DIFFERENTIAL ITEM FUNCTIONING IN THE PEABODY PICTURE VOCABULARY TEST – THIRD EDITION: PARTIAL CORRELATION VERSUS EXPERT JUDGMENT

A Dissertation

by

COLLEEN ADELE CONOLEY

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2003

Major Subject: School Psychology

DIFFERENTIAL ITEM FUNCTIONING IN THE PEABODY PICTURE VOCABULARY TEST – THIRD EDITION: PARTIAL CORRELATION VERSUS EXPERT JUDGMENT

A Dissertation

by

COLLEEN ADELE CONOLEY

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

Cecil R. Reynolds (Chair of Committee)	Maricela Oliva (Member)
Cynthia A. Riccio (Member)	Victor L. Willson (Head of Department)
Salvador Hector Ochoa (Member)	

December 2003

Major Subject: School Psychology

ABSTRACT

Differential Item Functioning in the Peabody Picture Vocabulary Test – Third Edition: Partial Correlation versus Expert Judgment. (December 2003)

Colleen Adele Conoley, B.A., Southwest Test State University

Chair of Advisory Committee: Dr. Cecil R. Reynolds

This study had three purposes: (1) to identify differential item functioning (DIF) on the PPVT-III (Forms A & B) using a partial correlation method, (2) to find a consistent pattern in items identified as underestimating ability in each ethnic minority group, and (3) to compare findings from an expert judgment method and a partial correlation method.

Hispanic, African American, and white subjects for the study were provided by American Guidance Service (AGS) from the standardization sample of the PPVT-III; English language learners (ELL) of Mexican descent were recruited from school districts in Central and South Texas. Content raters were all self-selected volunteers, each had advanced degrees, a career in education, and no special expertise of ELL or ethnic minorities. Two groups of teachers participated as judges for this study. The "expert" group was selected because of their special knowledge of ELL students of Mexican descent. The control group was all regular education teachers with limited exposure to ELL.

Using the partial correlation method, DIF was detected within each group comparison. In all cases except with the ELL on form A of the PPVT-III, there were no

significant differences in numbers of items found to have significant positive correlations versus significant negative correlations. On form A, the ELL group comparison indicated more items with negative correlation than positive correlation [χ^2 (1) = 5.538; p=.019]. Among the items flagged as underestimating ability of the ELL group, no consistent trend could be detected. Also, it was found that none of the expert judges could adequately predict those items that would underestimate ability for the ELL group, despite expertise. Discussion includes possible consequences of item placement and recommendations regarding further research and use of the PPVT-III.

ACKNOWLEDGMENTS

This dissertation would never have made it past the proposal stage without the hard work of my dear friend, Amanda Heidgerken. I appreciate her willingness to accompany me on several border trips to collect data. I would have never completed my data collection without her help.

I am also indebted to Cindy Gonzales and her parents Helen and Sam Gonzales. Mr. and Mrs. Gonzales were wonderful in opening their home to me and I appreciate Mrs. Gonzales helping me to gain access to Mission Independent School District.

Many other persons were major contributors to this process. I am grateful to all of the teachers, administrators, and children from Laredo, Mission, and Kerrville Independent School Districts for participating in my study. I also appreciate AGS allowing access to the standardization data. Without their approval, this study would not have been possible.

I would like to thank my committee members and my family for not giving up on me. I have been very lucky to have had so many wonderful mentors. On a daily basis I am reminded of how well I have been trained. Last, but not least, a special thanks to Richard Ziegler for constantly reminding me that my dissertation needed to be completed and sending me away for a week to finish writing it.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

The issue of bias in mental testing has a long history in the field of psychology. This history can be seen in the many chapters, reviews, and books dedicated to the topic (e.g., Camilli & Shepard, 1994; Cole & Moss, 1989; Kamphaus, 2001; Reynolds, Lowe, & Saenz, 1999). Most of these works have credited Binet around 1910 as the first to address the subject, referring to his question of whether he was testing "mental capacity" or environmental advantage.

This complex question has transformed itself considerably within the rather extensive literature base dedicated to the topic, as evidence has mounted to indicate differences in average performance between ethnic groups (Jensen, 1994). As stated here, Binet asked what construct was being measured. This question may not necessarily be an indication of bias. Since Binet, a number of explanations for these group differences have been offered and generally fall within four categories – genetics, environment, a combination of genetics and environment, and faulty tests (Brown, Reynolds, & Whitaker 1999). The fourth explanation – faulty tests – was investigated in the current study.

The issue of faulty tests has become more relevant in educational decisions and hiring and promotion practices since the Civil Rights Movement of the 1950s and 1960s

---

This dissertation follows the style and format of *Clinical Child and Family Psychology Review.*

(Jensen, 1980; Kamphaus, 2001; Reynolds et al., 1999). Court cases of this era attempted to bring equality to those groups considered to be in the political minority. As these issues were brought to the fore, research began to address and eliminate bias in mental testing. Of course, to do this, the notion of "bias" had to be clearly defined and possible sources of bias needed to be investigated (Jensen, 1980). In the following chapter this issue is addressed in greater detail.

Summary of Findings

Jensen's (1980) review of 20 years of research on test and item bias in mental testing revealed that well-constructed tests were not biased against native-born, English-speaking groups. Of particular relevance to this dissertation were findings regarding item bias or content validity. Jensen reported in *Bias in Mental Testing* (1980), or *BIMT* as commonly known in the literature, that when items were detected as being biased against a certain ethnicity (typically African American during this time frame), the number of items detected did not make a significant difference in overall score. Indeed, many times the number of items found to be biased against a focal group (e.g., African American) was also found to be biased against the reference group (i.e., white Americans). Jensen's conclusion regarding item bias has since been questioned by Camilli and Shepard (1987). They argued that his conclusion was primarily based on research using an inadequate method of detecting item bias, analysis of variance (ANOVA). Therefore, more research, using different methods, was needed to reexamine Jensen's conclusions.

Up until the mid 1980s, ANOVA was a popular method for detecting item bias. Since that time, it has fall out of popularity with researchers investigating item bias.

Brown et al. (1999) reviewed research since Jensen's *BIMT.* Their review provided continued support for Jensen's conclusions despite Camilli and Shepard's (1987) criticism of the use of ANOVA based methods to detect bias. Reynolds (2000a) directly addressed Camilli and Shepard's criticism of conclusions derived from ANOVA based methods. He stated that although Camilli and Shepard's article did demonstrate inadequacies of ANOVA for detecting item bias, other methods, including those thought to be superior by Camilli and Shepard (1994), also failed to demonstrate bias. However, although the lack of evidence regarding item bias had been clearly indicated by these authors, the need for continued research to ensure that items and tests were functioning similarly across ethnicities also was asserted.

Myths about Bias

Despite a large body of literature failing to identify bias against native-born, English-speaking minorities, certain myths regarding bias have continued to persist in policy and writings of many educational experts and psychologists (Reynolds, 2000b). Reynolds explored this phenomenon and suggested that these myths continue because of an inadequate knowledge base of tests and measurement, influence of the media, and/or the appeal of the "egalitarian fallacy" (defined by Jensen, 1980, as the belief that people are created equally in all respects, not just their value as human beings).

Regardless of the continued folklore of bias in mental testing, Reynolds (2000b) made a clear argument regarding the possible damage of accepting the notion of bias when bias does not exist. "For the racist people in our society, adoption of the cultural test bias hypothesis as true would be a major advantage, especially if it is false" (p.

148). He argued that large sums of money could be saved in additional educational support programs if the observed differences were only artifacts of testing. He argued that there would be no need to fund projects to investigate "nonexistent differences" (e.g., effects of lead, poverty, breast-feeding, maternal stress, etc.).

Although possible damaging effects of decision-making based on biased tests have been well-publicized through the media and policy statements (reviewed in Williams, Dotson, Don, & Williams, 1980), it has also been established that the assumption of bias where no bias exists could be just as damaging (Reynolds, 2000a). Therefore, it is important for research to continue to be conducted on commonly used tests regarding the possibility of bias.

Significance of the Problem and Rationale for the Study

Ethnic minority populations have continued to grow in public schools throughout the United States. In 1999, 61.9% of public education students were white, non-Hispanic; 16.5% were African American; 16.2% were Hispanic; and 5.5% were categorized as "other" (National Center for Education Statistics, 2001). Overall, in 1999 ethnic minority students totaled 38.1% of the student population; this was a 6.1% increase from 1989 (NCES). In addition to the increasing number of ethnic minority populations, numbers of linguistic minorities or English Language Learners (ELL) have increased. For the 1999-2000 school year the National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs (2000) reported an ELL enrollment of 4,416,580, a 104.97% increase from 1989. This growth of ELL is more astonishing when contrasted with the total enrollment growth of 24.21% from 1989

to 2000. As of 2000, ELL represented 9% of the public school enrollment in the United States. This growth in linguistic and ethnic minority populations has continued to show the need for continued research developing and examining assessment tools that provide an accurate representation of their levels of functioning.

Statement of the Problem

Both the assumption of bias where there is none, and failure to identify actual occurrences of bias can have undesirable consequences. Critics of mental testing have argued that biased tests will lead to missed opportunities (i.e., college admittance and job opportunities). Reynolds (2000a) explained that assuming bias when none is present can also have negative impacts such as denying access to special services or not identifying real problems. Although other methods of selection have been proposed, standardized, norm-referenced tests continue to be reliable predictors of school success, job performance (Reynolds, et al., 1999), and even outcomes of some medical procedures (Shapiro et al., 2000). Therefore, it is important that researchers continue to examine how tests and test items function for different groups.

While as indicated by numerous reviews of item and test bias, there have been no evidence of consistent or meaningful levels of bias in well-constructed tests (e.g., Brown, et al., 1999; Jensen, 1980). Nevertheless, lack of evidence these findings should not be interpreted as a reason to stop investigation in the area. Psychological standards created by the American Psychological Association (2002) state that it is the responsibility of the profession to administer unbiased tests. Since no particular set of items or content area has been found to be biased against a certain group (Reynolds et

al., 1999), no specific set of items can be identified to avoid in test construction. Therefore, it is necessary for new tests and items to continue to be thoroughly investigated for bias.

Purpose of Study

The Peabody Picture Vocabulary Test – Third Edition (PPVT-III) is a recently revised, single-word receptive vocabulary test. Previous editions have been among the most commonly used assessment tools used in schools (Stinnett, Havey, & Oehler-Stinnett, 1994).

Since item difficulty is closely related to the frequency of each word's typical usage, the test has been described as "culturally loaded" (Jensen, 1974). Therefore, both the original PPVT and its revised edition, the PPVT-R, have been the subject of previous test and item bias studies. The findings of these studies have yielded various results, depending on the version and the use of this test. Overall, little evidence of bias has been demonstrated by these studies. (These findings will be reviewed in the following chapter.) To date, item bias or "differential item functioning" (a term commonly used in the literature, because statistical methods do not indicate bias, only its potential) has been investigated in the PPVT-III only during test development/item selection using the Rasch method (a one-parameter IRT based method) and expert judgment (Williams & Wang, 1997). The current study was proposed to investigate differential item functioning on both current forms of the PPVT-III (Forms A & B) and to compare one sound method of detecting differential item functioning, partial correlation, with the commonly used method of expert judgment.

The current study had three goals. The first was to detect items that functioned differentially using a partial correlation method on both forms of the PPVT-III for each of three ethnic minority groups when compared to a focal group (i.e., whites in the standardization sample). The second goal was to attempt to find a consistent, meaningful pattern of items that functioned differentially within each group. The third goal was to compare the expert judgment method to the partial correlation method of detecting items that were likely to disfavor one of the comparison groups (i.e., ELL of Mexican descent), and to determine whether teachers with special knowledge about this group would be better than teachers without special knowledge in predicting items that were suspect of being biased. Based on previous finding in the literature the following was hypothesized:

Hypothesis 1

By using a partial correlation method to detect differential item functioning (DIF), items on both versions of the PPVT-III would be found to function differently for ethnic minority groups (i.e., focal groups) when compared to whites in the standardization sample (i.e., reference group). However, consistent with previous research (see Jensen, 1980 & Brown et al., 1999) the number of items identified as underestimating ability in a group would be similar to the number of items found to overestimate ability in a group. That is, the difference between the number of items functioning for a group and against a group would be stastically insignificant.

Hypothesis 2

Consistent with previous findings (Reynolds et al., 1999), no consistent pattern would be found across items identified as underestimating ability for any focal group.

Hypothesis 3

The expert judgment method would not adequately predict those items detected by the partial correlation method as underestimating ability of ELL through simple item examination. Furthermore, those teachers with special knowledge about ELL of Mexican descent ("experts") would not be better at predicting biased items than those without special knowledge ("non-experts").

CHAPTER II

REVIEW OF LITERATURE

Brief History of Test Bias

Questions regarding group differences in scores on standardized tests have been present for nearly a century. Binet became concerned when his test, which was developed in 1905 to measure innate ability, demonstrated significant group differences in overall scores among social status groups (Binet & Simon, 1911/1979). These differences were found across multiple countries, including the United States. Binet began to speculate that certain items were measuring cultural training, such as schooling and home environment, rather than innate ability. He eliminated categories of items that he believed to be dependent upon social standing; however, this attempt yielded little difference in the between group discrepancies of overall scores. Since this time, numerous researchers also have reported group differences in mental test scores and identified possible sources of group differences (Jensen, 1980).

Standardized Testing and Court Cases

Questions regarding bias and the validity of mental testing as related to diagnosis, placement, and selection became a topic of popular media, litigation, and debate during the Civil Rights' Movement (Camilli & Shepard, 1994; Jensen, 1980; & Reynolds et al. 1999). The Civil Rights' Movement of the late 1960s and 1970s focused on equality in occupational and educational opportunities. Since mental testing was and continued to be an important component in personnel selection/promotion and

educational placement, mental testing itself became a target of controversy, which at times was addressed in court disputes.

Hobson v. Hansen (1967) was the first court decision in which the appropriateness of standardized testing for use of minority placement was questioned (Jensen, 1980). Children in Washington D. C. schools were being placed into "ability tracks" based on standardized testing performance. This practice resulted in African American children being disproportionately placed in low "ability tracks." The plaintiffs argued that this placement was inaccurate and limited their children's opportunity to learn. Judgment was made in favor of the plaintiffs. The tracking system was found to be discriminating along racial and socioeconomic lines rather than ability to learn, thus, resulting in unequal educational opportunity. The presiding judge's rationale was that although ability tests tended to measure some constant trait within children, this trait was dependent on each child's previous learning opportunities and did not necessarily predict his or her ability to learn. Therefore the tracking system was abolished in Washington, D.C.

Diana et al. v. State Board of Education (1970) was another court case that challenged testing practices used to classify children. In the case of Diana et al., Spanish-speaking children of Mexican descent were qualified as "Educable Mentally Retarded" (EMR) based on their performance on an intelligence test administered in English. According to Jensen (1980), there was a disproportionate number of children of Mexican descent classified as EMR – three percent of the Mexican American children to one and a half percent of the white children within the district. This case was settled out of court.

The settlement stipulated that 1) children were to be tested in their primary language, 2) nonverbal items were to be used in the place of "unfair" verbal items, 3) bilingual children who were already qualified as EMR were to be retested with these new procedures, 4) state psychologists were to develop an IQ test in Spanish and normed with Mexican American children, 5) and any district with disproportionate EMR placement among races would need to submit an explanation for the disparity.

Larry P. et al. v. Wilson Riles et al. (1979) and *Parents in Action on Special Education (PASE) et al. v. Joseph P. Hannon (1980)*, were two other federal court cases concerned with determining whether individually administered intelligence tests were appropriate tools for the assessment of African American children. In both cases the assessment method under question was used to place children in either EMR or "Educable Mentally Handicapped" (EMH) classes. Although the major issues of the two cases were very similar, the decisions made by the two judges were different (see Sattler, 1981b, for an in depth analysis of the judgments).

In Larry P. v. Riles, Judge Peckham ruled that intelligence tests were culturally biased; Judge Grady in PASE et al., v. Hannon ruled that qualification decisions based partially on intelligence test performance were not culturally biased. A major point of disagreement involved the judges' understanding of how children were selected for special classes (Sattler, 1981b). Judge Peckham stated that the placement process was based on intelligence testing performance. In contrast, Judge Grady described the placement process as was mandated by Public Law 94-142, and involved multiple sources of information (e.g., prereferral teams). They also disagreed on the impact

"Black English" (BE) had on scored items. Peckham believed that BE negatively impacted the child's performance, while Grady stated that children were not penalized for nonstandard grammar on verbal sections. Both judges selected items they believed to be racially biased. Peckham selected items that had been questioned in the literature. Judge Grady, however, dismissed the expert testimony, and listed items he felt were biased based solely on his own opinion. Grady stated that these items should not be used with minority children; however, he ruled the number of biased items to be insignificant and that the remainder of the test was not biased against ethnic minorities.

Golden Rule Insurance Company et al. v. Washburn et al. (1984) was a case in which an insurance company alleged that the impact of the state insurance licensure examination was discriminatory because African American examinees failed at a higher rate than the white examinees. As part of the settlement, Educational Testing Service (ETS) agreed to make certain modifications in their test construction. They agreed to select items for which failure rates were no less than 40% for either white examinees or African American examinees, and item difficulties (i.e., *p*-values) differed no more than 0.15 between groups when such items were available. In a similar settlement, *Allen et al. v. Alabama State Board of Education (1985)*, it was decided that items would be selected if item difficulty between African American and white examinees differed no more than 0.05. This case also was settled out of court. Interestingly, both of the cases made allowance for actual differences between the groups by allowing some degree of difference in item difficulties or *p*-values; however, the amount of discrepancy allowed did not entirely resolve the problem of confounding item discrimination (Camilli &

Shepard, 1994). The impact of these judgments, as demonstrated by Linn and Drasgow (1987) in their analysis of the impact of the Golden Rule ruling, only resulted in the inclusion of items with weaker discrimination power leading to less reliable scores. This consequence was unfortunate, as the primary purpose of a test is to discriminate; indeed, a test would be useless if it did not (Jensen 1980). Also, paradoxically, the less reliable a test, the more likely it is to be biased (Linn & Werts, 1971).

These court cases demonstrated differing conceptualizations of fairness and intended purposes of mental testing. In Hobson v. Hansen, judgment was made based on the idea that performance on intelligence tests was dependent on environmental advantages and performance on intelligence tests was not necessarily predictive of ability to learn. In Diana v. State Board of Education, an out-of-court settlement was based on the decision that testing a child in a language she did not adequately understand was unfair. Judges presiding over Larry P. v. Riles and PASE v. Hannon evaluated an intelligence test at item level and made differing qualitative judgments regarding the fairness of the test. Golden Rule Insurance Company v. Washburn and Allen et al. v. Alabama State Board of Education settlements were based on the concept that item difficulties should not significantly differ for two groups. Consistent with Jensen's (1980) criticism of early bias studies, the verdicts in these court cases further demonstrated a difficulty with poorly defined concepts of bias.

Research in Bias

Jensen (1980) credited the first investigation of cultural bias in standardized tests in the United States to Davis and Eells in 1945. Jensen's review of studies following Eells' "pioneering study" found that terminology of "bias" and "unfairness" lacked conceptual clarity. According to him the term "cultural bias" was used to describe differences in mean levels of performance across socioeconomic status and ethnic groups' test scores. In fact, he claimed those who proclaimed "bias" against ethnic minorities and those of lower SES contributed the least to developing a working definition of "bias."

Definition and Terminology

Most would agree that developing a definition of "bias" was essential to studying it accurately. Yet creating a definition of the term bias has produced considerable debate (see Frisby, 1999; Jensen, 1980; Reynolds et al., 1999). Reynolds et al. clearly made an essential distinction between questions regarding actual test bias and bias in the manner in which tests have been used. The former was a question of test score reliability and validity, while the latter question was focused on the decision-making process in which tests may be used rather than on the test itself. They suggested that prior to selecting a decision-making system, the ultimate goal must first be identified, whether it was equality of opportunity, equality of outcome, and/or representative equality. Reynolds et al. stated that the best way to ensure unbiased selection in any decision-making system based on these goals was to use test scores that were equally reliable and valid for each group being evaluated.

Jensen (1980) presented three inadequate concepts of test bias that have continued to be important to conceptualizing bias (Brown et al, 1999; Reynolds et al., 1999; Reynolds 2000a): the egalitarian fallacy (i.e., all human populations are equal in an underlying trait), the culture-bound fallacy (i.e., items can be subjectively screened for their cultural-boundness, as in Larry P. v. Riles and PASE v. Hannon), and the standardization fallacy (i.e., tests are biased against any group other than those for which they were normed). In addition to these inadequate concepts, Jensen also cautioned against confusion of terms and reminded his readers that "discrimination" in a measurement sense means a reliable difference and that all tests discriminate and would be useless if they did not. In defining "bias," he also distinguished it from "fairness." He defined bias according to mathematical statistics in a purely objective, measurable manner: "'bias' refers to the *systematic* under- or overestimation of a population parameter by a statistic based on samples drawn from the population." He cautioned that biased tests could be used fairly, just as unbiased tests could be used unfairly.

By using a definition similar to the one presented by Jensen (1980), bias became a testable concept. As presented by Reynolds et al. (1999), bias in testing traditionally has been conceptualized as related to predictive/criterion validity, construct validity, and content validity. Predictive validity has been used to describe how performance on a test is predictive of performance or status on an external criterion. The difficulty with determining test bias with predictive validity methods has been that finding a measurable, unbiased, external criterion has been difficult (Camilli & Shepard, 1994). Construct validity has been used to describe how well a test measured theoretical traits.

Methods for assessing bias as related to construct validity have frequently used factor analysis; however, there has been no single method designated as adequately determining bias in a construct validity sense (Reynolds et al. 1999). The final category and most closely related to this study has been described as content validity:

> An item or subscale of a test is considered to be biased in content when it is demonstrated to be relatively more difficult for members of one group than for members of another when the general ability level of the groups being compared is held constant and no reasonable theoretical rationale exists to explain group differences on the item (or subscale) in question" (Reynolds et al., 1999 p.564).

Similar to the difficulty addressed regarding bias detection with construct validity, no single statistical method has been found to identify items accurately as biased.

Because no single method has been found to identify item bias accurately, Camilli and Shepard (1994) warned that detecting item bias by internal methods should involve more than statistical computations; professional judgment also should be considered. Professional judgment as discussed here should not be confused with expert judgment. In this context, professional judgment occurs post statistical computations in an attempt to explain why a particular item functioned differently for a particular subgroup. As Camilli and Shepard (1994) reviewed methods of detecting bias, they explained how several methods were capable of falsely identifying bias, as well as missing real occurrences of bias. Therefore the term "differential difficulty" was used to describe earlier methods derived from classical test theory and "differential item

functioning" (DIF) was used to describe methods derived from item response theory (IRT) and methods presented as contingency tables (CT).

Throughout the literature base the terms "item bias" and "DIF" have at times been used interchangeably. However, for the remainder of this dissertation the term "DIF" has been used to describe internal methods attempting to identify bias based on content validity. This distinction is made because as explained in the next section, none of the internal methods of identifying bias are perfect predictors of bias. At best, they only flag suspicious items and may miss actual occurrences of bias. This distinction is important so that readers are not mislead into assuming bias from DIF.

Content Validity Methods

Early Methods

Earlier methods of detecting item bias looked for differential difficulty as an indicator of bias. These methods of detection looked for items in which the difference between item difficulties for two groups was significantly different when compared to other items. Conceptually, these methods were similar to the settlement reached in Golden Rule Insurance Company et al. v. Washburn et al. and Allen et al. v. Alabama State Board of Education. The assumption was that bias produced different results for different populations (Camilli & Shepard, 1994).

Golden Rule. The Golden Rule procedure was named from *Golden Rule Insurance Company et al. v. Washburn et al.(1984)* from which it was developed. As previously stated, the Golden Rule procedure was to be performed during test construction. As part of the settlement, items were permitted if they met two criteria: 1)

between group *p*-value differences were less than 0.15 and 2) items with *p*-values were greater than 0.40 for both groups. This method reduced test reliability and predictive validity (as demonstrated by Linn & Drasgow, 1987). Thus, Camilli & Shepard (1994) recommended against using the Golden Rule procedure.

Transformed Item Difficulty and Delta Plots. According to Angoff (1982), transformed item difficulty (TID) or delta plots, were methods in which items were flagged if they exaggerated the difference between groups. Items with differences in item difficulty greater than was typical among the other test items were suspected of bias. This technique involved converting 1 – *p*-values to a standard score. Converting the difference to a standard score removed the curvilinear relationship between the sets of values and minimized the ceiling and floor effects of difficult and easy items (Camilli & Shepard, 1994). Angoff used the "delta" scale (μ= 13 and σ=4) and then plotted the scores on a bivariate plot. A simpler example, using *z*-scores, was offered by Camilli and Shepard. If all items on a test had a similar relative difficulty for two groups, then the plot would exhibit a 45° trend-line. Deviations from this line or outliers were thought to indicate bias. The TID index was calculated for each item as the perpendicular distance of each item from the trend-line. Items with significantly greater distances, in comparison to the other items, were identified as those with the greatest differential difficulty, and thus suspect of item bias.

Camilli and Shepard (1994) described this approach as conceptually appealing and easy to calculate; however, they criticized this approach as continuing to confound group mean differences with item discrimination. When two groups differed or were

unequal on an underlying trait, highly discriminating items would yield large discrepancies and, thus, appear biased. In addition to indicating bias erroneously (Type I error), Camilli and Shepard (1987) demonstrated how this method could miss real occurrences of bias (Type II errors) with a demonstration using item characteristic curves. In their demonstration, differences in difficulty occurring in regions other than between the means could obscure the differences and perhaps miss real instances of bias. Camilli and Shepard (1994) admitted that this occurrence would be rare in practice. It would mean that, at comparable ability levels, the reference group was more likely to answer an item correctly at a given level (e.g., low ability level) than the focal group, but the focal group was more likely to answer an item correctly at a different ability level (e.g., high ability level) than the reference group, then the differential difficulties would be missed. If visually displayed with item characteristic curves (explained later), it would be a scenario in which the curves crossed.

Adjustments were made to remedy the problem of item discrimination confounding differential difficulty. Angoff (1982) suggested matching groups with a relevant, external criterion prior to conducting the analysis. However, the practical implications of this method were quite difficult. An external criterion had to be established that was less biased than the test being analyzed (this proposal was similar to the criterion/predictive validity methods reviewed by Reynolds et al., 1999). Angoff also suggested using item-test correlations to provide a measure of the item's discriminating power. TIDs were divided by their item-test correlation. According to Camilli and Shepard (1994), the item-test correlations used to estimate an item's discriminating

power were point biserial correlations. Angoff (1982) and Camilli and Shepard (1994) acknowledged that point biserial correlations were notorious for being unreliable (related to the sample and the sample size to number of items ratio) and therefore Angoff's adjustments were unstable. Shepard, Camilli, and Williams (1985) compared the adjusted TID index and the TID index with an item response theory model and found that the original TID index was more consistent with the item response theory model than the adjusted index.

Analysis of variance. Analysis of variance (ANOVA) was one of the most commonly used methods for detecting item bias until the late 1980s (Camilli & Shepard, 1987). Consistent with the previously described methods, items with differential difficulties were identified as biased. With this method, item bias was indicated by a significant group-by-item interaction (Angoff & Ford, 1973). When Jensen published *Bias in Mental Testing* (1980) a large number of item bias studies had been performed using an ANOVA method to detect item bias. It was written that this method was "powerful"; the only way that it would miss actual occurrences of bias would be if bias were consistently demonstrated across all items in a test that would affect the items in the same way (Jensen, 1984).

Camilli and Shepard (1987) were unimpressed with this methodology and stated that criticisms of ANOVA had been rather "one-sided." As in the TID methods, methods using ANOVA could indicate false instances of bias in highly discriminating items. Camilli and Shepard (1987) also demonstrated that methods using ANOVA could miss real incidents of bias. They argued, "differential difficulty contributes more to the

between-groups effect than to the interaction" (p. 88), therefore potentially missing real occurrences of bias. Camilli and Shepard (1987) offered a heuristic demonstration with item characteristic curves, an algebraic demonstration, and a simulation with contrived data. Camilli and Shepard (1987) concluded that their simulation and equations did not provide evidence that bias existed in mental testing; however, it was their opinion that previous studies using ANOVA were to be disregarded and ANOVA as a method of detecting item bias should no longer be recommended.

Item Response Theory. Models of DIF based on item response theory (IRT) related the probability of a particular response on an item to overall examinee ability (Camilli & Shepard 1994). Ability or latent trait (θ) was defined as the construct that the test was attempting to measure. The principle unit of IRT was the item characteristic curve (ICC), which was a function that related the probability of a correct answer on an item to the ability measured by the test. Within the ICC, item discrimination (the *a* parameter), item difficulty (the *b* parameter), and the probability of a correct response from an examinee with low ability (the *c* parameter or the guessing parameter) were all represented.

Camilli and Shepard (1994) presented IRT as a gold standard of the statistical methods used to flag items as potentially biased. An advantage of IRT methods was that estimates of parameters (*a, b,* & *c*) were less confounded with sample characteristics than were those of classical measurement theory. Also, IRT allowed for DIF to be described more precisely, and the statistical properties of items could be more readily graphed with the IRT approach than with classical measurement theory.

As explained by Camilli and Shepard (1994), depending upon the format of the test items (i.e., multiple-choice vs. free-response) different equations could be used for detecting DIF, three-parameter, two-parameter, or one-parameter models. IRT methods were highly dependent on model selection and parameter selection; complex statistical programs helped with this process. The three-parameter model was the most commonly used of the three and was recommended for multiple-choice test items, so that the probability of an examinee guessing the correct response should be. The two-parameter model was useful in detecting DIF in free-response items. In free-response items the *c* parameter was equated to 0 and therefore was dropped from the equation resulting in a two-parameter model. One-parameter IRT models, when they accurately described the data, provided the most sensitive tests for DIF enabling them to compensate for incomplete data, which translated to requiring less data or smaller sample sizes. However, Camilli and Shepard advised against the one-parameter or Rasch method in detecting DIF with multiple-choice items because guessing is typically present. The benefit of three-parameter models has been in their generality.

There have been several three-parameter methods. Of these, Camilli and Shepard (1994) recommended the probability difference indices to measure DIF and the item drift method as a test of DIF. Unfortunately these three-parameter IRT methods were computer intensive and required a rather large sample size (Camilli & Shepard, 1994; Reynolds et al., 1999). Even Camilli and Shepard (1994) indicated that it is often hard to justify the resources needed for the three-parameter models in applied settings, but continued to recommend these models for research.

Beyond the complex statistics and large sample sizes required, IRT methods also have been criticized conceptually. Hunter and Schmidt (2000) argued that even in IRT methods, total test score has been used to estimate the underlying trait and therefore they have made the assumption that the total test was an errorless measure, and as in other measures of DIF, IRT has assumed unidimensionality of the measure. The following two methods more directly used total score as an estimate of underlying trait and also assumed unidimensionality.

Contingency Table. The term contingency table was used by Camilli and Shepard (1994) to describe the manner in which these methods of detecting DIF could be tabulated. Within this section they described methods for measuring DIF, proportion difference measures and Mantel-Haenszel (MH) Log Odds Ratio, and tests of DIF including the summed chi-squared method, the MH chi-square method, and a technique of logistic regression. The major advantages of the CT approaches over IRT approaches have been sample size and its easy implementation (Camilli & Shepard, 1994). When faced with smaller samples, Camilli and Shepard (1994) recommended the Contingency Table (CT) approaches for measuring DIF. These approaches did not require computer-intensive analysis.

CT approaches required smaller sample sizes because total score was used to estimate ability. Comparisons were made on each item for each total test score. No provision was made for guessing, and no provision is made for variation in discriminating power. Thus, according to Camilli and Shepard (1994), the weakness of CT approaches was that strong assumptions were made – guessing and discrimination

were assumed to be the same for any two groups on each item. Incorrect assumptions led to both Type I and Type II errors.

Partial Correlation. Another approach to measuring DIF was a partial correlation method that correlated item response and group membership, with total test score partialled out of equation. This technique was comparatively recent to the aforementioned methods. Darlington (1971) first argued that a test or an item could be consider biased if it had a significant partial correlation with subgroup standing (ethnicity, gender, or SES) when the criterion (i.e., test score) was held constant.

Stricker (1982) first used the partial correlation index to detect DIF, as Reynolds, Willson, and Chatman (1984) also independently developed the method. As presented by Stricker, the index was operationalized by the following formula (p. 263):

$$\mathbf{r}_{iS \cdot T\infty} = \frac{\mathbf{r}_{iS} - \mathbf{r}_{iT\infty}\,\mathbf{r}_{T\infty S}}{\sqrt{1 - \mathbf{r}^2_{iT\infty}}\;\sqrt{1 - \mathbf{r}^2_{T\infty S}}}$$

In this formula "r_{iS}" represented the correlation between the item responses and subgroup standing. The correlation between the item response and the total score, adjusted for item overlap and corrected for attenuation in the score was represented by "$r_{iT\infty}$;" "$r_{T\infty S}$" represented the correlation between the total score and subgroup standing. All correlations in this formula were product-moment correlation coefficients.

Stricker (1984) indicated many advantages of this method over the previously mentioned methods. In comparison to the IRT method, the partial correlation index was

far less costly in computer software and running time. Also, this partial correlation method requires a much smaller sample size, only 300 per subgroup. The partial correlation method was flexible in that any number of different subgroups could be accommodated simultaneously; it provided for a significance test and permitted a straightforward evaluation of effect size.

Although the method of detecting DIF through partial correlation was not frequently found in empirical literature, this method has been compared to other methods of detecting DIF: ANOVA (Reynolds, Willson, & Chatman, 1984), and ICC (three-parameter model) and item difficulty index (Stricker, 1982). In all comparisons, partial correlation functioned as well or better than the comparative methods (see Valencia, Rankin, & Livingston, 1995; Willson, Nolan, & Reynolds, 1989 for additional examples of partial correlation in DIF).

Distractor Analysis. Distractor analysis was a technique that inspected multiple-choice items by determining which distractors (incorrect choices) were more attractive to particular subgroups (for a more detailed explanation, see Scheuneman, 1982). According to Veale and Foreman (1983) this method was based on the idea that more can be learned from incorrect than correct responses. The assumption was that incorrect responses on a multiple-choice exam were not picked randomly. Instead, there was some sort of logic applied by the examinee in attempting to pick the correct answer.

Distractor analysis could have been used independently for bias detection or posteriori analysis. A major benefit of this approach over previously mentioned approaches was that it did not require the majority of the remaining items to be unbiased

(Veal & Forman, 1983). Another benefit of this approach was that it gave more insight into why an item was biased. This insight allowed the item to be modified rather than discarded completely. A major concern with distractor analysis as a method for detecting bias was that distractor attractiveness was also a function of examinee ability (Scheuneman, 1982). An examinee without any information on a topic may have randomly picked (i.e., breaking the underlying assumption of this technique); whereas, an examinee with limited knowledge might have employed a narrowing tactic. Therefore, it was also suggested, as in other methods, that distractor analysis be followed by careful item review. Also, this method is limiting given that most individually administered tests of intelligence and achievement are open-ended item formats and distractor analysis does not apply to such tests.

Expert Judges. The method of using expert judges involved selecting individuals who represented a gender group and/or a different ethnic minority group or had special expertise about a particular group. These individuals were asked to rate items on their offensiveness or likelihood of being biased against a particular group (Tittle, 1982). Obviously, this method was appealing to test developers in contrast to the more sophisticated and costly statistical methods (Plake, 1980). Its appeal has made it a widely used technique for test developers and was used in item selection on the PPVT-III to eliminate items thought to be offensive (Williams & Wang, 1997).

Although the use of expert judges has been useful in identifying items containing sensitive content and in increasing popular acceptability of a measure, empirical literature failed to support its utility as a method of detecting items that were biased

against a particular group (Camilli & Shepard, 1994). For example, Plake (1980) attempted to validate this method by comparing expert opinion to items detected by an ANOVA method for detecting DIF on the Iowa Test of Basic Skills. The results of her study found little relationship between items detected by her experts and items detected statistically.

Sattler (1981b) also became interested in this method when two federal judges deemed themselves experts and attempted to detect bias. As previously mentioned, both federal judges in *Larry P. et al. v. Wilson Riles et al. (1979)* and *Parents in Action on Special Education (PASE) v. Joseph P. Hannon (1980)* declared specific items from the Wechsler Intelligence Scales for Children to be "biased" against African American students. Sattler tested the validity of the judges' decisions by comparing item difficulty of African American performance to white performance. Results of this study found that the of the 11 items that the judges thought to be "biased" against African Americans, only six were more difficult for the African Americans group than for the white group. Sattler's (1981a) study also demonstrated that there were six additional items that were more difficult for African American subjects than the white subjects. Reynolds et al. (1999) briefly reviewed this "armchair analysis," reporting that although it had been shown to sort items at a level no better than chance, expert judges continue to be commonly used by test developers.

Summary. As presented in this section, a variety of methods have been proposed for detecting potentially biased items and each of these methods possesses its own strengths and weaknesses. Although the partial correlation method also has its own

strengths and weaknesses, it was chosen for this study three reasons: 1) This method allows for comparison of small sample sizes; 2) it attempts to control for ability level by partialling total score from the equation and therefore should not be as affected by large differences in underlying trait as older methods; and 3) it is not as affected by differences in subgroup size as older methods.

Research Findings of Item Bias

Reynolds et al. (1999) reported that despite countless studies attempting to find a consistent pattern of biased items, no pattern or type of item has been determined to be biased against any ethnic, cultural, or gender group. Even when early studies were developed to test hypotheses such as verbal items or items that required previous knowledge as being biased against African American examinees, no such trend could be found (Jensen, 1980). Bruce (1940) had psychologists classify 34 items of the Kuhlman-Anderson Intelligence Scale into three categories: 1) questions that required a previous knowledge base to answer correctly ("information"), 2) questions that required problem solving skills to answer correctly ("new situation"), or 3) questions that required both previous knowledge and problem-solving skills ("hybrid"). Results demonstrated that the African American sample performed similarly across all item types. McGurk (1975) reviewed 18 studies between 1951 and 1970 investigating differences in white versus African performance on verbal and nonverbal items. Findings of these articles found greater group differences on nonverbal than on verbal items. However, Jensen (1974) found that these effects disappeared when examinees are matched by mental age (i.e., intelligence).

There have been multiple complications in finding a source or pattern of biased items. One complication in finding a consistent pattern of biased items has been a problem common to multi-ethnic/multicultural research: within group differences were larger than between group differences. Jensen (1980) went so far as to emphasize that more than three times as much variance attributable to race and socioeconomic status combined could be attributed to between family differences of the same ethnicity and socioeconomic status. He further reported that the largest source of total variance was actually between siblings.

Another complication has been that when items have been identified as being biased or underestimating a particular group's ability, the number of items detected has been too small to find a consistent, interpretable pattern (Valencia, 1992). For example, when Reynolds, Willson, and Chatman (1984) attempted to find item bias on the both forms of the PPVT-R (i.e., form L and form M) with a partial correlation method, only a minimal number of the total items were detected as underestimating ability for the African American sample. Eleven items on form L and one item on form M were found to underestimate ability in the African American sample. Although the 11 items detected on form L as functioning against the African American sample was statistically significant when compared to the three items detected as underestimating ability for the white sample, the 11 items detected were insufficient for finding a trend of bias. The authors were not able to determine a trend of bias by analyzing these 11 items. Another example of using the partial correlation method to identify items with the potential of bias against African American children was Willson et al.'s (1989) study of differential

item functioning on the Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983). The K-ABC is composed of 10 different subtests and the items in each subtest are different from items in the other subtests regarding administration, response style, and mental processes that each subtest has been constructed to measure. The 23 items flagged as being potentially biased against the African American sample were scattered among nine subtests. Therefore, it was difficult to find a consistent trend of items that underestimated ability for African American children. However, on one subtest, Gestalt Closure, requiring a child to identify a degraded, black and white drawing (Kaufman & Kaufman, 1983) eight items were identified as being potentially biased against African American children. Although recommendations were made to further investigate Gestalt Closure, the overall effects, as in other studies of item bias, found no trend of bias.

There have been occasional studies that have found a large proportion of the item on a test to flagged for potentially being bias against a particular group. For example, Stricker (1982) compared three methods of identifying DIF in the GRE, including partial correlation, comparisons of subgroups' ICCs, and item difficulties. He found that the partial correlation index identified almost one-half of the items; although most indices were small in absolute size (i.e., less than 0.10). The ICC curves identified about a quarter of the items as being significant and the difficulty index identified less than a tenth of the items. This study presents another difficulty in detecting a consistent pattern of bias, the methodologies for detecting item bias.

As previously discussed, there has been and continue to be a variety of methods used to detect differential item functioning. Each has been demonstrated to have its own strengths and weaknesses and each provided different information. Therefore, different items could be detected by using different methods, and as stated by Camilli and Shepard (1994) an item that has been identified statistically as being suspicious does not indicate bias.

Peabody Picture Vocabulary Test – Third Edition

Originally developed in 1959, then revised in 1981, and more recently revised in 1997, the PPVT-III (Dunn & Dunn 1997) was designed as an individually administered, multiple-choice, single-word, receptive vocabulary test that required a nonverbal response. This test was been designed to measure receptive vocabulary and screen for verbal ability. Its two parallel forms each contain 204 items arranged in increasing difficulty and were designed for use with individuals from ages 2 ½ years to 90+ years. Administration time is relatively short, averaging 11 to 12 minutes. In order to respond to each item, the examinee is asked to point or say the number of one of four black and white pictures.

Item Bias on Previous Versions

As mentioned in Chapter I, the previous two versions of the PPVT-III (i.e., PPVT & PPVT-R) have been popular subjects of item bias research. The first version was frequently used a brief IQ measure and the PPVT-R continued to be used frequently throughout the 1980s and into the early 1990s as a measure of intelligence (Stinnett et al.

1994); although the IQ had been dropped in the revised edition to discourage its use as an intelligence test (Dunn & Dunn, 1981).

The earlier studies of bias in the PPVT compared it to other measures of intelligence. For example, Jensen (1974) compared the original version of the PPVT to the Raven's Colored Progressive Matrices, a nonverbal intelligence test with various methods of bias detection. Jensen was interested in comparing a "culture loaded" test to a "culturally reduced" test. The PPVT was seen as culturally loaded because item difficulty was highly correlated with frequency each word was used in common language and the Raven's Colored Progressive Matrices was viewed as being "culturally reduced." Jensen found that the African American sample performed similarly on both measures. However, he did find Mexican American Children to score significantly lower on the PPVT than on the Raven's Colored Progressive Matrices. This finding was attributed to the Mexican American children being bilingual. Neither the group-by-item interaction in analysis of variance nor in the item distractor analysis indicated bias.

Halpin, Simpson, and Martin (1990) used step-down hierarchical multiple regression procedures to investigate bias in predicting African American and white performance on the Wechsler Intelligence Scale for Children – Revised (WISC-R). Their findings failed to indicate bias for either group. Bracken and Prasse (1981) also compared total score means of the PPVT and PPVT-R to various tests of intelligence for white, Hispanic, and African American groups. Their finding did not support bias, but did suggest that neither the PPVT nor the PPVT-R should be used in the place of intelligence testing. Others also have recommended against using the PPVT for other

purposes than a receptive vocabulary test or language screen (e.g., Altepeter, 1989; Altepeter & Handal 1985; Dunn & Dunn, 1997; Maxwell & Wise, 1984; Strein & Ysseldyke, 1974).

Reynolds et al. (1984) compared a partial correlation to an ANOVA method of detecting DIF with African American children. Results of that study did not find any significant items with the ANOVA. However, with the partial correlation method form M was suggested over form L for African American populations. On form L 11 items were found to favor the white sample and only three were found to favor the African American sample. In comparison, on form M only one item favored the white sample and three favored the African American sample. Argulewicz and Abel (1984) used ANOVA to find item-by-group interactions with Mexican American children when compared to a white sample. The effects were small and the study concluded that neither form was biased against either group. None of these studies were able to determine a trend or category of items that consistently functioned differently for a particular group.

Bias and PPVT-III

A panel of six consultants, representing the perspectives of Asians, African-Americans, Hispanics, Native Americans, and women were asked to review the Peabody Picture Vocabulary Test – Revised (PPVT-R) items to identify offensive or biased material (Williams &Wang, 1997). These identified items were removed from the PPVT-III item pool. In addition to this panel, the Rasch model was used to identify biased items during the item tryout phase of the PPVT-III. The items identified by the Rasch model were also dropped from the national tryout pool.

In addition to the studies conducted during test construction, Washington and Craig (1999) conducted a study to determine whether the PPVT-III form B was biased against low SES African American children who speak "African American English." Children were selected from a preschool in Detroit (all children except four were low SES and African American). Washington and Craig compared the PPVT-III scores with assessments from the speech-language pathologist. The language assessments involved a Wh-Question Comprehension task with two pictures taken from the Bracken Concept Development program (Bracken, 1986). Their conclusion was that the PPVT-III was not biased against low SES African American children and they recommended its use with this population.

Summary

The question of bias has existed for a long time and ideas about what constitutes bias have been discrepant. Of most relevance to this study is item bias. Given that there is no statistic that adequately detects item bias, the term DIF (differential item functioning) has been chosen because it more accurately describes what is identified with these statistics.

The following study exams differential item functioning in the PPVT-III with a partial correlation method. This method was thought to be the best given the data available. To date, DIF on the PPVT-III only has been investigated during test development with a one-parameter IRT model, i.e., the Rasch method. Since different methods yield different results, the partial correlation method was used to detect DIF within each ethnic group comparison. As described in the following chapter, three

groups were provided from the standardization sample: African American, Hispanic, and white. A fourth sample was collected was collected by the author. This sample was composed of English language learners, whose first language was Spanish. To make this group more homogenous, only children of Mexican descent were included in this fourth group.

To date, no consistent pattern of items has been identified as being biased against a particular group. By the methods described in this and the following chapter, items with significant partial correlations are suspected of bias. An attempt has been made to determine a pattern among the suspicious items.

In addition to internal statistics used to identify DIF, another method is commonly used in test development. This method requires "experts" to identify items that may be potentially biased. Currently, this method has not been demonstrated to identify items at a level better than chance. This method will be reexamined with teachers, who have daily contact with these students.

In the following chapter, the methods used to conduct this study are presented. Included in Chapter III are the participants, procedures, and instrumentation.

CHAPTER III

METHODS

Participants

Normative Sample

Data for the Hispanic, African American, and white subjects for the study were provided by American Guidance Service (AGS) from the standardization sample of the PPVT-III (Dunn & Dunn, 1997) (as in Reynolds, Willson, & Chatman, 1984). Sample sizes are summarized in Table 3.1 and chronological ages are summarized in Table 3.2. Descriptive statistics regarding performance on both versions IIIA and IIIB have been provided in Tables 3.3 and 3.4.

English Language Learners

Three hundred children, who were English language learners (ELL), as indicated by their school districts, were recruited from school districts in Central and South Texas. The children selected for the study met six criteria: (1) they were designated as ELL by their school districts and received ELL support (e.g., English as a Second Language, two-way emersion, etc.); (2) Spanish was their first language; (3) they were attending first, second, or third grade; (4) they were of Mexican descent, (5) they could speak conversational English as demonstrated during informal conversation based solely on examiner discussion; and (6) their parents had signed a consent form. Each school provided dates of birth of ELL participants. A comparative white sample was formed by restricting the age range of the normative sample to 6 to 10 years, which approximated the age range of the ELL group and was designated as “Restricted-age white.” Sample

sizes and ages of these two groups have been presented in Tables 3.1 and 3.2, respectively. Performance on the PPVT-III has been provided in Tables 3.3 and 3.4.

TABLE 3.1. Sample Size

Groups	Sample Size
African American	494
Hispanic	352
White	1,753
ELL	300
Restricted-age White	349

TABLE 3.2. Group Chronological Ages in Years

Groups	Mean	Standard Deviation	Minimum	Maximum
African American	13.88	14.60	2	81
Hispanic	12.98	10.76	3	75
White	19.07	18.64	2	91
ELL	7.20	1.01	5	11
Restricted-age White	7.75	1.44	6	10

TABLE 3.3. Raw Score Points – PPVT-IIIA

Groups	Mean	Standard Deviations	Minimum	Maximum
African American	99.79	54.967	4	195
Hispanic	111.7	51.846	4	195
White	130.17	54.850	2	203
ELL	73.99	23.896	3	150
Restricted-age White	109.48	24.245	49	181

TABLE 3.4. Raw Score Points – PPVT-IIIB

Groups	Mean	Standard Deviations	Minimum	Maximum
African American	98.89	55.816	4	200
Hispanic	111.05	54.420	0	199
White	128.47	57.417	0	204
ELL	73.44	24.137	0	140
Restricted-age White	111.28	24.665	0	176

Expert Judges

Ten teachers were selected based on their availability to participate. These teachers were divided into two groups. In order to be selected for the first group, the teachers had to be fluent in both Spanish and English, be of Mexican descent, and teach an ELL class. All except one of the teachers were the classroom teachers of the ELL participants. These four teachers were the only teachers, of those who met the aforementioned criteria, who were available at this stage of the study. The fifth was a graduate student and was specifically recruited by the author of this study. There was no random selection of this sample given the difficulty in finding participants who met the criteria and were willing to volunteer their time to this study.

Teachers of the second group were chosen because they did not meet any of the criteria for the first group. That is, they did not speak Spanish, were not of Mexican descent, and did not teach in an ELL setting. All teachers taught in Central Texas schools. They were selected and approached by their campus administrators and agreed to participate. Random selection was not used. Years of teaching experience for both groups have been presented in Table 3.5.

Content Raters

Five participants were approached by the author of this study and asked to be content raters of the items on both forms of the PPVT-III. All five raters were selected because they were well-educated (i.e., each had advanced degrees); also, each participant had an education related career. Because this task required that the participants have a large vocabulary, they were all selected by the investigator, without random selection.

TABLE 3.5. Years of Experience Teaching

	Number of Years	Mean	Standard Deviation
Group 1 (ELL Teachers)		16.2	12.97
Teacher 1	12		
Teacher 2	36		
Teacher 3	1		
Teacher 4	20		
Teacher 5	12		
Group 2 (Regular Education Teachers)		17.6	10.43
Teacher 1	7		
Teacher 2	23		
Teacher 3	29		
Teacher 4	6		
Teacher 5	23		

Apparatus

Both forms of the Peabody Picture Vocabulary Test – Third Edition (PPVT-III) were under investigation. The standardization sample (n = 2,725) of the PPVT-III was selected to match 1994 U.S. Census data (Williams &Wang, 1997). Stratification within each age group was done according to ethnicity, socioeconomic status (SES), gender, and geographic region. This sample only included those individuals who could speak and understand English.

Reliability

Split-half coefficients for ages 2-6 through 90+ years ranged from 0.89 (at age 2-6) to 0.97 (at age 41-50) with a mean of 0.94 on form IIIA (Williams &Wang, 1997). On form IIIB, coefficients ranged from 0.86 (at 6-6 level) to .96 (at ages 4-6, 11, 13, & 61-90) with a mean of .94. In order to obtain alternate-form reliabilities, the sample was

given both forms of the PPVT-III in a counterbalanced design. The alternate form reliabilities ranged from 0.88 (at age 2-6) to 0.96 (at age 11) with a median of 0.94.

Validity

Four studies were conducted with standardization of the PPVT-III comparing scores from the PPVT-III with instruments of intelligence and oral language (Williams &Wang, 1997). In the studies that compared scores from the PPVT-III (forms IIIA & IIIB) with scores from cognitive assessments, scores from the Wechsler Intelligence Scale for Children – Third Edition (WISC-III; Wechsler, 1991), Kaufman Adolescent & Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993), and Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990) were used. The correlations for form IIIA and form IIIB with the WISC-III Verbal IQ were, respectively 0.91 and 0.92; with the WISC-III Performance IQ 0.82 and 0.84; and WISC-III Full Scale IQ 0.90 and 0.90. The correlations with forms A and B with the KAIT Crystallized IQ were, respectively, 0.87 and 0.91; with the KAIT Fluid IQ, 0.76 and 0.85; and with the KAIT Composite IQ, 0.85 and 0.91. The correlations for each form with the K-BIT Vocabulary were, respectively, 0.82 and 0.80; with the K-BIT Matrices 0.65 and 0.62; and with the K-BIT Composite 0.78 and 0.76.

Scores from the Listening Comprehension (LC) and Oral Expression (OE) Scales of the Oral and Written Language Scales (OWLS; Carrow-Woolfolk, 1995) were also compared with scores from both forms of the PPVT-III (Williams &Wang, 1997). The correlations ranged from 0.63 to 0.83.

Procedure

The procedures described in this section were approved by the Institutional Review Board (IRB) at Texas A & M University. Prior to conducting the current study, permission was obtained from campus administrators, parents of the children who participated in the study, and the teachers who participated in the study. Consent forms for the parents of the children participating in the study were written in English and in Spanish. Children were only tested if they gave verbal assent.

Partial Correlation

According to the technical manual of the PPVT-III (Williams & Wang, 1997), each member of the standardization sample was administered both forms of the PPVT-III (i.e., PPVT-IIIA & PPVT-IIIB) in a counterbalanced design. Members of the standardization sample were from the Northeast, North Central, West, and South regions of the United States of America.

Data from the ELL population were collected from South and Central Texas. Consistent with the standardization sample, both forms of the PPVT-III were administered in a counterbalance design to all participants. Campus administrators were generous in providing separate rooms for testing. Children were administered the PPVT-III according to the administration instructions provided in the *Examiner's Manual for the Peabody Picture Vocabulary Test – Third Edition* (Dunn & Dunn, 1997).

For each item, the correlations between group membership and total score, item response and group membership, and item response and total score were used to calculate a partial correlation between ethnicity and item performance, controlling for

differences in total score (i.e., raw score or total items correct). The African American and Hispanic groups from the standardization sample were compared to the white group from the standardization sample (ethnic minority group = 1 & white group = 0). The ELL sample collected for this study was compared to the white, age-restricted sample (ELL = 1 & age-restricted, white = 0). All items were coded as correct (1) or incorrect (0).

The null distribution was approximated by a significance test for a phi coefficient (see Stricker, 1982 or Reynolds, Willson, & Chatman, 1984), $\chi^2 = Nr^2$, and was tested against a chi-square at the 0.001 significance level with one degree of freedom. The "Critical Values of the Chi-square Distribution," (Siegel & Castellan, 1988, p.323) was used to obtain the value of χ^2 at 0.001 significance level with one degree of freedom (χ^2 = 10.83). The absolute values of obtained partial correlations needed to reject the null have been listed in Table 3.6. Significant positive correlations were interpreted as favoring the subgroup under investigation (focal group) and negative correlations were interpreted as disfavoring the focal group. Items identified as functioning against focal groups were reviewed subjectively for content to determine if a particular pattern based on content category could by determined.

TABLE 3.6. Values Needed to Achieve Significance

Sample	Absolute Values of Correlations
African American vs. white	0.0694
Hispanic vs. white	0.0717
ELL vs. white age-restricted	0.1291

Content Analysis

Items were all classified according to the content categories identified during the national tryout phase of the PPVT-III. Five participants performed this classification process and inter-raters reliabilities were computed with unweighted kappa coefficients. Results are summarized in Table 3.7. On items that were not agreed upon by all raters, the majority determined the category.

Items with significant partial correlations for each group were sorted by category in attempt to identify trends of items functioning against each group. Significance was tested with a chi-square test by using expected frequencies calculated from the entire test.

TABLE 3.7. Kappa Coefficients of Inter-rater Reliability

	Raters				
Raters	2	3	4	5	Composite
1	0.810	0.783	0.765	0.812	0.887
2		0.767	0.727	0.794	0.854
3			0.725	0.789	0.847
4				0.843	0.845
5					0.914

Expert Judges

Both sets of teachers were given instructions to review both forms of the PPVT-III. They were to mark each item that they believed to be biased against ELL students of Mexican descent. Bias was described to the teachers as being any item in which white

students would have an unfair advantage of knowing, for whatever reason, over ELL students of Mexican descent.

Comparison of methods

In order to determine the effectiveness of the expert judgment method, the results obtained from the partial correlation method and those items indicated by teachers as biased were be evaluated by unweighted kappa coefficients. Only the first 132 items of each version were compared, given that no items were identified by the partial correlation method after item 132 on either form and very few subjects responded to items past 132.

Research Questions

With the aforementioned methods, the following research questions were addressed: 1) Is there DIF as indicated by significant partial correlations? If so, is the discrepancy between the positive and the negative correlations significantly different with in each group comparison? 2) Is there a meaningful trend within the items identified as having a significant partial correlation? 3) Does either group of teachers adequately predict those items that underestimate ability for the ELL of Mexican descent group?

CHAPTER IV

RESULTS

Partial Correlation Results

African American

On form IIIA, 35 items were significant at the 0.001 alpha level. Twenty of these items favored the African American normative sample and 15 items favored white standardization sample. On form IIIB, 37 items were significant at the 0.001 alpha level. Twenty-one of these items favored the African American sample and 16 of these items favored the white sample. On both IIIA (χ^2 (1) = .714; p= .398) and IIIB (χ^2 (1) = .676; p= .411) the amount of items favoring the white group over the African American group was insignificant. On both forms the majority of the partial correlations fell between – 1.0 and +1.0. Only four items on version IIIA and seven items on IIIB had partial correlations greater than the absolute value of one. A summary of the frequency distributions was listed in Table 4.1. Correlation in either direction tended to be small.

TABLE 4.1. African American Frequency Distribution of Partial Correlation

	Frequency	
Partial Correlation Index	PPVT-IIIA	PPVT-IIIB
.10 to .19	3	4
.00 to .09	96	100
.00 to -.09	104	97
-.10 to -.19	1	3

Items that achieved significance at the .001 alpha level by the partial correlation method have been listed in Table 4.2 by item set. As stated in the methods section, African Americans were coded as "1" and "0" was used for the white sample. Therefore, positive correlations represented items that favored the African American sample; negative correlations represented items that favored the white sample. On form IIIA items disfavoring the African American sample tended to occur at the beginning and end of the test; while items favoring African Americans occurred within the middle region. Only Item Set 49-60 contained items that functioned both for and against the African American group. On form IIIB items favoring the African American group were found throughout the beginning and midsections of the test. As indicated by the items numbers in each set, there were 12 items per set. Seven items in Item Set 181-192 were indicated to have negative significant partial correlations. Therefore, over half of the items on this set were identified as functioning against the African American sample.

TABLE 4.2. African American Significant Items by Item Set (r >|.0694|)

	PPVT-IIIA		PPVT-IIIB	
Items	Positive	Negative	Positive	Negative
Start Ages 2-2 – 3 1-12				
Start Age 4 13-24		digging, feather, cage	shark	throwing, can, farmer
Start Age 5 25-36		shoulder, accident, penguin	dressing, desk	picking
37-48		tearing	lock	
Start Age 6 – 7 49-60	parachute, delivering	diving, writing	uniform, terrified	
61-72			hive	
Start Age 8 – 9 73-84	selecting		nutritious, annoying	
Start Age 10 – 11 85-96	reptile, polluting		deflated, calculating, cruiser	
97-108	rodent, valley			
Start Age 12 - 16 109-120	injecting, links, cooperative		scholar	
121-132	hazardous, isolation, coast, appliance, foundation		salutation, parallel, glider	banister

TABLE 4.2. Continued

	PPVT-IIIA		PPVT-IIIB	
Items	Positive	Negative	Positive	Negative
133-144	blazing, mammal, reprimanding, consuming, colt		irregular, composing, easel, lubricating, axle	
Start Age 17 + 145-156		ladle		orating
157-168				
169-180		derrick, entomologist		perusing
181-192		wildebeest, honing		stamen, pachyderm, expunging, deciduous, lamenting, perilous, converging
193-204		embossed		supine, pedagogue

Hispanic. On form IIIA, 32 items were significant at the 0.001 alpha level. Fifteen of these items favored the Hispanic normative sample and 17 items favored the white standardization sample. On form IIIB, 52 items were significant at the 0.001 alpha level. Twenty-seven of these items favored the Hispanic sample and 25 of these items favored the white sample. On both version IIIA (χ^2 (1) = 0.125; p= 0.724) and version IIIB (χ^2 (1) = 0.077; p= 0.782) the amount of items favoring the one group over the other was insignificant.

As with the African American sample, the majority of the partial correlations fell between –1.0 and +1.0 on both forms. On version IIIA, 10 items were greater than the absolute value of one and seven items on IIIB had partial correlations greater than the absolute value of one. A summary of the frequency distributions was listed in Table 4.3.

TABLE 4.3. Hispanic Frequency Distribution of Partial Correlation

	Frequency	
Partial Correlation Index	IIIA	IIIB
.10 to .19	3	6
.00 to .09	111	126
.00 to -.09	83	71
-.10 to -.19	7	1

TABLE 4.4. Hispanic Significant Items by Item Set ($r > |.0717|$)

	PPVT-IIIA		PPVT-IIIB	
Items	Positive	Negative	Positive	Negative
Start Age 2-6 – 3				
1-12				
13-24				throwing
Start Age 5				picking
25-36				
37-48			lock, fruit	
Start Age 6 – 7	delivering,		cobweb,	
49-60	rectangle		jogging, huge, uniform, statue, jewelry, terrified	
61-72	luggage, hydrant, calculator		hive, root, tugging, tornado	

TABLE 4.4. Continued

	PPVT-IIIA		PPVT-IIIB	
Items	Positive	Negative	Positive	Negative
Start Age 8 – 9 73-84			ankle, pair, walrus, directing	
Start Age 10 – 11 85-96	surprised, clarinet exhausted		shuttle, tropical, deflated, cruiser	
97-108	pedal, inhaling, valley, tubular, adjustable		sorting, greeting, hoof	
Start Age 12 – 16 109-120			harvesting, assisting	
121-132	hazardous, coast		salutation	
133-144				physician
Start Age 17 + 145-156		ladle		
157-168		confiding, primate		trowel, angler, nape, enumerating, submerging
169-180		pilfering, derrick, ascending, monetary, quintet, incarcerating		marsupial, siphoning, concave, trestle, receptacle, equestrian
181-192		gourmand, quiescent, honing, cupola		depleted, stamen, pachyderm, expunging, deciduous, gable, converging

TABLE 4.4. Continued

	PPVT-IIIA		PPVT-IIIB	
Items	Positive	Negative	Positive	Negative
193-204		embossed, perambulating cenotaph, osculating		copious, supine, succulent, pedagogue

Items that achieved significance at the 0.001 alpha level for the Hispanic sample have been listed in Table 4.4 by item set. On form IIIA, items identified as functioning for the Hispanic sample were located within the first 132 items and items functioning for the white sample were identified within items 145-193. In other words items functioning for the Hispanic sample were found in the first two-thirds of the test and items functioning against the Hispanic sample were found in the last third of the test. Findings on form IIIB were similar to those of the African American sample. That is, items found to be functioning against the Hispanic sample were found within the beginning and ending item sets. On both forms significant negative and positive partial correlations were not found within the same item set.

ELL of Mexican Descent. On form IIIA, 26 items were significant at the 0.001 alpha level. Seven of these items favored the ELL of Mexican descent sample and 19 items favored white standardization sample. This finding was significant at the 0.05 level (χ^2 (1) = 5.538; p=0.019). On form IIIB, 32 items were significant at the 0.001 alpha level. Fourteen of these items favored the ELL sample and 20 of these items favored the white sample. This finding was not significant (χ^2 (1) = 1.059; p=0.303). A

summary of the number of items identified as significant at the 0.001 alpha level on both forms of the PPVT-III is listed in Table 4.5.

TABLE 4.5. ELL Frequency Distribution of Partial Correlation

	Frequency	
Partial Correlation Index	IIIA	IIIB
.20 to .29	4	3
.10 to .19	10	25
.00 to .09	96	84
.00 to -.09	32	31
-.10 to -.19	11	16
-.20 to -.29	18	7
-.30 to -.39	7	2

TABLE 4.6. ELL Significant Items by Item Set ($r >$ |0.1291|)

	PPVT-IIIA		PPVT-IIIB	
Items	Positive	Negative	Positive	Negative
Start Age 2-6 – 3 1-12			baby, money	
Start Age 4 13-24			kangaroo	
Start Age 5 25-36			desk	
37-48			time	triangle
Start Age 6 – 7 49-60		diving, drilling, hook	lock, uniform, statue	cobweb, wrist, binoculars
61-72	awarding, calculator	vehicle, hydrant, signal, squash, frame	liquid, brain, root, tornado	
Start Age 8 – 9 73-84	heart	towing, horrified, wrench	pair	ankle, antlers, nutritious, jaw, cliff

TABLE 4.6. Continued

	PPVT-IIIA		PPVT-IIIB	
Items	Positive	Negative	Positive	Negative
Start Age 10 – 11 85-96	Reptile	tambourine, interviewing, pitcher, polluting	deflated	shuttle, tropical, angle, shore
97-108	demolishing, fern, hurdling	pedal, inhaling, tusk		canine, arctic, colliding
Start Age 12 – 16 109-120		fragile		gnawing, beverage
121-132				banister, hovering

Although version IIIA was found to have significantly more items favoring the white sample than the ELL sample, Table 4.6 demonstrated that items favoring the ELL sample and the white sample were distributed throughout the test. Items sets containing items 49-60 and 109-120 contained items only functioning against the ELL population. Otherwise, the item sets that contained items with significant negative correlations also contained items with significant positive correlations. On version IIIB, items found to function for the ELL sample tended to occur within the first part of the test; whereas, items functioning against the ELL population occurred later in the test. There was overlap. Within several item sets items were identified as functioning for and against the ELL group. No significant items were found as significant after item 132 on either version. However, items identified toward the end of the tests should be interpreted with caution given that the majority of the sample had reached their ceiling by item 122 (i.e., 2 SD, see Tables 3.3 & 3.4).

Content Analysis Results

An attempt was made to identify bias from DIF by comparing the frequencies of items within each category on both tests combined, as designated by sorters, to the frequencies of items with significant negative partial correlations. Table 4.7 displays the frequencies of each category found on both versions combined. These frequencies were calculated twice; the first was for the total number of items used for the African American and Hispanic samples and the second for items 1-132, on each version, for the ELL sample. The data displayed in Tables 4.8 provide the results of the chi-square analyses of each category. Frequencies from Table 4.7 were used to calculate expected frequencies.

As shown in Table 4.8, one category from the African American sample was identified as containing significantly more items with significant partial correlations than expected. Three of the 14 items were detected as having significant negative partial correlations within the "Workers" category. This finding was significant at the 0.05 alpha level. Regarding the Hispanic sample, none of the categories was identified as containing significantly more items than expected based on overall frequencies of the combined versions. Within the ELL of Mexican descent sample, two categories were identified as containing significantly more items than expected. Both of the items within the "Foods" category were identified, which was significant at the 0.001 alpha level and four of ten items from the "Geographical Scenes" category were identified, which was significant at the 0.05 alpha level.

TABLE 4.7. Content in Versions A & B Combined

Categories	Percentage of Items	Number of Items	Percentage of Items 1-132	Number of Items 1-132
Action	25.2	103	26.5	70
Adjectives	12	49	9.8	26
Animal	11.8	48	12.5	33
Body Parts	3.7	15	4.5	12
Books	0.5	2	0.4	1
Building	4.7	19	3.4	9
Clothing & Accessories	1.2	5	1.1	3
Emotions	2	8	0.8	2
Food	1	4	0.8	2
Fruits & Vegetables	1.5	6	1.9	5
Geographical Scenes	3.4	14	3.8	10
Household Objects	4.2	17	4.9	13
Musical Instruments	1.2	5	1.9	5
People	2.7	11	1.9	5
Plants	3.2	13	3.4	9
Shapes	4.4	18	3.8	10
Tools	9.8	40	10.2	27
Toys	0.5	2	0.8	2
Vehicles	2.5	10	3.4	9
Workers	3.4	14	3.4	9
Other	1.2	5	0.8	2
Total	100.1	408	100	133

TABLE 4.8. Significant Negative Partial Correlations by Category

	African American			Hispanic			ELL		
Category	*n*	$\chi^2(1)$	***p***	*n*	$\chi^2(1)$	***p***	*n*	$\chi^2(1)$	***p***
Action	12	2.430	0.119	15	2.058	0.515	9	0.203	0.652
Adjectives	2	0.861	0.354	5	0.000	0.985	2	1.019	0.131
Animals	5	0.531	0.466	4	0.207	0.649	4	0.186	0.667
Body Parts	1	0.021	0.884	1	0.218	0.641	3	1.045	0.307
Books	0	--	--	0	--	--	0	--	--
Buildings	1	0.155	0.693	4	2.318	0.128	1	0.094	0.759
Clothing & Accessories	0	--	--	0	--	--	0	--	--
Emotions	1	0.252	0.615	0	--	--	0	--	--
Foods	0	--	--	0	--	--	2	10.829	0.001
Fruits & Vegetables	0	--	--	1	0.243	0.622	1	0.106	0.744
Geographical Scenes	0	--	--	0	--	--	4	5.023	0.025

TABLE 4.8. Continued

	African American			Hispanic			ELL		
Category	*n*	$\chi^2(1)$	***p***	*n*	$\chi^2(1)$	***p***	*n*	$\chi^2(1)$	***p***
Household Objects	1	0.076	0.783	2	0.035	0.852	2	0.005	0.945
Musical Instruments	0	--	--	0	--	--	1	0.106	0.744
People	0	--	--	2	0.737	0.391	0	--	--
Plants	2	1.107	0.293	2	0.356	0.551	0	--	--
Shapes	1	0.105	0.746	1	0.434	0.510	2	0.213	0.645
Tools	2	0.384	0.536	4	0.004	0.952	6	1.205	0.272
Toys	0	--	--	0	--	--	0	--	--
Vehicles	0	--	--	0	--	--	2	0.402	0.526
Workers	3	3.882	0.049	1	0.143	0.705	0	--	--
Other	0	--	--	0	--	--	0	--	--

Expert Judgment Results

As demonstrated in the Tables 4.9 and 4.10, all of the kappa coefficients were extremely low. The "expert" teacher group kappa coefficients ranged from -0.058 to 0.242 and the kappa coefficients in the "non-expert" group ranged from -0.092 to 0.071. The fifth non-expert teacher indicated that she did not believe any items to be biased against ELL of Mexican descent because they all had access to television. Therefore her endorsements remained constant and no kappa could be calculated.

TABLE 4.9. Expert Teachers

Teacher	PPVT-IIIA	PPVT-IIIB
1	0.072	-0.058
2	0.119	-0.041
3	0.049	0.083
4	-0.006	0.052
5	0.242	-0.022

TABLE 4.10. Non-expert Judgment

Teacher	PPVT-IIIA	PPVT-IIIB
1	-0.052	-0.080
2	0.051	0.071
3	-0.071	-0.041
4	-0.092	0.011
5	--	--

Summary of Results

Using the partial correlation method, DIF was detected within each group comparison. In all cases except with the ELL on form A of the PPVT-III, there was no significant difference in number of item found to have significant positive correlations versus significant negative correlations. On form A the ELL group comparison indicated more items with negative correlation than positive correlation (χ^2 (1) = 5.538; p=0.019). Among the items flagged as underestimating ability of the ELL group, no consistent trend could be detected. Also, it was found that none of the expert judges could adequately predict those items that would underestimate ability for the ELL group, despite expertise. Discussion includes possible consequences of item placement and recommendations regarding further research and use of the PPVT-III.

CHAPTER V

DISCUSSION AND SUMMARY

Summary of Study and Findings by Hypothesis

There were three purposes of this study. The first was to detect DIF on both forms of the PPVT-III with African Americans, Hispanics, and ELL of Mexican descent when compared to the white normative sample by a partial correlation method. The second was to determine whether a trend based on item content could be determined among the items found to have significant partial correlation for each focal group. The third purpose of this study was another attempt to validate an expert judgment method of DIF by using teachers as judges and to determine whether special knowledge of one group allowed better prediction over the control group.

Partial Correlation

In order to find items that functioned against the African American and Hispanic sample, standardization data were provided courtesy of AGS. A white sample, acting as a reference group, also came from these data. Three hundred, first through third grade, ELL of Mexican descent were recruited from south and central Texas for this study.

All subjects were administered both versions of the PPVT-III in a counterbalanced design. Administration rules, including basal and ceiling rules, were followed according to the examiner's manual (Dunn & Dunn, 1997). The African American, Hispanic, and ELL samples were each compared to a white sample. DIF was detected with a partial correlation between race and item performance, controlling for difference in total score. Significant positive correlations indicated items that functioned

for (i.e., favored) the ethnic minority group under investigation and significant negative correlations indicated items that functioned for the white reference sample.

Findings revealed significant positive and negative correlations with each of the groups (i.e., African American, Hispanic, & ELL) when compared to the white comparative group. On both forms of the PPVT-III there was no significant difference found between the amount of items that displayed significant positive correlations versus negative correlations for African American and Hispanic samples. Similar findings were indicated on form IIIB with the ELL sample. However, on form IIIA there were significantly more items found to function against the ELL sample when compared the white sample.

Findings of Hypothesis One. Consistent with previous findings in other studies attempting to detect DIF, and ultimately item bias (see Jensen, 1980; Brown et. al, 1999), the first hypothesis was accurate for the African American and Hispanic groups when compared to the white sample. Form IIIB with the ELL sample was also consistent with this hypothesis. However, the first hypothesis (the null) was rejected on form IIIA with the ELL sample. That is, there were significantly more items found to function against the ELL group than were found to function for the ELL group (χ^2 (1) = 5.538; p=.019).

Content Analysis

After the partial correlations were calculated, an attempt was made to compare significant positive and negative correlations according to content categories previously used for creating items for the PPVT-III (Williams & Wang, 1997). Items with

significant positive and negative partial correlations were compared based on frequency of occurrences in each of these content categories. Based on their occurrences, relative to overall item occurrences, no apparent trends emerged.

Findings of Hypothesis Two. Consistent with previous literature (Reynolds, et al. 1999) and the second hypothesis, no consistent trend or patterns could be determined based on the content of the items identified to have statically significant partial correlations.

Expert Judges

A total of 10 teachers were recruited to participate in this study. Five of the teachers were selected because they taught ELL students, spoke Spanish, and were of Mexican descent. All teachers lived in south or central Texas. Five other teachers were recruited as "non-expert" controls. These teachers did not teach ELL students, did not speak Spanish, and were not of Mexican descent.

All teachers were given the first 132 items of both versions of the PPVT-III and asked to identify which items would be biased against ELL students of Mexican descent. Their identified items were then compared to items that were identified as functioning against the ELL sample by the partial correlation method with an unweighted kappa coefficient. All coefficients were low (range .242 to -.080).

Findings of Hypothesis Three. Consistent with the final hypothesis and other related findings in the literature (for review see Reynolds et al., 1999), items that function differently could not be detected by subjective viewing of the items, even with previous experience with or special knowledge of a special population.

Discussion

In contrast to Reynolds et al. (1984), more items were detected as significant at the .001 alpha level with the partial correlation method on the PPVT-III than on the either form of the PPVT-R. However, similar to their findings on form L of the PPVT-R with the African American sample, the PPVT-III form A contained significantly more items with significant negative correlations than positive correlations for the ELL group. Therefore, version IIIB was thought to be superior to IIIA for use with ELL of Mexican descent.

The location of significant items (both positive and negative) was a bit concerning, particularly on form IIIA with the African American sample. The second, third, and fourth item sets all contained items that functioned against African Americans. The fifth item set contained items that functioned both for and against the African American sample. The seventh item set through the twelfth item set only contained items that functioned for the African American sample. Although these findings are unlikely to be of major consequence in clinical practice because of the small correlations, the findings are of significant concern in longitudinal research. Therefore, version IIIB was thought to be superior to version IIIA with preschool through fourth grade African American children.

In addition to trends detected by location of items or by item sets, with the African American sample it was demonstrated that there were significantly more items than expected based on overall frequencies detected with significant negative partial

correlations within the "Workers" category. No categories were detected with the Hispanic sample. Two categories were detected with the ELL sample: "Foods" with significance at the 0.001 alpha level and "Geographical Scenes" at the 0.05 alpha level. In all cases, these significant finds were based on a small number of items and therefore further research needs to be conducted prior to making any conclusions about these categories and possible bias. In qualitatively examining the items, there were no items that appeared to measure something significantly different than vocabulary. The PPVT-III has been conceptualized as a culturally loaded test because item difficulty is directly related to occurrence of the word in language (Jensen, 1980). However, the PPVT-III has been constructed to assess a culturally loaded construct, receptive vocabulary.

Another purpose of this study was to determine whether special expertise of a particular group would aid in predicting bias. Neither group under investigation reliably predicted those items that functioned against ELL. This finding was not surprising. As stated by multiple authors (e.g., Camilli & Shepard, 1994; Jensen 1980; Reynolds et al., 1999) expert panels should continue to be used with the specific purpose of identifying items likely to be offensive. However, it should not be assumed that being of a certain ethnic background bestows special knowledge of what is offensive to remaining members of an ethnic background.

Limitations

Limitations of this study were similar to those limitations in other studies attempting to detect item bias through DIF. The partial correlation method used in the current study identified items with significant correlations between item group and item

response while controlling for total score. This method assumed the total score to be unbiased or as equally predictive of the underlying trait for one group as the other group. This logic has been questioned by Hunter and Schmidt (2000). They argued the claims of unbiased possibly containing biased items were logically flawed; especially when the same researchers have claimed that a large amount of biased items indicate a biased test. This argument continued to demonstrate that studies such as the one presented in this dissertation only detect DIF and not bias. However, items that function differently for one group than another were suspect for bias (Reynolds et al., 1999).

Another limitation to this study was a lack of homogeneity within groups. As shown in Tables 3.2 and 3.3 the within group standard deviations for total scores are large. In the African American, Hispanic, and white groups the age range is also quite large and contributes to the wide range of test scores. These large standard deviations likely have impacted the results and therefore should be interpreted with caution. Also, the samples within this study were not examined by gender, which could have also masked significant finding.

Although the age ranges were restricted for the comparison of the white group to the ELL of Mexican descent, another cautionary statement also needs to be made. There was no formal test of first language or second language proficiency. Therefore, it is likely that these children had different levels of proficiency in each language.

Implications and Recommendations

Within this study, several items were identified with the partial correlation method as functioning differently for each of the focal groups when compared to the

reference group. However, no trends were identified by content or subjective rationale as to why certain items were identified as functioning for a group versus those that were identified as functioning against that same group. Although several items were identified as functioning against each group, bias was neither assumed nor denied as occurring in any of the items tested.

Although bias could not be found or discounted within the current study, DIF was detected as occurring within each focal sample. The partial correlation method did yield two unexpected and noteworthy findings, however. The first of these findings was that more items were found to be functioning against the ELL group than for the ELL group on version IIIA; therefore version IIIB should be used instead of IIIA when testing ELL children of Mexican descent within the primary grades. The second of these findings was that there was a peculiar group of significant partial correlations according to item sets. The most alarming was demonstrated on version IIIA with the African American sample. It was therefore recommended that extreme caution be used when measuring progress with the PPVT-III as when using any standardized test to measure progress. Based on the current findings, if progress monitoring has to be done with the measure, IIIB was recommended for use with elementary-aged African American children, especially in research applications.

Further research is needed to determine the significance of item location in differential items functioning. If when compared to the entire standardization sample, instead of a selected comparative group (i.e., the white sample), there were significant trends based on location, then there could be significant implications regarding progress

monitoring. For example, if a trend similar to that found on IIIA with the African American sample were found when compared to the entire standardization sample (i.e., negative items early and positive items later), then growth or progress demonstrated in testing could be an artifact of the test rather than actual growth or progress. For example if the PPVT-IIIA was used as a program evaluation tool to assess growth within an African American sample and baselines were assessed at the age of five years and then progress monitored at a three-year reevaluation at the age of eight years, then growth beyond the expected trajectory (based on standard scores) could be an artifact of the test rather than actual growth or progress through the program. Of course this scenario was one of test use and demonstrated Jensen's (1980) statement that even unbiased tests could be used in unfair ways. Ethically, progress or assessment of language and/or vocabulary would not be made with the results of one measure, consistent with Judge Grady's decision referring to PL 94-142.

Although constructing tests that are culturally sensitive is extremely important, as well as ethical (APA, 2002) test results need context specific interpretation by trained professionals. Jensen (1980) made the point, "We must distinguish between tests and testing practices; between current *de facto* uses (and abuses) and possible optimal uses; and between tests and testing as they are today and as they might be in the future" (p.41). An examiner needs to interpret test scores with consideration given to the background of the child and the purpose of the assessment. This means that no test, including the PPVT-III should ever be used to evaluate a person or a program in isolation.

Interpretation of PPVT-III results in research as well as clinical practice needs to happen with data from other relative sources, including other tests and history.

REFERENCES

Allen et al. v. Alabama State Board of Education et al., U.S. District Court Middle District of Alabama. Northern Division Docket No. 81-697, (1985).

Altepeter, T. S. (1989). The PPVT-R as a measure of psycholinguistic functioning: A caution. *Journal of Clinical Psychology, 45*, 935-941.

Altepeter, T. S., & Handal, P. J. (1985). A factor analytic investigation of the use of the PPVT-R as a measure of general achievement. *Journal of Clinical Psychology, 4*, 540-543.

American Psychological Association. (2002). *American Psychological Association ethical principles of psychologists and code of ethics*. Retrieved May 20, 2003, from http://www.apa.org/ethics/code.html

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96 116). Baltimore: The Johns Hopkins University Press.

Angoff, W. H., & Ford, S. R. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10,* 95-106.

Argulewicz, E. N., & Abel, R. R. (1984). Internal evidence of bias in the PPVT-R for Anglo-American and Mexican-American children. *Journal of School Psychology, 22*, 299-303.

Binet, A., & Simon, T. (1979). The development of intelligence in children (the Binet-*Simon scale)* (E. S. Kite, Trans). New York: Arno. (Original work published in 1911).

Bracken, B. A. (1986). *Bracken concept development program*. San Antonio, TX: Psychological Corporation.

Bracken, B. A., & Prasse, D. P. (1981). Comparison of the PPVT, PPVT-R, and intelligence tests used for the placement of black, white, and Hispanic EMR students. *Journal of School Psychology, 19*, 304-311.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since bias in mental testing. *School Psychology Quarterly, 14*, 208-238.

Bruce, M. (1940). Factors affecting intelligence test performance of Whites and Negroes in the rural South. *Archives of Psychology*, *252*, 99.

Camilli, G., & Shepard, L. A. (1987). The inadequacy of ANOVA for detecting test bias. *Journal of Educational Statistics*,*12*, 87-99.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Carrow-Woolfolk, E. (1995). *Oral and Written Language Scales: Listening Comprehension and Oral Expression.* Circle Pines, MN: American Guidance Service, Inc.

Cole, N.S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement* (pp. 201-219). New York: Macmillan.

Darlington, R. B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement, 8*, 71-82.

Diana V. California State Board of Education, U.S. District Court for the Northern District of California (consent decree) (1970).

Dunn, L. M., & Dunn, L. M. (1981). *Examiner's manual for the Peabody Picture Vocabulary Test – Revised*. Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Dunn, L. M. (1997). *Examiner's manual for the Peabody Picture Vocabulary Test – Third Edition*. Circle Pines, MN: American Guidance Service.

Education for All Handicapped Children Act of 1975, 20 U.S.C. § 1401 *et seq*.

Frisby, C. L. (1999). Straight talk about cognitive assessment and diversity. *School Psychology Quarterly*,*14*, 195-207.

Golden Rule Insurance Company et al. v Washburn et al., 419-76 *stipulation for dismissal and order dismissing case, file in the Circuit Court of the Seventh Judicial Circuit, Sangamon County, IL (1984).

Halpin, G., Simpson, R. G., & Martin, S. L. (1990). An investigation of racial bias in the Peabody Picture Vocabulary Test – Revised. *Educational & Psychological Measurement*, *50*, 183-189.

Hobson v. Hansen, 269 F. Supp. 401 (D.C. 1967).

Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law*, *6*, 151-158.

Jensen, A. R. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs*, *90*, 185-244.

Jensen, A. R. (1980). *Bias in mental testing.* New York: The Free Press.

Jensen, A. R. (1984). Test bias: Concepts and criticisms. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 507-586). New York: Plenum Press.

Jensen, A.R. (1994). Race and IQ scores. In R. J. Sternberg (Ed.), *Encyclopedia of human intelligence* (Vol. 2, pp. 899-907) New York: Macmillan.

Kamphaus, R. W. (2001). Culture and bias. In *Clinical assessment of child and adolescent intelligence, second edition*. Heights, MA: Allyn & Bacon Needham.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children*. Circle Pines, MN: American Guidance Service, Inc.

Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman Brief Intelligence Test.* Circle Pines, MN: American Guidance Service, Inc.

Kaufman, A. S., & Kaufman, N. L. (1993). *Kaufman Adolescent and Adult Intelligence Test.* Circle Pines, MN: American Guidance Service, Inc.

Larry P. et al. v. Wilson Riles et al., 343 F. Supp. 1306 (N. D. Cal. 1972) (preliminary injunction), aff'd, 502 F. 2d 963 (9th Cir. 1974); 495 F. Supp. 926 (N.D. Cal. 1979) (decision on merits), aff'd, (9th Cir. No. 80-427 Jan. 23, 1984). Order modifying judgment, C-71-2270 RFP (9th Cir. 1986).

Linn, R. L., & Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. *Educational Measurement: Issues and Practices*, *6*, 13-17.

Linn, R. L., & Werts, C.E. (1971). Consideration for studies of test bias. *Journal of Educational Measurement*, *8*, 1-4.

Maxwell, J. K., & Wise, F. (1984). PPVT IQ in adults: A measure of vocabulary, not of intelligence. *Journal of Clinical Psychology*, *40,* 1048-1053.

McGurk, F. C. J. (1975). Race differences – twenty years later. *Homo*, *26*, 219-239.

National Center for Education Statistics (2001). *Participation in education: Racial/ethnic distribution of public school students.* Retrieved April 28, 2002, from http://nces.ed.gov/programs/coe/2001/section1/tables/t03_1.html

National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs. (2000). *The growing numbers of limited English proficient students.* Retrieved April 28, 2002, from http://www.ncbe.gwu.edu/

PASE (Parents in Action on Special Education) v. Joseph P. Hannon, 506 F. Supp. 831 (N. D. Ill. 1980).

Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the test validation process. *Educational and Psychological Measurement, 40,* 397-404.

Reynolds, C. R. (2000a). Methods for detecting and evaluating cultural bias in neuropsychological tests. In E. Fletcher-Janzen, T. L. Strickland, & C. R. Reynolds (Eds.), *Handbook of cross-cultural neuropsychology*. New York: Kluwer Academic / Plenum Publishers.

Reynolds, C. R. (2000b). Why is psychometric research on bias in mental testing so often ignored? *Psychology, Public Policy, and Law*, 6, 144-150.

Reynolds, C. R., & Brown, R. T. (1984). Bias in mental testing: An introduction to the issues. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing.* New York: Plenum.

Reynolds, C. R., Lowe, P. A., & Saenz, A. L. (1999). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 549-595). New York: John Wiley & Sons, Inc.

Reynolds, C. R., Willson, V. L., & Chatman, S. P. (1984). Item bias on the 1981 revisions of the Peabody Picture Vocabulary Test using a new method of detecting bias. *Journal of Psychoeducational Assessment*, *2*, 219-221.

Sattler, J. M. (1981a). How good are federal judges in detecting differences in item difficulty on intelligence tests for ethnic groups? *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*, 125-129.

Sattler, J. M. (1981b). Intelligence tests on trial: An "interview" with Judges Robert F. Peckham and John F. Grady. *Journal of School Psychology*, *19*, 359-369.

Scheuneman, J. D. (1982). A Posteriori Analyses of Biased Items. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore: The Johns Hopkins University Press.

Shapiro E., Krivit W., Lockman L., Jambaque I., Peters C., Cowan M., et al. (2000). Long-term effect of bone-marrow transplantation for childhood-onset cerebral X linked adrenoleukodystrophy. *Lancet*, *356 (9231)*, 713-8.

Shepard, L.A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22*, 77-105.

Siegel, S., & Castellan, N. Jr. (1988). *Nonparametric statistics for the behavioral sciences – Second edition*. Boston: McGraw Hill.

Stinnett, T. A., Havey, J. M., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: A national survey. *Journal of Psychoeducational Assessment, 12*, 331-350.

Strein W., & Ysseldyke, J. E. (1974). Process- and product-dominant testing of disadvantage and nondisadvantaged Appalachian children. *Exceptional Children*, *40*, 451-451.

Stricker, L. J. (1982). Identifying test items that perform differentially in population subgroups: A partial correlation index. *Applied Psychological Measurement*, *6*, 261-273.

Stricker, L. J. (1984). The stability of a partial correlation index for identifying items that perform differentially in subgroups. *Educational and Psychological Measurement*, *44*, 831-837.

Tittle, C. K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore: The Johns Hopkins University Press.

Valencia, R. R. (1992). Explaining cultural bias in educational test: How important is "Opportunity to Learn?" *Child Assessment News*, *2*, 8-11.

Valencia, R. R., Rankin, R. J., & Livingston, R. (1995). K-ABC content bias: Comparisons between Mexican American and white children. *Psychology in the Schools*, *32*, 153-169.

Veale, J. R., & Foreman D. I. (1983). Assessing cultural bias using foil response date: Cultural variation. *Journal of Educational Measurement*, *20*, 249-258.

Washington, J. A., & Craig, H. K. (1999). Performances of at-risk, African American preschoolers on the Peabody Picture Vocabulary Test – III. *Language, Speech, and Hearing Services in School*, *30*, 75-82.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children – Third Edition*. San Antonio, TX: The Psychological Corporation.

Williams, K. T. & Wang, J. (1997). *Technical references to the Peabody Picture Vocabulary Test – Third Edition (PPVT-III).* Circle Pines, NM: American Guidance Service, Inc.

Williams, R. L., Dotson, W., Don, P., & Williams, W. S. (1980). The war against testing: A current status report. *Journal of Negro Education*, *49*, 263-273.

Willson, V. L., Nolan, R. F., Reynolds, C. R., & Kamphaus, R. W. (1989). Race and gender effects on item functioning on the Kaufman Assessment Battery for Children. *Journal of School Psychology*, *27*, 289-296.

VITA

NAME:		Colleen A. Conoley
DATE OF BIRTH:		May 1, 1974
ADDRESS:		8506 South 143rd Street Omaha, NE 68138
EDUCATION:	B.A.	Southwest Texas State University San Marcos, TX (Psychology, 1996)
PROFESSIONAL EXPERIENCE:		Intern Munroe-Meyer Institute Omaha, NE, 2000-2001 Pediatric Neuropsychology Fellow University of Minnesota Minneapolis, MN, 2001-2003
PUBLICATIONS:		Conoley, J. C., & Conoley, C. A. (2001). Systemic interventions for safe schools. In J. H. Hughes, A. M. LaGreca, & J. Conoley (Eds). *Handbook of psychological services for children and adolescents* (pp. 439 – 454). New York: Oxford University Press.