective qualities sufficiently to provide useful interpretations. To date, programs have been developed to deal with structured problems for which the answer can be determined by linear, step-by-step, and convergent methods (Dreyfus and Dreyfus 1988). However, for the interpretation of unstructured tests, computerized linear sequential methods are ineffective. In the past decade, computer researchers have attempted to develop computer programs with near-human qualities, known as expert systems or, more globally, as artificial intelligence. Some of these programs have been successful in interpreting responses to structured problems, but to date there are no programs that can emulate the complex human ability to deal with unstructured problems. Kramer (1988) points out that a common problem with computer interpretations is what is known as the "Barnum effect"; computer interpretations are often very general statements about the examinee that could apply to almost any human being.

These criticisms have led some researchers to denounce the use of computers in assessment (Dreyfus & Dreyfus, 1988). However, most researchers take the position that computers are appropriate to use for test administration and scoring but inappropriate to use for interpreting test results, which is better left to expert judgment (Maddux & Johnson 1993). As Maddux and Johnson (1993) have stated,

> "The computer is here to stay. Indeed, it is destined to proliferate enormously. It is such a powerful tool that pressure for implementation has become an irresistible force. Those who wish

> to abolish the computer will fail. The question is not whether to involve computers in assessment, but how to do so intelligently" (p. 194).

If Maddux and Johnson (1993) are correct in their prediction that computers will continue to be used in assessment, it becomes increasingly important that this is carried out in a way that is advantageous and psychometrically sound. For the purposes of the present study, a set of guidelines was developed on the basis of the recommendations of the APA and researchers in this field. These guidelines are presented in the following section.

Guidelines for Developing Computerized Tests from Conventional Tests

Because of the dangers of computerized testing that have been identified in the literature, it is important that rigorous standards be adhered to in the development of computerized test instruments. On the basis of the recommendations of the APA (1986) and subsequent research, the following guidelines have emerged and will be followed in the present study.

First, the conventional instrument should be standardized with established validity and reliability. If the conventional instrument has weaknesses they will only be duplicated in the computer version (Watkins & Kush, 1988).

Second, only appropriate instruments should be computerized. The presentation, scoring and/or interpretation procedures should be easily adaptable to a computerized format. Examinee response modes should also be appropriate for computerization (Schuerholz, 1984-1985). Specifically, the instrument should lend itself to being computer-adaptive (CA) and not merely computer-based (CB). That is, items should be ordered dynamically with regard

to the examinee's previous responses (CA), rather than linearly like a paper pencil test that has merely been transferred to a computer screen (CB) (Wise and Plake, 1989). The computer version of the test should be highly visual and audible to take advantage of the computer's multimedia capabilities and promote a high level of examinee motivation (Malone, 1981). The computerized version should have a complex scoring system requiring a number of calculations (Maddux & Johnson, 1993). The accessibility of instruments that require simple examinee response modes may be increased through computerization. The simpler the response, the greater the opportunity to take advantage of computerized adaptive devices (Schuerholz, 1984-1985).

Third, the administration of the computerized instrument should duplicate that of the conventional instrument as closely as possible to maximize its validity. Any departure from the conventional test should be demonstrated not to affect test scores significantly (APA, 1986).

Fourth, test developers should minimize differences across computer administrations due to individual differences in equipment. Differences in audio quality, speed, color, resolution, and fonts, should be controlled (APA, 1986; Madsen, 1986).

Fifth, computerized tests that use their conventional counterparts' normative, validity and reliability data should be established as being equivalent to their conventional versions. The rank order of scores should be similar, as should means, dispersions and shapes of the scores' distributions (APA, 1986).

In summary, test selection and computer version development should be

based on rational criteria and the psychometric properties of computerized tests should be established before such tests are administered in educational or clinical settings. It is only through following rigorous guidelines that the value of computerized testing can be adequately assessed and the effects of computerization on test performance evaluated.

Purpose of the Present Study

The purpose of the present study was to provide a preliminary assessment of the validity of the computerized test that was developed according to the guidelines presented in the previous section. Specifically, the concurrent validity of this test was investigated within a preschool population.

Rationale for the Selection of the Test Instrument

During the planning of this study regarding the psychometric properties of a computerized test, a variety of instruments were reviewed. An instrument that was found to meet the standards presented previously was the Peabody Picture Vocabulary Test - Revised (PPVT-R; Dunn & Dunn, 1981). The PPVT-R is an individually administered, norm-referenced, wide-range, power test of hearing vocabulary, designed for educational, clinical, vocational, and research uses. The reasons for its selection will be described in this section, with references to the five guidelines previously identified.

First, the PPVT-R is a well established instrument with demonstrated validity and reliability (Bracken & Prasse, 1984; Robertson & Eisenberg, 1981; Stevenson, 1986; Tillinghast, Moorrow & Uhlig, 1982). The PPVT-R is used frequently in clinical and educational settings. For example, it is currently being

used by psychologists (Childers, Durham, & Wilson, 1994), speech-language pathologists (Wagner, 1994); and educators conducting language screenings (Majsterek & Lord, 1991). The PPVT-R is also used in research; a search of PsychLIT and ERIC databases revealed that at least 37 studies using it have been published since 1990.

The test has also been demonstrated to be reliable. In a study by Robertson & Eisenberg (1981) of its psychometric properties, the overall split-half reliability using the Spearman-Brown formula was found to be .80; the overall alternate-form reliability was .84; and the delayed retest reliability was .78. A study by Carvajal, Hayes, Miller and Wiebe (1993) comparing scores on the PPVT-R with scores on the Wechsler Intelligence Scale for Children-III (WISC-III) revealed an overall correlation of .70. Miller and Lee (1993) applied a structural equation model to compare the acquisition order of words to the 175 words used in the PPVT-R. The authors concluded that the model provided further evidence for the construct validity of the test.

Second, the PPVT-R lends itself to computerization. The administration and scoring procedures of the PPVT-R are easily adapted to the computer, and the examinee's response mode can be identical to that required by the conventional administration of the test. Furthermore, the examinee's response mode is simple; mere pointing is required. In addition, the computerized instrument is computer adaptive; each item is administered with regard to the examinee's previous responses. The computerized version can have both visual and audio components to take advantage of the computer's multimedia

capabilities. In the conventional administration of the PPVT-R, calculations are needed to determine chronological age, for on-going scoring, to obtain a raw score, a standard score equivalent, a percentile rank, a stanine and an age-equivalent score. Therefore, a computerized version has the potential to improve examiner accuracy.

Overton and Scott (1972) have stated that the PPVT is ideal for being immediately adaptable to the computer environment. Their early attempts and those of Knights, Richardson and McNarry (1973) have been described previously. Although the testing apparatus in these two studies was complicated, the administration, the examinee's response repertoire, and the scoring, were all converted to the computerized medium. Elwood and Clark (1978) further demonstrated the suitability of computerizing the PPVT. They used computer-controlled technology and stated, "the experiment demonstrated that it is feasible to use a computerized method" (p. 46). In a more recent study, Maguire, Knobel, Knobel, Sedlacek and Piersel (1991) chose the PPVT-R because the simple examinee response mode is easy to adapt to computerized binary microswitches enabling access to a wide variety of special needs populations. The equipment used by Maguire and his colleagues was makeshift, however. The researchers merely hung transparencies of the PPVT-R plates in front of the monitor and scores were not computer generated. The use of more sophisticated equipment and programming could enhance the psychometric soundness of a computerized PPVT-R.

Third, the PPVT-R is ideally suited to the computer medium for the

computer administration of the test will closely match the conventional administration. This will be describe fully in Chapter 2.

Fourth, with the use of modern equipment and programming, the PPVT-R can be designed to minimize differences across different equipment. Speed, colour, resolution, and fonts can be controlled.

Finally, although the psychometric properties of a well-constructed computerized PPVT-R have not been established to date, the present study will constitute one step toward the evaluation of this test's validity. The equivalence of the computerized and conventional versions will be assessed and examinees' performance on each will be analyzed.

Limitations of the PPVT-R.

It is noted in the PPVT-R manual that a limitation of the test is its brevity and simplicity which may lead to casual administration and scoring. The test authors also warn that the PPVT-R is a measure of hearing vocabulary and that overgeneralizations should not be made to broader linguistic or cognitive ability. A variety of well-established tests are available that can provide a thorough assessment of linguistic or cognitive ability. The PPVT-R, because of its brevity, is best used as a screening device for receptive vocabulary in conjuction with a battery of other screening measures. However its circumscribed focus makes it appropriate for computerization and for an initial study of computerizing a conventional test.

Hypothesis

On the bases of findings demonstrating the appropriateness of the PPVT-R

for computerization and the similarity in administration of the computerized and conventional versions of the test, it was hypothesized that the mean raw score obtained on the computerized version of the PPVT-R would not differ significantly from the mean raw score obtained on the conventional PPVT-R when both versions were administered to a preschool sample. Second, it was hypothesized that scores obtained on the computerized version will correlate positively and significantly with scores obtained on the conventional version.

CHAPTER II

Method

Subjects

The number of subjects needed was calculated using a power analysis for a one-tailed Pearson's r for 99% power at the 5% significance level. The critical effect size was set at .70. According to the results of this analysis, 25 subjects were required (Kraemer & Thiemann, 1987). However, as it was feasible to recruit a larger number of subjects, 30 male and 30 female subjects were assessed. After testing was complete, seven subjects were eliminated from the study because errors were found in the paper and pencil test results. The remaining sample consisted of 28 females and 25 males. All subjects were between the ages of 3-0 and 5-0 and were selected from four day care centres in Winnipeg, Manitoba. Only children who had English as a first language were included in the study. Autistic children and children with physical or sensory handicaps were excluded.

The examiner for both conventional and computer versions was the author of the present study. The author was thoroughly familiar with the test and had previous experience in the test's administration while supervised by a registered child clinical psychologist.

Materials

PPVT-R standard format. The PPVT-R is an individually administered, norm-referenced, wide-range test of hearing vocabulary, available in two parallel forms designated L and M. Each form contains 5 training items, followed by 175

test items arranged in order of increasing difficulty. Each item consists of four simple, black-and-white illustrations presented together as a "plate." The examinee's task is to select the picture considered to illustrate best the meaning of a stimulus word presented orally by the examiner. The test is designed for persons 2 1/2 through 40 years of age who can see and hear reasonably well and understand Standard English to some degree. The PPVT-R was standardized nationally across America on a carefully selected sample of 5,028 persons - 4,200 children and adolescents, and 828 adults. The standardization sample included 800 children between the ages of 3.0 and 4.11.

The PPVT-R is administered while the examiner and examinee are seated on either side of a corner of a table. The test is presented as a standing binder that contains 175 pages or plates. The examiner presents a page of the binder and instructs the subject to identify the illustration that best matches the stimulus word (e.g., "POINT TO BALL."). The examinee responds by pointing to one of the four illustrations on the page. The examiner records the response on a score sheet, makes an additional mark if it is incorrect, and continues to the next plate. A basal of 8 consecutive correct responses is established and the test continues until the examinee gives 6 incorrect responses out of 8 consecutive items (ceiling). The examiner then calculates a raw score by subtracting the number of errors from the highest incorrect item number of the lowest ceiling group.

PPVT-R computerized format. The computerized version of the PPVT-R Form L was developed by the author to run on any standard Macintosh

computer. The Macintosh platform was chosen so that the program could be written in HyperCard, a programming language that uniquely suited the style of the programming required. The plates of the PPVT-R were scanned into the computer and adjusted so that they were displayed on the monitor in the same dimensions and resolution as the originals. The 180 stimulus words were digitally recorded on the computer in accordance with the pronunciation guide found in the PPVT-R manual. The entire set of statistical tables found in the PPVT-R manual, which are needed to score the test, was entered into the computer to provide computerized scoring. A program was then written that would administer and score the test and provide a printout of the results.

The initial screen of the PPVT-R computer program prompts the examiner to input information about the examinee via the keyboard (see Figure 1). The chronological age of the examinee is automatically calculated by the computer after the examiner has entered the date of testing and the examinee's date of birth, eliminating examiner error. This is important because the selection of the first picture plate to be administered is based on the examinee's age. The examiner then gives verbal instructions to the examinee according to the original manual and, with the use of a touch-sensitive screen, presses the practice button to start the test. The first practice plate is displayed on the monitor and the stimulus phrase "point to doll," is heard through the speaker (see Figure 2). The examinee is required to touch the picture that corresponds to the word. The response is automatically recorded and the examiner is free to observe the examinee. (If the examiner or examinee prefers, a mouse, keyboard, or adaptive

Figure 1. Initial input screen of the computerized PPVT-R.

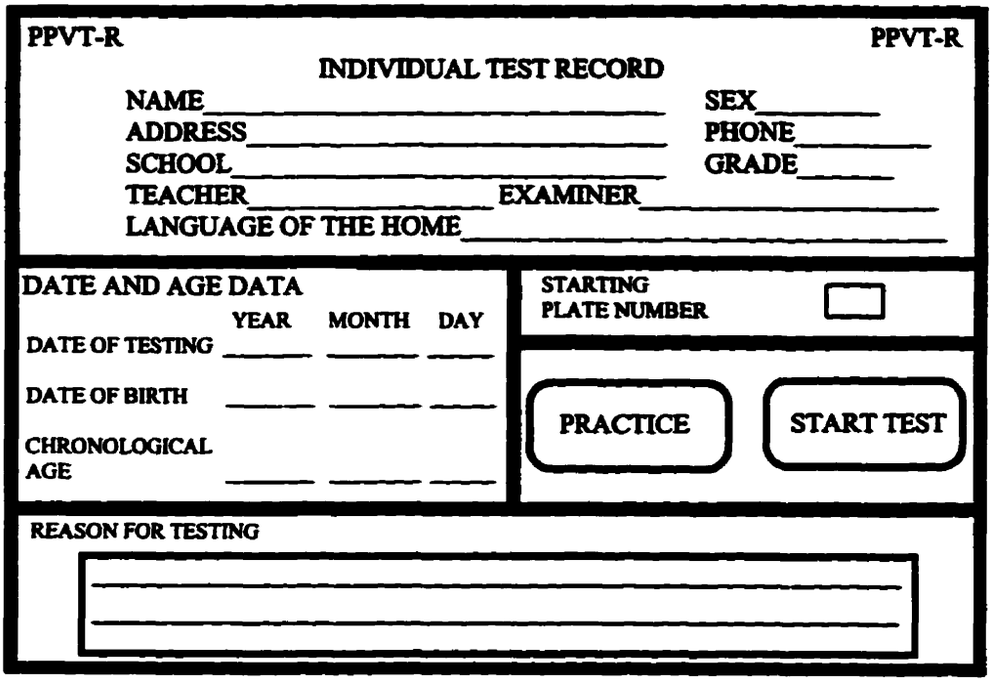
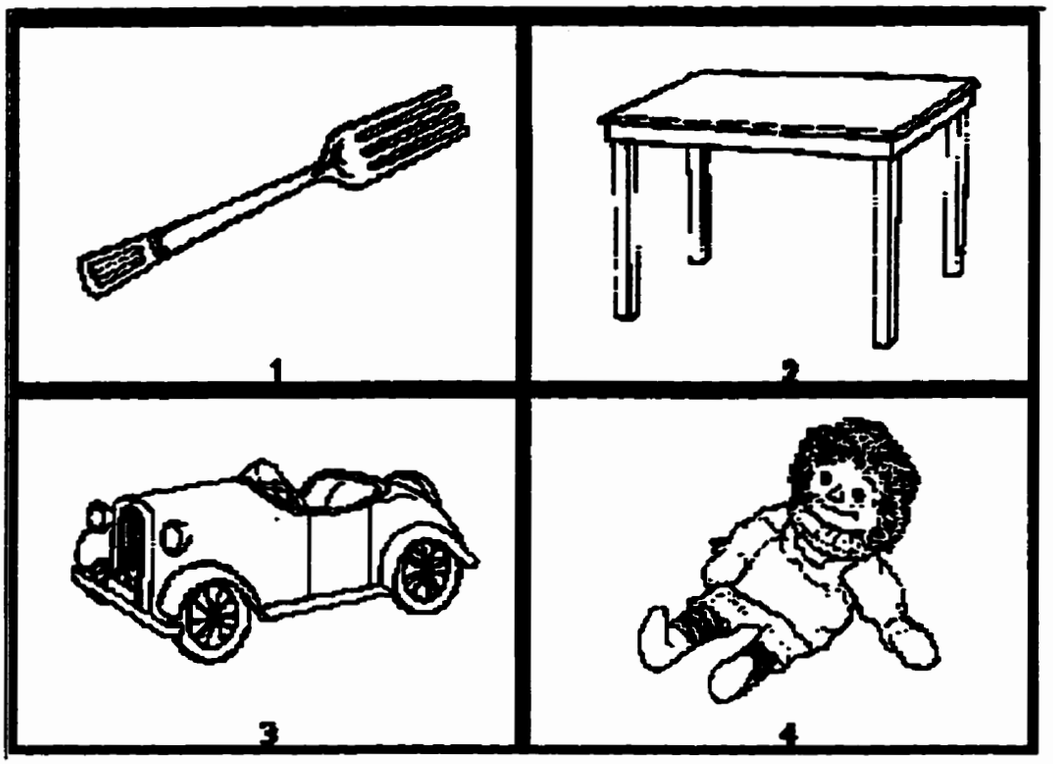


Figure 2. Initial practice plate of the computerized PPVT-R.

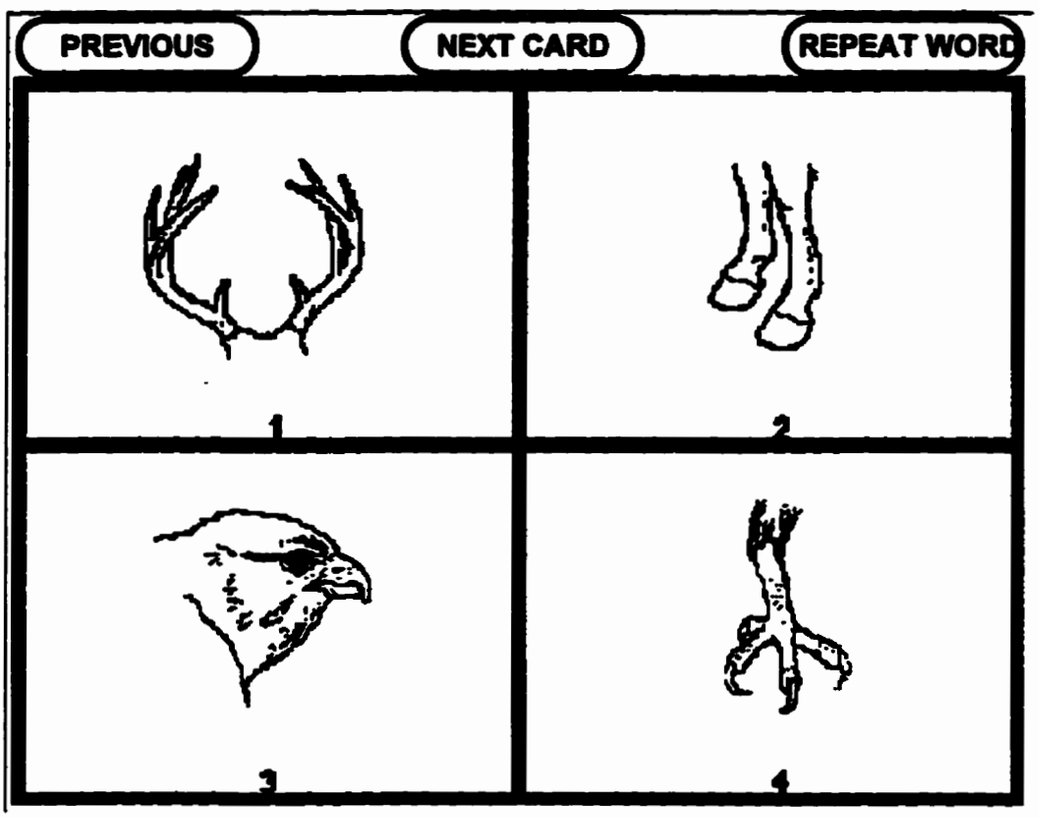


device can be used.[1]) For each of the test plates, if the stimulus phrase was not heard adequately the examinee may press the "repeat word" button at the top of the screen (see Figure 3). The computer will then play the stimulus phrase again. The programme was also designed to allow examinees to change a previous response by pressing the "previous" button, or to skip an item and proceed with the next one by pressing the "next card" button. These are features of the original test that have been maintained in the computer version.

The program was designed to be computer adaptive (CA). Specific algorithms were written to select each plate for presentation on the basis of the examinee's previous responses, as would be done by the examiner in the original paper and pencil version. The timing of administration is standardized regardless of the speed of the computer that is being used. The minimum standard is a Macintosh computer with an 030 processor running at 16 mhz. Any Macintosh computer with greater speed will present the plates at the same rate. Each plate is presented two seconds after the examinee responds to the previous plate. This two second interval is maintained throughout the test regardless of the order of the presentation of the plates. This design maximizes efficiency, standardization, and fluidity of administration.

Due to the adaptive nature of the program, the computer will end the test with different items for different examinees. When the test ends, a plate displays the message, "THE TEST IS FINISHED." The computer also plays this message audibly. At this point, the computer begins to calculate the examinee's basal, ceiling, and error scores, the raw score, the standard score equivalent, the

Figure 3. A typical plate of the computerized PPVT-R.



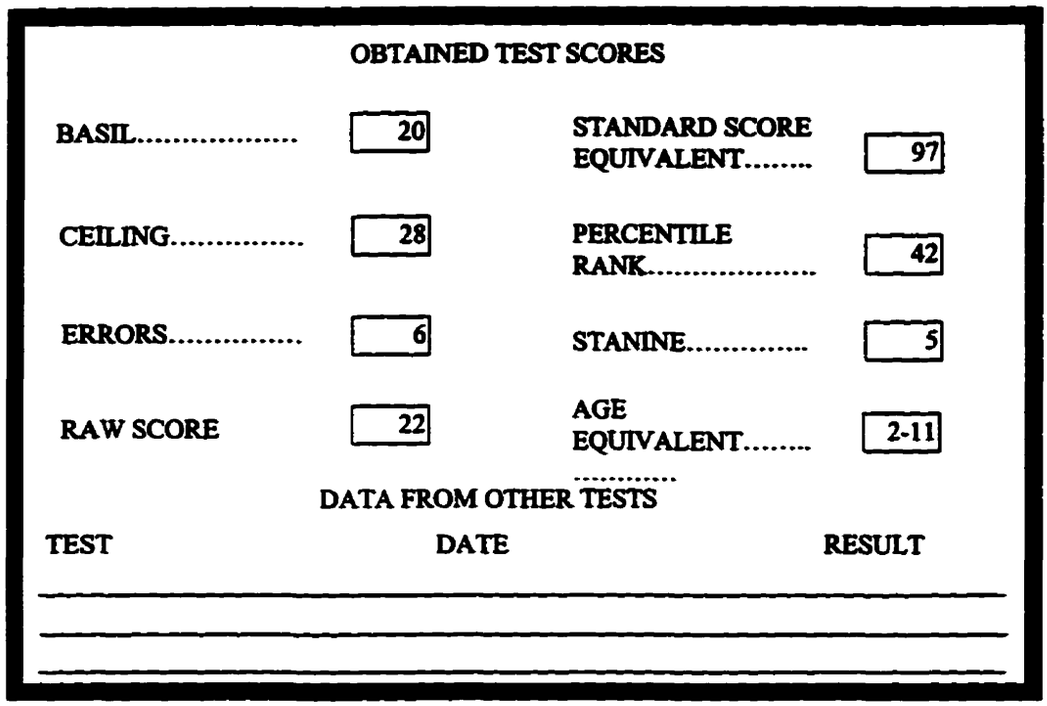
percentile rank, the stanine, and the age-equivalent score. The computer then creates an individualized computer file for the examinee that contains the initial demographic material recorded at the beginning of the test, a record of the examinee's responses, all calculated scores, a graph of the True Score Confidence Band, and templates for written comments on observations or recommendations (see Figures 1, 4, 5, 6, & 7). This set of procedures virtually eliminates examiner error in calculating scores and completes the examinee's results file in less than one minute. The results can be stored on disk or printed out as a hard copy file.

Procedure

Obtaining consent. The author of the present study contacted and met with day care centre directors to describe the study and to request their participation. After receiving consent from the directors, information letters and consent forms (see Appendix A) were distributed to the appropriate parents in the day care centre. An empty envelope was provided so that parents could return their consent forms to the director of the day care centre.

Test administration. A within-subjects design was used. The conventional and computerized versions of the PPVT-R Form L were given in counterbalanced order, with half of the subjects receiving the conventional version first and the other half receiving the computerized version first. There was a one-week interval between administrations. The limited attention span of three-and-four-year-olds precludes administration of both tests in the same session. A one-week delay was expected to minimize practice effects. Subjects

Figure 4. Example of an obtained test scores sheet of the individual results file of the computerized PPVT-R.



OBTAINED TEST SCORES

BASIL.....	20	STANDARD SCORE EQUIVALENT.....	97
CEILING.....	28	PERCENTILE RANK.....	42
ERRORS.....	6	STANINE.....	5
RAW SCORE	22	AGE EQUIVALENT.....	2-11

DATA FROM OTHER TESTS

TEST	DATE	RESULT

Figure 5. Observations, performance evaluation, and recommendations sections of the individual results file of the computerized PPVT-R.

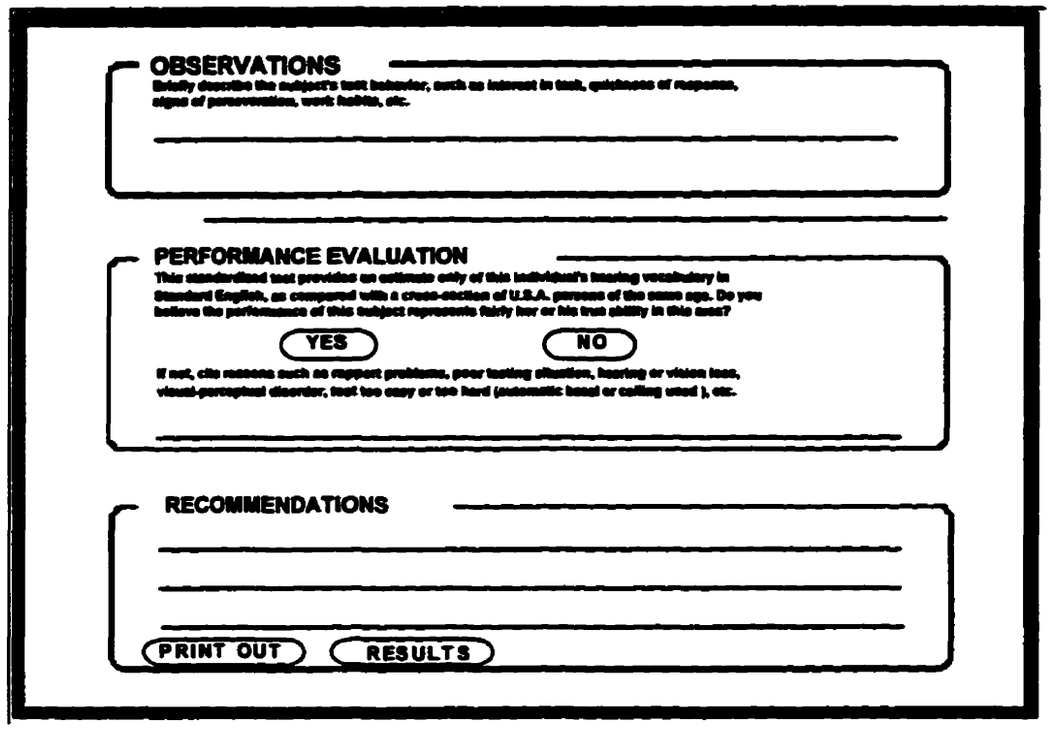
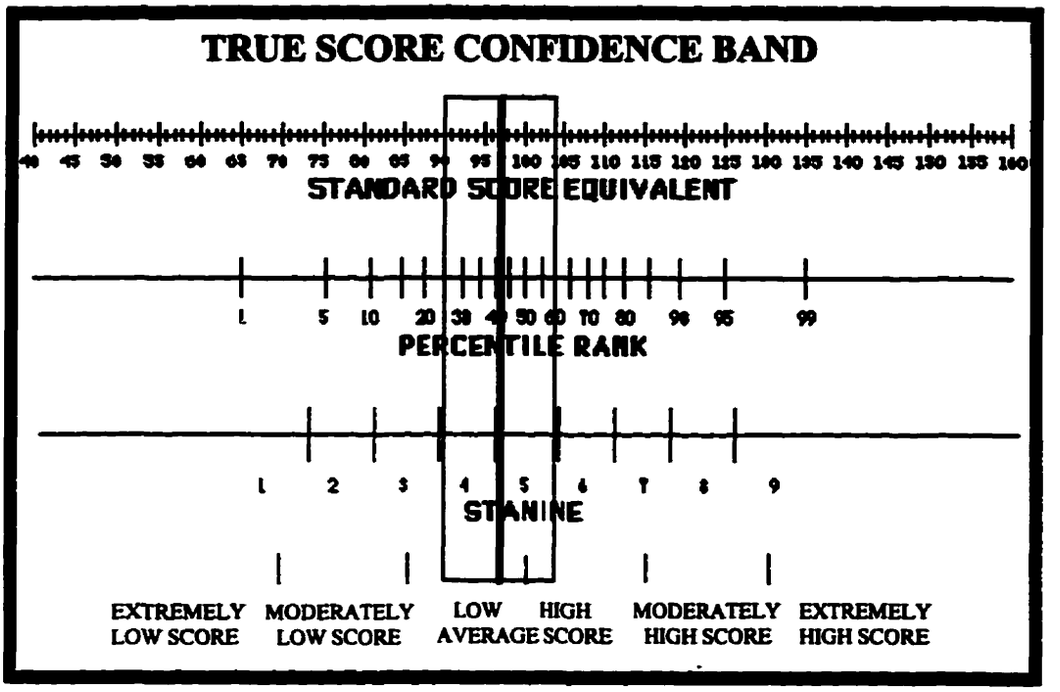


Figure 6. Response sheet of the individual results file of the computerized PPVT-R.

plate number	word	response key	response	error	plate number	word	response key	response	error
001	bus	(4)	4	0	016	feather	(1)	1	0
002	hand	(1)	1	0	017	empty	(3)	3	0
003	bed	(3)	3	0	018	fence	(4)	4	0
004	tractor	(2)	2	0	019	accident	(2)	2	0
005	closet	(1)	1	0	020	net	(2)	2	0
006	snake	(4)	4	0	021	tearing	(4)	3	x
007	boat	(2)	2	0	022	sail	(1)	3	x
008	tire	(3)	3	0	023	measuring	(2)	2	0
009	cow	(1)	1	0	024	peeling	(3)	3	0
010	lamp	(4)	4	0	025	cage	(1)	3	x
011	drum	(3)	3	0	026	tool	(4)	2	x
012	knee	(4)	4	0	027	square	(4)	2	x
013	helicopter	(2)	2	0	028	stretching	(1)	2	x
014	elbow	(2)	2	0					
015	bandage	(2)	2	0					

Figure 7. The true score confidence band sheet of the individual results file of the computerized PPVT-R.



were randomly assigned to one of the two order-of-testing conditions.

The two versions of the test were administered to children in their day care centre. The conventional version of the PPVT-R was administered as described in the manual. The computer version was administered as described in the previous section. The investigator brought the same MacIntosh Si computer to each day care centre.

Data Analysis. A Pearson product-moment correlation coefficient was calculated to determine the strength of association between subjects' performance on the two versions of the test and means and standard deviations of raw scores on each version of the test were calculated. A paired two-tailed t-test was also used to determine if there was a significant difference in subjects' performance between the conventional and computerized versions of the test.

CHAPTER III

Results

Hypothesis 1

It was predicted that scores obtained on the CPPVT-R would correlate positively and significantly with scores obtained on the PPVT-R. This hypothesis was strongly supported (see Table 1). A statistically significant correlation was found between the CPPVT-R and the PPVT-R ($p < .001$). Therefore, the CPPVT-R appears to have concurrent validity.

In order to examine the concurrent validity of the CPPVT-R within subgroups of the sample, four Pearson Product-Moment correlation coefficients were computed between the two versions of the test for girls, boys, three-year-olds and four-year-olds, respectively. Because the number of correlational tests being conducted inflates the chance of a Type I error, the Bonferroni correction was applied. The original significance level of .05 was divided by the total number of correlational tests (5) to give a corrected significance level of .01. As Table 1 reveals, the correlations between the two version of the test were highly significant among all subgroups ($p < .001$). Therefore, the CPPVT-R is also concurrently valid for all subgroups of the sample.

Hypothesis 2

It was expected that the mean raw score obtained on the CPPVT-R would not differ significantly from the mean raw score obtained on the PPVT-R. This hypothesis was not supported. The sample performed at a higher level on the PPVT-R than the CPPVT-R ($p < .01$). Means, standard deviations, and t - values

Table 1

Pearson Product-Moment Correlations Between CPPVT-R and PPVT-R Raw Scores for Total Sample and Subgroups

Group	*r*	*n*
Overall	.88*	53
Gender		
Female	.85*	28
Male	.91*	25
Age		
36 - 47 Months	.83*	26
48 - 60 Months	.82*	27

* $p < .001$

are presented in Table 2. To examine this difference more fully, the sample was subdivided by gender and by age. Four additional t-tests were conducted to assess the roles of these variables in accounting for differences in performance on the two versions of the test. The Bonferroni correction was also applied here; the original significance level of .05 was divided by the number of t-tests performed (5) to yield a significance level of .01

As Table 2 shows, the performances of female children on the two versions of the test did not differ significantly ($p > .01$), nor did those of male children ($p > .01$). Among four-year-old children, mean scores on the 2 versions of the test did not differ significantly ($p > .01$). However, three-year-olds performed significantly better on the PPVT-R than the CPPVT-R ($p > .01$). Therefore, age appears to play an important role in children's performance on the CPPVT-R.

Table 2

Mean CPPVT-R and PPVT-R Raw Scores, Standard Deviations and t-Values for Total Sample and Subgroups

Group	M	SD	t-value	n
Overall				
CPPVT-R	34.74	16.76	-3.24*	53
PPVT-R	38.32	14.98		
Gender				
Female				
CPPVT-R	37.46	16.82	-2.67	28
PPVT-R	42.00	15.04		
Male				
CPPVT-R	31.68	16.39	-1.83	25
PPVT-R	34.20	14.08		
Age				
36 - 47 Months				
CPPVT-R	24.88	13.32	-3.38*	26
PPVT-R	30.12	13.74		
48 - 60 Months				
CPPVT-R	44.22	14.06	-1.30	27
PPVT-R	46.22	11.62		

*$p < .01$

CHAPTER IV

Discussion

The purpose of the present study was to provide a preliminary assessment of the validity of a computerized test that was developed according to guidelines established for the computerization of conventional tests. Specifically, the concurrent validity of a computerized version of the PPVT-R was to be investigated within a preschool population.

Observational Findings

Strengths of the CPPVT-R. Although the sample assessed in the present study was small and non-representative, the initial assessment of the CPPVT-R looks promising. The computer and the computer program performed flawlessly over 60 trials in 4 different day care centres. The program proved to be child-proof as the children could inadvertently touch the keyboard or the touch screen and not affect the testing procedures or results. The touch-screen also seemed especially ergonomic and accessible for the children. Most of the children tested were excited to start "the computer game," and the "talking computer with pictures to touch" also seemed to motivate the older children throughout the administration of the test.

The accuracy of the CPPVT-R was also demonstrated. The picture plates were presented in the correct adaptive order for all test administrations. Test results were calculated automatically on the computer immediately following each test and were found to be error-free. The program also performed perfectly when previous responses were revisited and changed during testing.

As the administrator of the CPPVT-R merely had to initiate the testing, he had a greater opportunity to observe the children, as well as to provide feedback and reinforcement in a very natural and effective manner. Further, accurate results were calculated immediately and a complete file on the child could be printed out within minutes.

Limitations of the CPPVT-R. The current version of the CPPVT-R is flawed in terms of the time delay between item presentations. The time delay was estimated to match the time it would take an administrator of the conventional test to record an answer on the score sheet and turn the page to the next item presentation. However, during test administration in the present study, it was observed that the two-second delay exceeded the time needed to record responses by paper and pencil and therefore extended the length of administration for the CPPVT-R.

Another observed limitation of the CPPVT-R was that the younger children occasionally became distracted from the testing task because of interest in the digital voice coming from the computer. As stated in the introduction, a certain level of curiosity or incongruity is important in motivating children. However, in this case the digital voice may have exceeded the appropriate level of incongruity for the younger children.

Statistical Findings

The findings of the present study indicate that the CPPVT-R has concurrent validity for boys and girls in the 3-to 4-year-old age group. In fact, the correlations obtained in the present study are higher than the alternate form

reliability correlation coefficient obtained for the same age group (r = .78; Dunn & Dunn, 1981). Further, the correlation coefficients obtained between the CPPVT-R and the PPVT-R in the present study are higher than the average correlation coefficients for alternate form reliability calculated in a review of 24 studies of the PPVT-R (r = .83; Braken, Prasse, & McCallum, 1984). Finally, the correlation coefficients obtained in the present study are comparable to that obtained in a previous comparative study of conventional and computerized versions of the PPVT-R (r = .91, Maguire et al., 1991).

The significant difference in overall raw score means between the PPVT-R and the CPPVT-R demonstrates that the children performed at a significantly higher level on the conventional version of the PPVT-R. This finding does not support the hypothesis that the two tests would yield equivalent scores. The difference in scores appears to be largely attributable to an age effect; while the older children performed at similar levels on the two tests, the younger children performed better on the original version than on the computerized version.

It is possible that this age difference is at least partially due to the timing difference between the two versions of the test. As research has shown that the ability to attend develops with age (Levy, 1980), the younger children may have received lower scores on the CPPVT-R because their attention may have waned in the longer computer version, while older children, having longer attention spans, may not have been affected by the extended length of the test.

Another possible explanation for the age difference may be the incongruity of the digital voice of the CPPVT-R. For some of the younger

children, the digitally recorded voice seemed distracting, which may have affected test results. The older children however, seemed only moderately curious about the digital voice and stayed on-task.

Limitations of the Present Study

The sample of the present study was small (53 subjects) and selective (day care children aged three-and-four-years). Therefore, the results of this study should be generalized cautiously.

The within-subjects design of the study may also have inherent problems such as history, maturation, and testing effects. For example, the children's vocabulary could have increased within the week between testing sessions either as a result of participation within the day care program or from maturational growth. This may have affected the children's performance on the second testing. The children's familiarity with the test items from the first testing could also have affected their performance on the second administration of the test. A between-subjects design, however, would require extensive matching procedures to eliminate potential confounding variables.

Directions for Future Research

The findings of the present study raised several questions that need to be addressed in future research. First, the effect of the length of the test on young children's performance needs to be investigated. This issue raises an interesting theoretical question. In psychological testing for young children, "time in administration" is not standardized across administrations. In the case of the PPVT-R, for example, the time taken to administer the test will vary greatly

across examiners and situations. Individual administrators will have different rhythms which will affect the length of the test. Differences in the length of administration of individual test items will also fluctuate as examiners are required to perform different tasks, such as turning back seven pages versus moving forward one page, or calculating basals or ceilings. These differences in timing may not affect the performance of older children, but may affect that of younger children whose attending skills are not as well developed. Further research is needed in this area to determine if the duration of a test is a factor that affects young children's performance. Additional research with computerized tests that could eliminate this timing variable and increase standardization of administration is also warranted.

A follow-up to this study should also be conducted to determine if three-year-old children would perform equivalently on the two versions of the test if the length of the two tests were controlled. Additional research on the CPPVT-R should also be conducted with response-limited individuals to examine if the computer can significantly enhance performance. Finally, the CPPVT-R should be tested on a larger, more representative sample to assess equivalency to the PPVT-R more fully.

Conclusion

Overall, the CPPVT-R was found to be practical, and to have concurrent validity. The results also suggest that the guidelines proposed in this study are effective for use in the development of computerized tests. In the future, it will be crucial that evaluations of computerized tests are conducted frequently in order

to keep pace with advancements in computer technology. Test developers need to be aware of technological advancements in order to provide the most reliable and valid tests possible. However, researchers should keep in mind that not all technological advancements will be useful. The computer may still be appropriate only for certain tasks in combination with a human examiner. The task for researchers in this field is to find the best balance between effective computer technology and human interaction.

Appendix A

Information Letter to Parents

Dear Parent,

Did you know that the average 3-year-old child can understand 900 words? Or that this number doubles over the next 18 months? The work of many researchers has helped us to understand how children come to understand what we are saying to them and why some children may have difficulty in understanding language or expressing themselves.

Researchers have used a variety of measures to assess children's understanding of words. These measures are constantly being improved so that our understanding can be more and more precise. One of the ways that researchers are currently using to measure children's understanding of language is through computer programs. However, we do not have a good understanding of how useful these programs are compared to conventional paper-and pencil measures.

I am currently conducting a study of the usefulness of one computer program for assessing children's understanding of words and I would like to request your child's participation. The study would take about 30 minutes of your child's time and can be conducted in the day care centre. I would see your child twice - once to give the paper-and-pencil version of the measure (15 minutes) and once to give the computer version of the measure (15 minutes). Each measure would simply ask your child to point to a picture that matches a word. The two measures would be given one week apart. This study has been approved by the Faculty of Human Ecology Ethics Review Committee and is supported by the Director of the day care centre.

Your child does not need to have any computer experience in order to participate. However, for this study, all of the children must speak English as their first language, be either 3 or 4 years old, and not have been diagnosed with any sensory or motor impairments.

If you decide that you would like your child to participate, just sign the attached consent form, place it in the envelope provided, and return it to the day care centre. If you decide not to have your child participate, this will not affect in any way the services that your child receives at the day care centre.

Thank you for considering my request.

Sincerely,

Andrew Robson, M.A. Candidate

Parental Consent Form

I _____ allow my child _____ to participate in the research investigation entitled, " A Computerized Measure of Children's Receptive Vocabulary" which is being conducted by Andrew Robson, a graduate student at the University of Manitoba. I understand that the purpose of the study is to investigate the usefulness of a computer program designed to measure children's understanding of words. If I consent, my child will spend 15 to 20 minutes matching pictures on cards to words and, the following week, my child will spend 15 to 20 minutes matching pictures on a computer to words. I understand that the results of this investigation will be confidential, will not be released without my written consent and will be destroyed when the study is completed. I further understand that if the results of this study are published, neither I or my child will be identified in any way. I understand that I will receive a summary of the findings of the study. I understand that I may withdraw my child from this study at any time, without penalty, even after signing this form. If I have any questions, I may contact Andrew Robson at 474-9225.

_____	_____	_____
Signature of Parent or Guardian	Phone number	Date

Endnote

[1]In the adaptive response mode for examinees with motor handicaps, a rectangular highlighting frame is added to the plate that circulates through the four pictures. It starts by highlighting the frame for picture one, then moves to picture two, to three, to four, then back to one, and so on. The examinees are required to activate a binary microswitch when the highlighting frame is on the picture of their choice. This simple response makes available a wide variety of adaptive response devices, such as large button switches, foot pedal switches, breath activated switches, eye blink switches, or any other type of switch that can provide simple binary input.

REFERENCES

Abkarian, G.G., King, P., & Krappes, T.L. (1987). Enhancing interaction in a difficult-to-test child: The PPVT-R TV technique. Journal of Learning Disabilities, 20(5), 268 - 269.

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments. (1985). Standards for educational and psychological testing. Washington, DC: Author.

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessments. (1986). Guidelines for computer-based tests and interpretations. Washington, DC: Author.

Argulewicz, E. N., & Kush, J. C. (1983). Equivalence of forms L and M of the PPVT-R for use with Anglo-American and Mexican-American learning disabled students. Psychological Reports, 52(3), 827-830.

Beaumont, J. G. (1981). Microcomputer-aided assessment using standard psychometric procedures. Behavior Research Methods and Instrumentation, 13(4), 430-433.

Berlyne, D. E. (1966). Curiosity and exploration. Science , 153, 25-33.

Bing, S. B. & Bing, J. R. (1984). Concurrent validity of the PPVT-R for college students. Psychological Reports, 55(3), 863-866.

Bracken, B. A., & Prasse, D. P. (1984). Peabody Picture Vocabulary Test-Revised: An appraisal and review. School Psychology Review, 13(1), 49-60.

Brahan, J. W., & Brown, W. C. (1972). The NRC computer aided learning cooperative research project. Proceedings of the Canadian Symposium on Instructional Technology. Calgary: 1972.

Bunderson, C., Inouye, D., & Olsen, J. (1989). The four generations of computerized education measurement. In Robert Linn (Ed.), Educational measurement, (3rd edition), (pp. 367-407). New York, NY: American Council of Education MacMillian.

Burke, R. L. (1982). CAI sourcebook. Englewood Cliffs, NJ: Prentice-Hall.

Butzin, S. M. (1990). Project CHILD: "Not boring school, but work that's fun and neat." Computing Teacher, 17, 20-23.

Carvajal, H., Hayes, J. E., Miller,H. R., & Wiebe, D. A. (1993). Comparisons of the vocabulary scores and IQs on the Wechsler Intelligence Scale for Children-III and the Peabody Picture Vocabulary Test--Revised. Perceptual and Motor Skills, 76(1), 28-30.

Childers, J. S., Durham, T. W., & Wilson, S. (1994). Relation of performance on the Kaufman Brief Intelligence Test with the Peabody Picture Vocabulary Test--Revised among preschool children. Perceptual and Motor Skills, 79, 1195-1199.

Choi, H., & Proctor, T. B. (1994). Error-prone subtests and error types in the administration of the Stanford-Binet Intelligence Scale: Forth edition. Journal of Psychoeducational Assessment, 12(2), 165-171.

Collins, M., & Odell, K. (1986). Computerization of a traditional test for nonverbal visual problem solving. Cognitive Rehabilitations, 4(5), 16-18.

Condry, J. C. (1977). Enemies of exploration: Self-initiated versus other-initiated learning. Journal of Personality and Social Psychology, 35, 459-477.

deCharms, R. (1968). Personal causation. New York: Academic Press.

Dreyfes, H. L., & Dreyfus, S. E. (1988). Mind over machine: The power of human intuition and expertise in the era of the computer. New York: The Free Press.

DuBois, P. H. (1970). A history of psychological testing. Boston: Allyn & Bacon.

Dunn, L. M. (1959). Peabody Picture Vocabulary Test. Circle Pines, MN. American Guidance Service.

Dunn, L. M. (1981). Peabody Picture Vocabulary Test-Revised. Circle Pines, MN. American Guidance Service.

Eberly, C. G., & Cech, E. J. (1986). Integrating computer-assisted testing and assessment into the counseling process. Measurement and Evaluation in Counseling and Development, 1, 18-28.

Eisele, J. A., & Aram, D. M. (1993). Differential effects of early hemisphere damage on lexical comprehension and production. Special Issue: Acquired childhood aphasia. Aphasiology, 7(5), 513-523.

Elwood, D. L. (1969) Automation of psychological testing. American Psychologist, 24, 287-289.

Elwood, D.L. (1972). Test retest reliability and cost analysis of automated and face to face intelligence testing. International Journal of Man-Machine Studies, 4, 1-23.

Elwood, D. J., & Griffin, R. H. (1972). Individual intelligence without the examiner: Reliability of an automated method. Journal of Consulting and Clinical Psychology, 38, 9-14.

Elwood, D. L., & Clark, C. L. (1978). Computer administration of the Peabody Picture Vocabulary Test to young children. Behavior Research Methods and Instrumentation, 10(1), 43-46.

Endicott, J., & Spitzer, R. L. Current and past psychopathology scales (CAPPS): Rationale, reliability, and validity. Archives of General Psychiatry, 27, 678-687.

Evan, W. M., & Miller, J. R. (1969). Differential effects on response bias of computer vs. conventional administration of a social science questionaire. Behavioral Science, 14, 216-227.

Finney, J. C. (1966). Programmed interpretation of MMPI and CPI. Archives of General Psychiatry, 15, 75-82.

Foreman, N., & Hemmings, R. (1987). The Gollin Incomplete Figures Test: A flexible, computerized version. Perception, 16(4), 543-548.

Gorham, D. R. (1967). Validity and reliability studies of a computer based scoring system for inkblot responses. Journal of Consulting Psychology, 31, 65-70.

Groeneweg, G., Conway, D. G., & Stan, E. A. (1986). Performance of adults with developmental handicaps on alternate forms of the Peabody Picture Vocabulary Test. Journal of Speech and Hearing Disorders, 51(3), 259-263.

Hagan, D. (1984a). Jason says "yes." Pointer, 28(2), 40-43.

Harter, S. (1981). A new self-report scale of intrinsic verus extrinsic orientation in the classroom: Motivational and informational components. Developmental Psychology, 3, 300-321.

Herz, M. L., Endicott, J., Spitzer, R. L. & Mesnikoff, A. (1971). Day vs. inpatient hospitalization: A controlled study. American Journal of Psychiatry, 127, 1371-1382.

Hunt, J. McV. (1965). Intrinsic motivation and its role in psychological development. In D. Levine (Ed.), Nebraska Symposium on Motivation (Vol. 13. pp. 189-282). Lincoln, NE: University of Nebraska Press.

Johnson, J. (1979). Technology. In T. Williams & J. Johnson (Eds.), Mental health in the 21st century (pp. 7-9). Lexington, MA: D.C. Health.

Johnson, D. E., & Mihal, W. L. (1973). Performance of blacks and whites in computerized versus manual testing environments. American Psychologist, 694-699.

Kagan, J. (1972). Motives and development. Journal of Personality and Social Psychology, 22, 51-66.

Kinzie, M. B., Sullivan, H. J., & Berdel, R. L. (1992). Motivational and achievement effects of learner control over content review within CAI. Journal of Educational Computing Research, 8(1), 101-114.

Kleinmuntz, B., & McLean, R. S. (1968). Diagnostic interviewing by digital computer. Behavioral Science, 13, 75-80.

Knights, R. M., Richardson, D. H., & McNarry, L. R. (1973). Automated vs. clinical administration of the Peabody Picture Vocabulary Test and the Coloured Progressive Matrices. American Journal of Mental Deficiency, 78(2), 223-225.

Kramer, J.J. (1988). Computer-based test interpretation in psychoeducational assessment: An initial appraisal. Journal of School Psychology, 26, 143-153.

Lasky, D. I., Felice, A., Moyer, R. C., Buddington, J. F., & Elliot E. S. (1973). Examiner effects with the Peabody Picture Vocabulary Test. Journal of Clinical Psychology, 29(4), 456-457.

Lepper, M. R. (1985). Microcomputers in education: Motivational and social issues. American Psychologist, 40, 1-18.

Levy, A., & Barowsky, E. I. (1986). Comparison of computer administered Harris-Goodenough Draw-A Man Test with standard paper-and pencil administration. Perceptual-and-Motor Skills, 63, 395-398.

Levy, F. (1980). The development of sustained attention (vigilance) and inhibition in children: Some normative data. Journal of Child Psychology and Psychiatry, 21, 77-84.

Maddux, C. D. (1984). Breaking the Everest Syndrome in educational computing: An interview with Gregory Jackson and Judah L. Schwartz. Computers in the School, 1(2), 37-48.

Maddux, C.D. & Johnson, L. (1993). Best practices in computer-assisted assessment. In H. B. Vance (Ed.), Best practices in assessment for school and clinical settings.(pp. 177-200). Brandon, Vermount: Clinical Psychology Publishing Company, Inc.

Madsen, D. H. (1986). Computer-assisted testing and assessment in counseling: Computer applications for test administration and scoring. Measurement and Evaluation in Counseling and Development, 1, 6-14.

Maguire, K. B., Knobel, M. M., Knobel, B.L., Sedlacek, L. G., & Piersel, W.C. (1991). Computer-adapted PPVT-R: A comparison between standard and modified versions within an elementary school population. Psychology In The Schools, 28,199-205.

Majsterek, D. J. & Lord, E. N. (1991). An evaluation of the PPVT-R and VMI for screening preschoolers who are at risk for reading disabilities. Diagnostique, 16(2-3), 173-179.

Malone, T. W. (1981). What makes things fun to learn? A study of intrinsically motivating computer games. Pipeline, 49, 50-51.

Martin, T. A., & Wilcox, K.L. (1989). HyperCard administration of a block-design task. Behavior Research Methods Instruments and Computers, 21(2), 312-315.

McCledon, S. L. (1989). First grade spelling success with keyboarding. The Computing Teacher, 17, 35-36.

McMillan, J. H., & Forsyth, D. R. (1991). What theories of motivation say about why learners learn. New Directions For Teaching and Learning, 45, 39-51.

Meier, S. T. & Geiger, S. M. (1986). Implications of computer-assisted testing and assessment for professional practice and training. Measurement and Evaluation in Counseling and Development, 2, 29-37.

Miller, C., Witt, J., & Finley, J. (1981). School psychologists' perceptions of their work: Satisfactions and dissatisfactions in the United States. School Psychology International, 2, 1-3.

Miller, L. T., & Lee, C. J. (1993). Construct validation of the Peabody Picture Vocabulary Test-- Revised: A structural equation model of the acquisition order of words. Psychological Assessment, 5(4), 438-441.

Moe, K. C., & Johnson, M. F. (1988). Participants' reactions to computerized testing. Journal of Educational Computing Research , 4(1), 79-86.

Monsaas, J. A., & Engelhard, G. (1990, October). Attitudes toward testing practices as cheating and teachers' testing practices. Paper presented at the Annual Meetings of the Georgia Educational Research Association, Carrollton, Georgia.

Overton, G. W., & Scott, K. G. (1972). Automated and manual intelligence testing: Data on paralle forms of the Peabody Picture Vocabulary Test. American Journal of Mental Deficiency, 76(6), 639-643.

Paitich, D. A. (1973). A comprehensive automated psychological examination and report (CAPER). Behavioral Science, 18, 131-136.

Pearson, J. S., Swenson, W.M., Rome, H.P., Mataya, P., & Brannick, T. L. (1964). Further experience with the automated Minnesota Multiphasic Personality Inventory. Mayo Clinic Proceedings, 39, 823-829.

Piotrowski, Z A. (1964). A digital computer administration of inkblot test data. Psychiatric Quarterly, 38, 1-26.

Pressey, S. L. (1926). A simple apparatus which gives tests and scores-and teaches. School and Society, 23, 373-376.

Pressey, S. L. (1927). A machine for automatic teaching of drill material. School and Society, 25, 549-552.

Reiner, H., & Hartshorne, T. (1982). Job burnout and the school psychologist. Psychology in the schools, 19, 508-512.

Reschly, D., & Grimes, J. (1990). Best practices in intellectual assessment. In A. Thomas & J. Grimes (Eds.), Best practices in school psychology-11 (pp. 425-439).

Robertson, G. J., & Eisenberg, J. L. (1981). Peabody Picture Vocabulary Test - Revised: Technical Supplement. Circle Pines, MN: American Guidance Service.

Ronau, R. N., & Battista, M. T. (1988). Microcomputer versus paper-and-pencil testing of student errors in ratio and proportion. Journal of Computers in Mathematics and Science Testing, 7, 33-38.

Saccardi, M. (1991). The interactive computer: Authors and readers online. School Library Journal, Oct. ,36-38.

Schuerholz, L. J. (1984-1985). The use of technology and media in the assessment of exceptional children. Diagnostique, 10, 197-208.

Seymour, S. L., Sullivan, H. J., Story, N. O., Mosley, M. L. (1987). Microcomputers and continuing motivation. Educational Communications and Technology Journal, 35, 18-23

Skinner, B.F. (1954). The science of learning and the art of teaching. Harvard Educational Review, 24, 86-97.

Skinner, B. F. (1958). Teaching Machines. Science, 128, 969-977.

Slate, J. R. & Hunnicutt, L.C. (1988). Examiner errors on the Wechsler Scales. Journal of Psychoeducational assessment, 6, 280-288.

Stevenson, J. D. (1986). Alternate form reliability and concurrent validity of the PPVT–R for referred rehabilitation agency adults. Journal of Clinical Psychology, 42(4), 650-653.

Thiemann, S. & Kraemer, H. C. (1984). Sources of behavioral variance: Implications for sample size decisions. American Journal of Primatology, 7(4), 367-375.

Tillinghast, B. S., Morrow, J. E., & Uhlig, G. E. (1982). Validity of the Peabody Picture Vocabulary Test–Revised using California Achievement Tests as criteria. Educational and Psychological Research, 4(3), 147-152.

Van Merrienboer, G., Jeroen, J., & Jelsma, O. (1988). The Matching Familiar Figures Test: Computer or experimenter controlled administration? Educational and Psychological Measurement, 48(1), 161-164.

Wagner, P. A. (1994). Adaptations for administering the Peabody Picture Vocabulary Test–Revised to individuals with severe communication and motor dysfunctions. Mental Retardation, 32(2), 107-112.

Walsh, W. B., & Betz, N. E. (1995). Tests and assessments. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Watkins, M. W., & Kush, J. C. (1988). Assessment of academic skills of learning disabled students with classroom microcomputers. School Psychology Review, 17(1), 81-88.

Weizenbaum, J. (1976). Computer power and human reason. New York: W. H. Freeman.

White, R. W. (1959). Motivation reconsidered: The concept of competence. Psychological Review, 66, 297-333.

Whitten, J., Slate, J.R., Jones, C. R., Shine, A. E., & Raggio, D. (1994). Examiner errors in administration and scoring the WPPSI-R. Journal of Psychoeducational Assessment, 12, 49-54.

Wilson, S.L., Thompson, J. A., & Wylie, G. (1982). Automated psychological testing for the severly physically handicapped. International Journal of Man Machine Studies, 17(3), 291-296.

Wise, S. L., & Plake, B. S. (1989). Research on the effects of administering tests via computer. Educational Measurement: Issues and Practice, 8, 5-10.