Summer 2021

Evaluating the Florida Postsecondary Education Readiness Test for Bias: An Adaptation of the Meade-Fetzer Updated Cleary Method

Ryan C. Criss
Shawnee State University

SHAWNEE STATE UNIVERSITY

Evaluating the Florida Postsecondary Education Readiness Test for Bias: An Adaptation of the Meade-Fetzer Updated Cleary Method

A Thesis

By

Ryan C. Criss

Department of Mathematical Sciences

Submitted in partial fulfillment of the requirements

for the degree of

Master of Science, Mathematics

6 July 2021

Accepted by the Graduate Department

Dustin G. Dahm 7/31/2021

Graduate Director, Date

The thesis entitled **'Evaluating the Florida Postsecondary Education Readiness Test for Bias: An Adaptation of the Meade-Fetzer Updated Cleary Method'** presented by **Ryan C. Criss**, a candidate for the degree of **Master of Science in Mathematics,** has been approved and is worthy of acceptance.

7/31/2021
Date

Douglas D. Dankel, Ph.D.
Graduate Director

6 July 2021
Date

Ryan Clay Criss
Student

ABSTRACT

In 2008, the Florida Senate enacted a College and Career Readiness Initiative which saw the creation of a statewide common placement test: the Postsecondary Education Readiness Test, or PERT. However, as of 2021, there is only one publicly available study on the predictive validity of the PERT and none on the potential for Test Bias. This study aimed to detect Adam Meade and Michael Fetzer's definition of Test Bias, a "systematic error in how a test measures performance for a particular group," using their updated version of the standard and widely-accepted Cleary Regression Method. The expectation was that either Florida's efforts in the realm of access to higher education would be validated by this study or the need to more deeply research the PERT would be made obvious. A Composite PERT score, created by averaging the scores of the three PERT subsections Math, Reading, and Writing for each student, was used to predict first-year college GPA for the demographic variables Gender and Race using Gender, Composite, and the Gender-Composite interaction effect in accordance with the Cleary Method. While no evidence was found that Test Bias exists for Female, Black, or Underrepresented race students, there was evidence that PERT exhibits Test Bias for Hispanic students. This does not imply that the PERT should be immediately discontinued, but that more studies need to be conducted in order to elucidate this Test Bias. Such studies also need to be made publicly available in order to assuage any concerns the Florida public may have towards this ubiquitous exam.

ACKNOWLEDGMENTS

I would like to thank Drs. Douglas Darbro, John Whitaker, David DeSario, and Philip Blau for their excellent teaching and patience with me during my course of study at Shawnee State University; my director Chris, manager Debi, and coworker Tina for working with me throughout this graduate program; Anthony, Bob, and Michele for welcoming me as part of their lives; Jenna for her understanding and friendship; capitalism for forcing me into graduate studies a year ahead of schedule; and Dr. Mary Clark of Polk State College's Institutional Research department for her support in this project and others. Stephen also deserves acknowledgement for being a significant motivating force in my life.

But above all else, I would like to thank my mother, without whom absolutely none of this would be possible. You never get enough credit, mom.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: Introduction

Dissatisfied with outcomes to that point, the Florida state senate enacted a College & Career Readiness Initiative in 2008 (FL SB 1908). The Initiative had many paths, one of which was the statewide common placement test: the Postsecondary Education Readiness Test, or PERT. Florida's PERT was a combination of efforts from assessment company McCann and a huge coalition of Florida secondary and postsecondary educators. They piloted their effort in the summer of 2010 and released the PERT as a placement tool across the state by summer 2011. This replaced the state's use of Accuplacer as a placement tool, though the PERT cutoff scores deciding a student's placement in composition and mathematics were based on previous Accuplacer results. While this was done to ensure continuity, research on Accuplacer's predictive ability (Medhanie, 2012) makes such a choice suspicious, but this study is not focused on PERT's creation; it is focused on its outcomes.

This study seeks to start a conversation about PERT. While many larger and more well-known college entrance exams (SAT, ACT, etc.) have had decades of studies on their reliability, bias, and efficacy, PERT is essentially a baby next to these older siblings. Because a more intensive study into the test's question bank is not possible, the discussion will be started by observing whether or not PERT exhibits a bias with respect to gender and/or race. A more hopeful outcome is expected: that no evidence of bias is found and Florida's efforts are validated. However, should the existence of bias be discovered, there should be no condemnation of the exam itself. Instead, a necessary critical evaluation of the PERT and Florida student outcomes should be started to ensure the mission of the original Initiative can come to pass.

Background of the Problem

Standardized testing for college entrance is a relatively new phenomenon. Originally, every college and university in the United States had their own entrance exam. The first standardized test collated these entrance exams into one long test. In 1926, the Scholastic Aptitude Test, or SAT, became the first national standardized test for college admissions not created by a college or university (Atkinson, 2009; Nettles, 2019).

Questions of bias in the SAT have been the subject of numerous studies. Some decry the exam as racist (Freedle, 2003; Santelices, 2010), others say it is as fair as fair can be (Brothen, 2003; Kobrin, 2007), and some say it benefits minorities over white students (Cleary, 1968). Questions about whether or not the SAT is an accurate predictor of college success led to the creation of other standardized tests. The ACT, for example, started as and still is a main competitor with the SAT. Accuplacer arose in 1985 and set the trend for computerized adaptive tests, though Accuplacer has been shown to underpredict success in college (Belfield, 2012). Indeed, even the efficacy or reliability of computer adaptive testing is still under study as of the year 2020.

To answer if standardized exams are a biased tool, one must first answer the question "how is bias defined?" Statisticians are familiar with the term bias as a measurable degree of error but the colloquialism relating bias as "unfair" seems to be the main driver of research in educational testing. For example, if a student is placed in a class lower than their skill level because the test predicts different outcomes with race as a significant factor but the student still succeeds, graduates, and attains a six-figure salary, is the test really "unfair" and worth discrediting? This was a typical rejoinder in decades past, but the explosion in remediation and

the high cost of tuition coupled with opportunity cost for those not from wealthy families have rendered this point moot. More appropriate would be to define bias like the statisticians do, an immutable concept not subject to the whims of social connotation. Meade and Tonidandel in their 2010 paper "Not Seeing Clearly with Cleary" have struck an appropriate balance with their definition. They clearly label Test Bias as the "systematic error in how a test measures a particular group." This does not index bias based on student outcome but instead is a systematic way to present evidence that a test favors or disfavors a particular group or groups. This is excellent for initial research and for pilot programs of new standardized tests. However, one still has to understand the two ways a test can be biased: through measurement bias or predictive bias.

The differences between measurement bias and predictive bias mostly come from the source of the bias. Measurement bias happens when items on the test itself favor groups with specific culture, upbringing, or experiences such as the infamous "oarsmen : regatta" question on the SAT. This type of bias requires more rigorous methods such as differential item functioning or item response theory to detect. This is beyond the scope of the present study as the methods take too long and require access to the PERT's questions that Florida and McCann are unlikely to allow. Instead, predictive bias will be explored here. Predictive bias means that fitting a common regression line for two or more groups gives a statistically significant difference between predicted response and observed response. Methods in predictive bias focus on student outcomes to index bias from the test, such as Cleary's 1968 study which predicted first-year college GPA based on high school rank and SAT score for black students and white students.

In 1968, T. Anne Cleary set out to prove the SAT was not biased. She concluded that the SAT overpredicted first-year GPA for black college students and that "any bias was in favor of the [black] student." Her methodology has come to be called the Cleary 1968 Method and is encoded in the AERA/APA/NCRE *Standards* (1999) for evaluating the existence of predictive bias, albeit implicitly since the original study contains quite racist language by today's standards. However, her study has several problems. First, the researchers decided if the student was white or black based on black and white photographs because colleges did not record race data at the time. If they could not determine the race, it was defaulted to white. Second, there was no accounting for variation of SAT scores within the black and white subsets that may have altered the regression equations. Finally, there was no accounting for gender differences if there even were male and female students studied. Cleary herself acknowledged that no general conclusions could be drawn from her study, but it was still elevated in spite of these problems. Her method went almost totally unchanged for nearly forty years.

Adam Meade and Michael Fetzer of University of North Carolina recognized Cleary needed an update. They recognized that differences in test scores within groups needed to be accounted for as they played a pivotal role in creating the coefficients of the regression equations. Meade and Fetzer also emphasized a need to be careful with language, calling for researchers to standardize their definition of bias without the connotation of "racist" or "sexist" or "unfair" that the word has in everyday speech. Separating "bias" from "fairness" is important to dealing with the tests in good faith, and so they proposed their own definition of Test Bias as given earlier. Such a simple update removes the student outcome indexing from regular regression methods and provides a more suitable platform to raise concerns about assessment

instruments. This is where PERT and the history of standardized testing will collide, with this study applying Meade and Fetzer's methodology to PERT outcome data.

Statement of the Problem

The central focus of this study is to examine the PERT for bias now, after ten years in use, using Meade and Fetzer's revised Cleary method. For one, there are decades of publicly available research for the SAT, ACT, ACCUPLACER, and other national standardized tests; PERT does not yet have those. That can be taken to mean that studies on the validity, reliability, and "fairness" of the PERT have not yet been conducted or at least not readily available for whatever reason. Simple addition shows that the more college courses a student has to take the higher the overall tuition bill will be. Placement at the remedial (or developmental) level in Florida incurs an additional tuition cost while also not giving students college credit, extending both the cost and time to completion. This is a potential economic burden that may or may not be clear to students placed at that level. Because this test has played a significant role in many young Floridians' lives since 2011, it is well overdue for these kinds of validations to be made publicly available.

Specifically, this study will examine the existence of Test Bias according to Meade and Fetzer's 2009 definition in their article proposing an update to the Cleary 1968 method. They define Test Bias as a "systematic error in how an assessment measures performance for a particular group" (Meade, 2009) in order to keep that definition similar to (albeit less rigorous than) the statistical concept of bias. Examinations will be made into whether or not the assessment predicts differently on the basis of race, gender, and the interaction between race and gender. It is important to note, and will be noted several times, that no charged statements can be

made if an indication of bias is found as this is only a preliminary study. More robust methods from experts in test theory and differential item functioning will be required in future studies.

Purpose of Study

This is a quantitative study using regression on PERT scores to predict first-year college GPA in order to detect Test Bias by Meade and Fetzer's revised Cleary method. PERT consists of three sections, Reading, Writing, and Math, so these three scores will be combined into a composite score and used as a predictor variable as is standard for placement tests. The sample will come from students that tested between 2014 and 2017 at Polk State College, a Florida College System institution located in central Florida. The other independent variables are race, gender, and the race-gender interaction effect. Differences in demographics by region mean that some race groups will have to be combined with others to have a detectable statistical outcome, but central Florida's composition at least guarantees a representation for the race groups white, black, and Hispanic. Gender will be limited to the male-female dichotomy for similar representation concerns. Race and gender are the two largest and most common demographic groups for which bias exists, so it makes sense to start with these for a preliminary study like this one.

Significance of Study

Florida has put enormous time, effort, and money into reforming higher education and access to it for its citizens. Part of its efforts resulted in the PERT as a common statewide placement test. It is the sincerest intention of this study to validate Florida's efforts and effect change in other states based on what has been observed. If, however, this study detects unsavory

issues and subsequent studies confirm that there is cause for concern, then it is hoped that Florida moves away from standardized tests for placement in favor of more qualitative methods. Whatever the outcome, it is most important that an instrument of assessment with such impact on citizen's lives be evaluated and the results available in the public domain.

Primary Research Questions

The overall research question can be summarized as "does PERT exhibit Test Bias as described by Meade and Fetzer?" To answer such a question takes many tests, so this study will be limited to the two largest demographic groups: race and gender. Due to PERT being divided into three subtests and students being permitted to skip a subtest if they just want to start at the lowest level for which they can earn college credit due to Florida law, it will be necessary to create a composite of PERT scores for each student. Also, PERT asks testers to self-identify their race and gender so those that chose not to identify in either category will obviously need to be excluded.

The following research questions need to be answered to answer the larger question:

1. Does PERT exhibit Test Bias with respect to gender?
2. Does PERT exhibit Test Bias with respect to race?
3. Does PERT exhibit Test Bias with respect to the race-gender interaction?

The process to examine these questions is lengthy. In short, if a common regression line does not provide the best fit of the entire data set, underlying factors will be examined. If group mean differences cause separate group regression lines with equal slopes but differing intercepts to be the best fits for the data, then PERT will be said to exhibit Test Bias. For example, if males

and females have statistically significant differences in PERT composite scores but their first-year college GPA is shown to not be statistically significantly different and the regression lines predicting them are shown to be "unequal" by the Cleary method, then it will be said that PERT exhibits Test Bias with respect to gender. Strong evidence suggesting Test Bias does not exist for race and does not exist for gender precludes the need to examine the interaction effect between race and gender, as noted in the Hypotheses section.

Hypotheses

Meade and Fetzer's update to Cleary's method requires accounting for differences within social groups before examining differences between social groups. For this study, that means ensuring the mean composite PERT scores for each social group are roughly equal, after which regression models for each group are created. These regression lines are first checked to make sure the intercepts are roughly equivalent and then regression coefficients corresponding to membership in a particular social group are tested to ensure there is not a significant difference. The primary hypothesis of this study is that Florida's PERT does not exhibit Meade and Fetzer's definition of Test Bias according to their updated Cleary method for race or gender.

Because each social group requires multiple models to be constructed for this process, the race-gender interaction will be tested if and only if Test Bias is weakly detected. For example, if males and females are shown to have similar placement rates but the regression coefficient for, say, Black students is barely within rejection criterion (within 0.001 of significance level), then race-gender regression models will be created and analyzed for the Black male and Black female students only. This was chosen because strong evidence that no Test Bias exists negates the need

to further subdivide the data to chase geese while weak evidence supporting it indicates a need to examine the underlying factors.

Research Design

This study is an exploratory analysis of prior tests results used to predict first-year GPA and test hypotheses. No persons will be interacted with directly but the data will be collected from the Institutional Research and Effectiveness Department of Polk State College, a Florida College System institution. Data points will come from a random sample of students that took the PERT between 2014 and 2017 with the following demographic breakdown:

22.5% Black	26.2% Hispanic	45.6% White	5.7% All Other Races
	43.6% Male	56.4% Female	

All participants will be above the age of 18, though the student body represented range from 16 to 85. No instruments are used as this is only a study of test scores taken long before the formulation of this study.

The procedure will unfold in four phases:

i. Collect PERT scores and first-year GPA for all students that enrolled at Polk State College between 1 August 2014 and 1 August 2017

ii. Test for significant differences in mean PERT scores between each group. This will mean an independent samples t-test for gender and ANOVA for race. (Predictor Test)

iii. Regress first-year GPA based solely on group and test for significant differences between the regression lines. (Criterion Test)

iv. Create multiple regression lines predicting first-year GPA from PERT scores and group and test for significant differences between the regression lines. (Intercept Test)

Interpretation of Tests of Bias will be made according to the chart from Meade and Fetzer's study, reproduced here (Table 1).

Table 1.

Interpretation of Tests of Bias

Intercept Test	Predictor Test	Criterion Test	Conclusion
Not significant	Not Significant	Not Significant	No bias
		Significant	No bias
	Significant	Not Significant	No bias
		Significant	No bias
Significant	Not Significant	Not Significant	Bias
		Significant	*No Bias*
	Significant	Not Significant	Bias
		Significant	*Maybe Bias*

Note: Italicized conclusions indicate different interpretations than those under the Cleary (1968) procedure.

Meade, A. W., & Fetzer, M. (2008, April). A New Approach to Assessing Test Bias

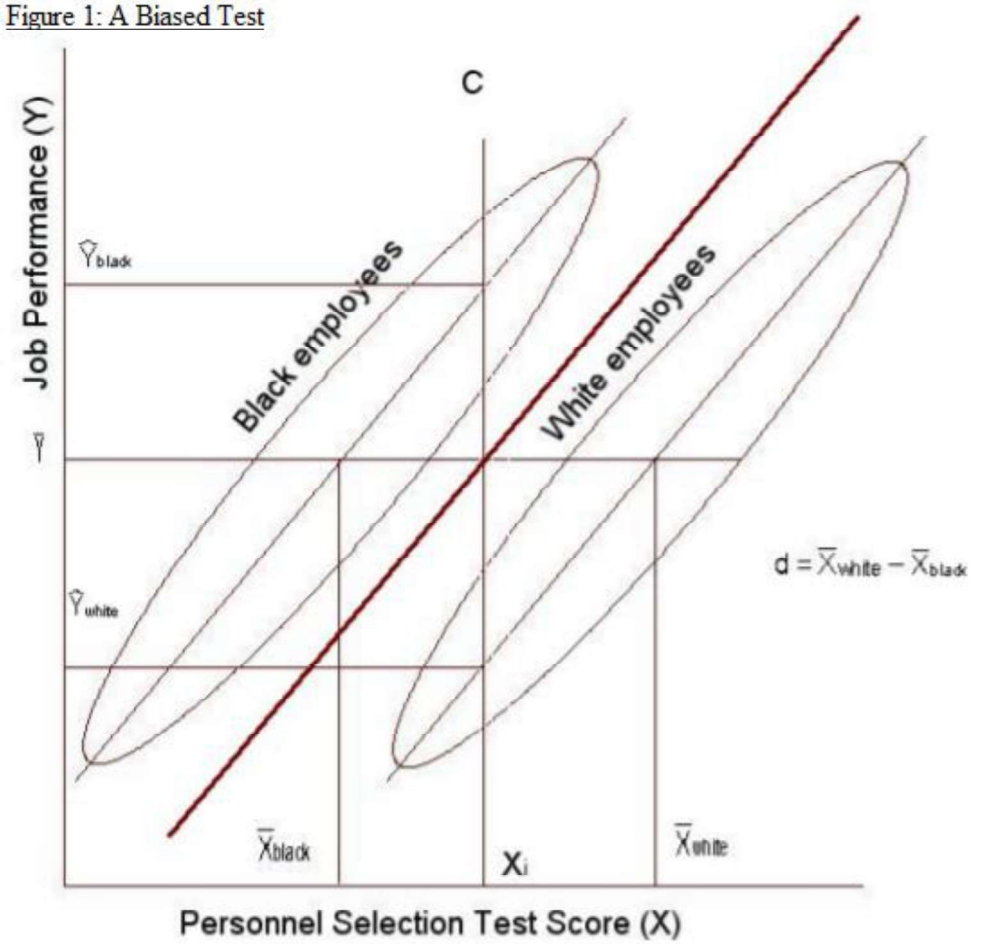
Theoretical Framework

T. Anne Cleary set her method on solid ground and her work has in turn become the solid ground upon which many other studies of bias have been based. Cleary's underlying statistical theory is based on the Gulliksen-Wilks Process, which tests if regression lines administered to different groups (but using the same predictors to predict the same criterion) are essentially equal. The paper Gulliksen and Wilks published even includes an illustrative example of predicting average college grades based on SAT scores for veterans and non-veterans (Gulliksen, 1950). This process is sequential, testing the hypothesis of equal variances between the groups before testing the hypothesis of the regression planes being parallel and ending at testing the hypothesis that the regression planes are identical. Any one test along the way resulting in statistical significance stops the process and concludes the two regression lines are not equal. Cleary adapted this process to racial groups, replacing the complicated tests using variance-covariance matrices with functionally equivalent hypothesis tests for equality of slopes and intercepts of the regression equations drawn separately for each group.

Intercepts and slopes in regression lines can differ for a variety of reasons, not all of which can constitute Test Bias. A major point of Cleary's method is that once the slopes are shown to be equal, statistically significant differences in intercepts indicate bias in the test. Meade and Tonidandel showed that, for equal slopes, intercepts can differ due to group mean differences on the predictor, group mean differences on the criterion, or group mean differences on both the predictor and criterion (Meade, 2010). They also showed that there can be group mean differences on both the predictor *and* the criterion but the regression lines can still have equal slopes and equal intercepts, matching Cleary's definition of unbiased. Of these, Meade and

Fetzer conclude that differing intercepts due to group mean differences on the predictor are the most indicative of bias, as this is the case where "groups differ on the test because of the influence of a construct that is irrelevant for criterion performance." (Meade, 2009) That is, a systematic error in how the test measures a particular group is altering the predicted outcome variable despite the true differences in the outcome variable not being significant. Figure 1 (Russell, 2000) below illustrates this for workplace assessment between two race groups.

Figure 1: A Biased Test



Russell, C.J. (2000) *The Cleary Model* White Paper

Here, the job performances are roughly equal between white and black employees, but a bias in the personnel selection test indicates that the white employee is more likely to be selected for hire due to being above the cutoff line (vertical line with C at the top). The common regression line (dark red, middle of plot) predicts higher Job Performance with a higher Personnel Selection Test Score. However, the regression line for black employees to predict their true Job Performance would need a different intercept than the one for white employees due to their group's poor performance on the Personnel Selection Test. This can accurately sum up the theory behind Meade and Fetzer's update to the Cleary method: if group mean differences in a predictor require separate regression lines for each group in order to best fit the data, then Test Bias is exhibited by the predictor.

This study places itself on Meade and Fetzer's shoulders. Using their method, variation within the social groups under consideration will be accounted for before drawing regression lines. The need for separate regression lines for social groups due to group mean differences in PERT scores indicate that Test Bias has been detected. Detection of Test Bias will be an indicator for more research to be conducted and cannot be viewed as a definitive ruling against PERT's use. More details about Meade and Fetzer's updated Cleary method is available in Chapter 3: Methodology.

Assumptions, Limitations, and Scope

The assumptions, limitations, and scope of this study are presented threefold. First, since the scores are being sampled from different points in time over a three year period, it is assumed that each sample is independent and that the underlying population has a normal distribution. Second, for brevity this study is limited to observing only the effect on prediction from race and

gender, with interaction effects between race and gender being considered. Therefore, any effects due to age, socioeconomic status, and whether or not the student transferred will not be taken into account. Any difference due to preparation from private tutors, the GPA if a student transferred halfway into their freshman year, or additional life experiences (which alone could probably constitute its own study) will be ignored for the purposes of this study. Finally, the scope of this project is very narrow: does PERT need to be more closely examined for bias? No claims of fairness or appropriate use of the exam can be derived from reading only this paper. This study takes general techniques from the sphere of predictive bias and applies them to detect a particular definition of bias: that of unequal prediction. More robust methods from the theory of measurement bias will be needed for future studies, as such techniques are beyond the scope of this paper and, indeed, this author.

One additional thing to note is that this study will examine predictions based on regressing first-year college GPA from PERT placement scores. This is standard practice in differential prediction techniques as first-year college GPA is usually the criterion the SAT and ACT predict for large university admissions offices. While universities can be selective, community colleges cannot, so placement tests are used in lieu of admissions tests. First-year college GPA may not be the full picture of academic achievement for many students, but the scope of this study is heavily limited and so it is necessary to adhere to tradition just this once. Future research may be able to align all facets of student life for more accurate models, but that is not the focus of a preliminary paper such as this one.

Definition of Terms

a. First-year college GPA - a student's GPA after completing two semesters, not necessarily consecutively, at Polk State College; if only one semester is recorded, the GPA for that semester is considered the First-year GPA while a GPA of 0.00 will be used for students for which Polk State College has PERT scores but the student did not attend any Polk State College classes or withdrew.

b. PERT - Postsecondary Educational Readiness Test; an untimed computer adaptive test consisting of a writing, reading, and math section used for college placement across Florida. NOTE: as of 2014, FL Senate Bill 1720 states that any active duty military or any student graduating with a standard high school diploma from a Florida high school after 2007 is exempt from taking PERT for placement; however, many student types such as non-traditional or students that graduated from a Florida high school before 2007 are still required to take PERT for placement.

c. Test Bias - the systematic error in how a test measures performance for a particular group; this is Meade and Fetzer's definition and will have both words capitalized when referenced throughout this text.

d. Group - a dichotomous variable indicating membership in a certain demographic such as race or gender; gender is split into the groups "male" and "female" while race is split into the groups "Black", "White", "Hispanic", and "Underrepresented" for the purposes of this study.

e. Predictive bias - the situation where an assessment predicts different outcomes based on membership in a particular group.

NOTE: this is not necessarily an indicator that a test is unfair, just that different groups are expected to have different outcomes. Fairness is an unexpectedly complicated issue and beyond the scope of this study.

Summary

The rise of standardized testing, like the rise of data and predictive analytics, has been meteoric and unique to the last century. History has seen colleges and universities creating the College Board to align their varied entrance exams into one national standardized test for admissions. Complex factors led to admissions being more certain and so these standardized tests became assessments of skill level, tools to categorize students and determine their starting point. This led to the current horizon of state-centric assessments, which has been around for high school graduation but is relatively new for college placement. But concerns over the possibility of occult segregation by these tools created a whole new field of assessment bias.

Assessment bias is a response to the rise of standardized testing. The field is split between measurement bias, methods to detect if properties of the assessment itself favor one group over another, and predictive bias, methods of determining if an assessment predicts different outcomes based on membership in a social group. T. Anne Cleary's 1968 study provided a framework lauded for detecting predictive bias, as she applied it to the SAT and determined bias did not exist or that it benefitted the minority group if it did. In part, this was due to the entwining of the definition of bias in the colloquial sense as "fairness" and the statistical term bias which is much more rigorously (and objectively) defined. Her method was updated by Meade and Fetzer in 2009 to account for variations that Cleary either was not aware could occur in the sixties or did not have the tools to assess. This update, while not violently upending the

field, untangled the ideas of bias as fairness from the statistical definition and provided a new structured definition for Test Bias, arguably their most important contribution. Yet Meade and Fetzer's update does not seem widely applied as of January 2021.

Next, this study will apply the Meade and Fetzer update to Florida's Postsecondary Educational Readiness Test. The literature review in Chapter 2 will focus on predictive bias, starting with the history of standardized testing in the United States, narrowing down to predictive bias as a field, and finally tapering down to regression methods, culminating in Cleary's work and its update. Chapter 3 will detail the methodology of this study, introducing each step in Meade and Fetzer's paper as applied to PERT scores collected from Polk State College. Results will be discussed in Chapter 4, noting the important outcomes and announcing what was detected according to the methodology. Finally, Chapter 5 will summarize the paper by briefly recounting what led to this point and suggesting paths forward. At this point and in several other places throughout this paper, the researcher would like to note that if Test Bias is detected it is **NOT** a definitive, once-and-for-all claim. Test Bias under Meade and Fetzer's definition is exploratory, meaning any announcement of its detection is preliminary and should incite a call for further exploration.

CHAPTER II: Literature Review

For context behind PERT and the method to evaluate bias, an overview of the relevant literature is presented here. This chapter will begin with a brief overview of standardized testing in the United States. Next, differential prediction and the several methods to detect bias will be introduced. Cleary's method of detecting bias will be discussed along with Meade and Fetzer's update to her powerful method. Finally, the Initiative that resulted in PERT will be briefly summarized as well as a few things learned since PERT's adoption in 2010. At the end, the highlights of the literature review will be summarized with the poignant points needed for Chapter 3: Methodology given particular attention.

Standardized Testing in the United States

When the 20th century gave way to the 21st, a retrospective of standardized testing was needed, which Richard Atkinson and Saul Gesier delivered. In 1901, The College Entrance Examination Board administered the first set of admissions tests standard to several colleges, called, confusingly, the "college boards" (Atkinson, 2009). These "college boards" were based on the curriculum of the consorted colleges and assessed students' preparation to tackle said curriculum. The SAT arrived in 1926 promising something new: a standardized assessment of students' general ability in an "easily scored, multiple choice instrument" (Atkinson, 2009). Problematically, an exam like the first iteration of the SAT is nothing more than an IQ test, born out of practices begun during World War I and deeply entrenched in the utopian eugenics movements of the time. Yet the concept of the SAT "resonated strongly with the meritocratic ethos of American college admissions" (Atkinson, 2009) and so began a new era of American higher education. Between 1926 and 1996, the SAT evolved from an aptitude test to a

generalized reasoning ability test to a critical thinking test, though the test still claims to be a "gauge [of] students' generalized analytic ability" (Atkinson, 2009). But Atkinson notes that the SAT actually had an adverse impact on low-income and minority students, ranking them for admissions in a lower caste than their high school GPA alone. Admittedly, Atkinson claims without citation or evidence that "high-school grades are the best indicator of student readiness for college and standardized tests are … a supplement" simply because high school GPA is, as Atkinson put it, a "repeated sampling of academic performance for an individual." This result may be indicated in other literature, but as Korbin et al noted in 2007, "there is substantial evidence supporting the notion of grade inflation," a factor which diminishes the reliability of high school GPA as a predictor. Still, Atkinson is correct in noting that the poor statistical power of the SAT renders the test about as useful as inflated high school grades. Even the updated SAT-R, an exam an hour longer and with a timed written essay, is not a better predictor of first-year college GPA (Atkinson, 2009). Then, in 1959, the ACT was introduced to compete with the SAT. The ACT was intended by its founder, E.F. Lindquist, to be an achievement test as a direct contrast to the concept of the SAT as an "intelligence" test. However, the ACT is still a statistically normalized test used as a tool to compare students for admission rather than measuring individual students' achievement ability. This "theoretical framework war" between the ACT and the SAT created spin-offs like a character on a sitcom, spawning the SAT II subject tests and the College Board's Advanced Placement Exam program. Atkinson believes this to be foolish, saying that "[t]he best examples of pure achievement tests now available are . . . standards-based assessments developed . . . to articulate clearer standards for what students are expected to learn" (Atkinson, 2009). This same notion was a motivating factor behind Florida's creation of PERT: an assessment tied to skills standards created by and for the state's K-12

system that could reliably place students at the level most appropriate for them in college. However, as even Atkinson notes, “prediction has captivated American college admissions” and so tests of placement or admissions are mostly designed in the same way: to predict performance.

Michael Nettles (Nettles, 2019) also prepared a retrospective on testing in the United States’ education system. He noted that standardized testing began as a way to reduce the burden of preparing for multiple entrance exams on both the students and the colleges. These exams also served a wide array of purposes: admissions, placement into remediation, guiding the choice of major, and many others (Nettles, 2019). However, the rise of standardized testing paved the road for controversy, as the SAT was designed by noted eugenicist and terrible racist Carl Brigham, infamous for his unfounded claim in a research paper that “future blended Americans will be less intelligent” than the segregated society of 1923 (Nettles, 2019). Standardized tests do show significant racial disparities in performance between white and black examinees on the SAT, ACT, and GRE, the three main admissions tests in the United States. Large shares of these differences are not able to be explained by family income, parents, or school district-level factors, commonly cited factors in the race group differences (Nettles, 2019). Nettles notes that exams have been an intricately woven part of the fabric of higher education in America since Harvard was founded and “score differences are not proof of bias.” Indeed, proving bias had to be developed using undeniable statistical techniques.

In their 2007 report for The College Board, Kobrin et al (2007) summarized nearly two decades of research on SAT performance for different subgroup populations. They noted that subgroup differences remained consistent between 1987 and 2007 despite the changes in 1994 and 2005 intended to make the exam more equitable. Female students traditionally scored higher

than their male peers on the verbal section of the SAT while the male students were able to shine in the mathematics section. However, the SAT is a better predictor of how female students will fare in college than it is for males, as a regression equation for females with just high school GPA underpredicts first-year college GPA and adding SAT scores corrects the underprediction (Kobrin, 2007). In terms of racial and ethnic differences, the SAT was found to predict the college success of Asian and white students better than black and Hispanic students and SAT scores were more highly correlated with high school GPA and first-year college GPA for Asian and white students than for black and Hispanic students. In fact, the SAT was found to overpredict first-year college GPA for black and Hispanic students (Kobrin, 2007). Kobrin noted that "the theory of stereotype threat is often cited as a contributor to mean differences." Stereotype threat, a theory by Claude Steele and Joshua Aronson, is the disruptive effect caused by experiencing a threatening (or high stakes) situation that highlights the awareness of a stereotype for minorities about whom a negative stereotype exists; Kobrin also stated that there has been no consistent demonstration of the existence of stereotype threat (Kobrin, 2007). A more likely explanation can be found in socioeconomic differences, as Kobrin's review of the literature showed that "the SAT is a better measure of parental income than of verbal or math ability." Sadly, the question about parental income is voluntary and since 2003 fewer and fewer test-takers have been willing to divulge that information. The most interesting part of Kobrin's report is that she compared SAT trends to trends from other exams like ACT and GRE and found that these SAT differences were consistent across the other exams as well. For this reason, it is expected that Florida's PERT will show a similar difference between race and gender without exhibiting Test Bias as Meade and Fetzer defined it. But in order to detect bias, there first need to be methods available to test for bias.

Bias in Selection and Differential Prediction

Historically, researchers agreed that a definitive way to determine if social groups were being segregated from the college experience by these standardized entrance exams. Methods such as differential item functioning and item response theory were developed as ways to assess biased constructs in the tests themselves (Meade, 2010) but such methods are usually conducted "in-house" by the makers of the exams. The field needed a way to assess cause for concern from only the test scores and outcomes, giving rise to the field of differential prediction, sometimes known as selection bias.

As early as 1972, several methods were available to detect selection bias from scores and outcomes, enough for Nancy Cole to compile them in a report for American College Testing. Cole outlined and discussed six methods common at the time: the quota model, the regression model, the Darlington model, the Employer's model, the Thorndike model, and the equal opportunity model. Each model had its own definition of bias that it was seeking to detect, usually conflated with "fairness" in terms of outcome. The quota model, for example, determined that a method of selection was unbiased (really they meant "fair") if the proportion of candidates selected was close to a proportion represented in the population under consideration. This would mean that a college for which 96% of the student body consisted of female students could be excused for only admitting 4% of males that applied simply because that proportion matched the established proportion of the student body. Obviously, this would not be ideal in every situation, as the method arose during Segregation and could be levied as an excuse to keep the status quo intact. Colleges in particular were more inclined to use the regression model, sometimes called the Cleary model (discussed later), because it was considered that (usually taxpayer-funded)

school resources should be used on those most likely to succeed. The Darlington model combined the spirit of regression and quota models to select those most likely to succeed from a pool of candidates matching a specified proportion. Use of the Darlington model first required the selector to decide "if there is special value in the selection of members of some cultural group" (Cole, 1972) and then regressing scores separately for each group to determine the most desirable candidates from each group. Darlington's model in not so much a detector of bias as it is a justifier of it, as when there is no reason to favor a certain population subgroup the model reduces to the regression model. The Employer's model is similarly unoriginal, as its definition of "unfair discrimination exists when persons with equal probabilities of success on the job have unequal probabilities of being hired" (Cole, 1972) is simply the Cleary definition with jargon for hiring practices inserted instead of generalized. The Thorndike model and the Equal Opportunity model are reworked definitions of the Cleary definition in terms of probabilities and are unnecessary to detail here. Overall, the models assess the chances of a candidate not being selected for something based on their membership in a population subgroup. Cole then presented several scenarios where someone needed to be selected from a pool of applicants and applied each model introduced to the situation. Using data from a 1970 Bowers paper, Cleary's 1968 paper, and a 1971 Temp paper, Cole determined that it was very rare for minority regression lines to actually be higher but parallel to majority regression lines or for their probabilities of selection to be lower than their probability of success (Cole, 1972). Her conclusion was that which method would be most appropriate to use would be based on the selector's need to preserve the rights of those not selected such as experienced in affirmative action situations. She did conclude that minorities were more likely to be favored by regression models such as the Cleary method, but overall it stands to be the best test of bias in selection.

John W. Young in 2001 compiled a comprehensive review and analysis with Jennifer Kobrin on differential prediction. In it, he reviewed significant meta-analyses before discussing differential prediction among social groups himself. First among the meta-analyses was Robert Linn's 1973 paper that summarized the work of Cleary, Temp, Davis, and Thomas. Linn documented the findings that a regression model based on white male student data would overpredict black male student outcomes and underpredict female outcomes, the first review paper to do so. Young used this to reject the hypothesis that the SAT was biased against minorities and that actual grades for these groups were lower than predicted. Next, Young reviewed Breland's 1979 monograph which itself summarized 35 regression model studies that mostly focused on race differences. Breland found 29 instances of significant differences between white students and a minority group in his report, confirming the result Young found in Linn that said the regression models based on white males over predicts outcomes for minorities. Young's reviews of a 1983 Duran article and a 1983 Wilson article report much of the same, seemingly cementing Young's stance that the SAT is unbiased. However, when he discusses the data himself he reports that the SAT has very low correlation to first-year college GPA for black and Hispanic students, the opposite finding for white and Asian students. Young notes that this creates a complex relationship between race and the prediction models. He is not afraid to admit that he does not know what it means if SAT overpredicts success but the SAT and first-year college GPA are nearly uncorrelated like happens for black and Hispanic students (Young, 2001). His discussion about differences between males and females did not delve too deeply, noting that females were routinely underpredicted by regression models though not to the same magnitude of difference as seen for the racial comparisons. Young concluded that the SAT was a case of differential prediction that was diminishing as time went on, though he was unable to

identify the cause of diminishment. . Nonetheless, group differences continue to occur but the validity of the test is still maintained (Young, 2001).

Differential prediction, however, is not always a case of bias. In some instances, like with the Darlington model, it may be favorable to choose more from the minority than from the majority. When it becomes an issue, however, is when selections are made based on this differential prediction unknown to the selector or, worse, with the selector being fully complicit and intending to segregate social groups. The regression model developed by Cleary is beneficial, and may be argued to be the best choice, for examining bias in the selection process.

Cleary in 1968 to Meade and Fetzer in 2009

T. Anne Cleary set the standard for regression-based bias testing in her seminal 1968 paper. In it, Cleary modeled separate regression lines for black students and white students in three integrated colleges. She used SAT score and high school rank-in-class to predict first-year college GPA for both groups. Using a framework built around the Gulliksen-Wilks Procedure (1950), Cleary tested the equality of slopes for the black students model and the white students model and if that was not statistically significant then she tested the equality of the intercepts. Bias was said to be found if the intercept difference was statistically significant. Her definition of bias in this paper, stated as "consistent nonzero errors of prediction," was found for only one of the three schools, which determined the test was "biased in favor of the [black] student" (Cleary 1968). This method has since been hailed as the golden standard for detecting bias in an assessment, being enshrined in national standards (AERA et al, 1999) and used to evaluate everything from admissions testing to employee evaluations. However, Cleary noted that "general conclusions cannot be reached" by her initial study and indeed there are a few flaws in

her method. One such flaw is the identification of race. Because colleges did not record race data at the time, researchers were shown black and white pictures of the students under study and identified the race on their own judgment; indeterminable photos defaulted to white (Cleary, 1968). Another flaw is that there is no accounting for in-group variation, which Adam Meade and Michael Fetzer corrected in their revision to the Cleary method (Meade, 2009). Regardless, Cleary's method is both the most-well known and most widely implemented test of bias using methods from differential prediction.

However, William Terris argued that Cleary's model is fundamentally wrong from the start and the method itself is biased against minorities. Terris' argument hinged on the fact that Cleary's model assumes the criterion or dependent variable is unbiased and is looking to detect bias in the predictor or independent variable, something Terris argued could not always be assumed. Therefore, if a test is unbiased it "should produce identical regression lines for all groups" (Terris, 1997), something Terris says cannot happen due to specific variation and factorial asymmetry. In essence, Terris laid out the technical reasons why selection bias can not properly be detected by the Cleary model if there is any measurement error in the criterion or predictor due to specific variation or factorial asymmetry. He backed his claims with graphs depicting the mathematics of how these hidden factors rotate regression lines for subgroups and could lead to false negatives for bias detection. Problematically, however, his mathematical proofs used no real-world data and he made no attempt to distinguish the effects of factorial asymmetry or specific variation in his graphs, a fact he stated in his publication (Terris, 1997). For this reason, it is hard to know if the claims he levied against Cleary's model would ever happen in real world scenarios. Still, he was aware that the traditional regression model needed to be updated and recognized that differences in distribution of predictor values for population

subgroups needed to be accounted for in future modeling situations. This is exactly what Meade and Fetzer did a decade later.

Adam Meade and Michael Fetzer recognized the Cleary method, powerful as it was, needed an update. Their paper revising her method changed very little of the process, but its most important contribution was that it defined bias in a way separate from fairness that could be more readily used. Test Bias, in their definition, is the "systematic error in how an assessment measures performance for a particular group" (Meade, 2009). This definition allowed them to separate out possible causes of differing intercepts and bolster the Cleary method, which they applied to job performance assessment data for black and white clerical workers. By noting that differences within the population subgroups affect the calculation of the regression coefficients, Meade and Fetzer were able to show that an assessment was not suitable for use if, under the Cleary method, the intercepts and slopes have a statistically significant difference but the criterion predicted does not. This was important, as a test cannot be called biased if it predicts different results for different groups and those differences are true observed phenomena. For that reason, it is Meade and Fetzer's method, including their definition and determination process for Test Bias, that is used in this paper.

Florida's Postsecondary Education Readiness Test

The Postsecondary Education Readiness Test came to be as a result of Florida's 2008 College and Career Readiness Initiative. Florida was seeking to align its K-12 education standards to its college and university systems entrance requirements to bridge the gap in student achievement (FLDOE, 2010). Faculty from colleges and universities gathered with teachers from the K-12 system and matched the expected college developmental education outcomes with high

school exit standards to achieve a common definition of "college and career ready" (FLDOE, 2010). This started with Florida's Statewide Course Numbering System, a uniform course code database that facilitated transfer of credits between any Florida college and university that offered the same course code. As such, the introductory college courses such as ENC1101 - College Composition and MAT1033 - Intermediate Algebra can be assumed to be the same at any public college or university in Florida, permitting the workshop group to align those competencies to high school standards. PERT grew from this alignment and the Florida Legislature's requirement that the state college system implement a statewide common placement exam (FL Senate Bill 1720). The group proposed a computer adaptive test with math, reading, and writing sections, each having 30 questions and no time limit, which they piloted in the summer of 2010. McCann, the company utilized for assistance in designing the test, provided psychometricians to help define cut scores for placement levels. These cut scores were based off the ACCUPLACER cut scores previously used in Florida and were intended to be interim only until enough data had been collected to update these cut scores. As of January 2021, these cut scores have not been updated but there is no indication why the update has not happened.

Mokher and Leeds determined that Florida's efforts had no substantial impact on longer term college success. Their 2019 paper used regression discontinuity analysis to measure persistence toward a college degree, which they defined as a dichotomous outcome variable of "still in college" or "not still in college" two years after high school graduation (Mokher, 2019). The result was that increased administration of PERT in high schools to show deficiencies that needed bolstering before college and then using PERT for placement in college only increased persistence by 1.8% (Mokher, 2019). Mokher and Leeds do not fully understand the reason behind this lack of increased persistence, but they attribute it to a deficit in the theory of action

behind the College and Career Readiness Initiative: showing students that they are not college-ready in 12th grade is more likely to demoralize than to motivate. However, the authors acknowledge that persistence is a multifaceted variable and note that PERT indicating a student is "college-ready" does not necessarily mean that they will be successful in college. Of additional interest is the chart in Appendix A of their study that found effect sizes greater than 0.05 standard deviations from baseline for black, Asian, Hispanic, and female students but the authors did not discuss what this could mean. The present study will use those population subgroups in an attempt to locate Test Bias in the PERT, so it is of interest that those subgroups differ from the white, male baseline in effect size. Such a finding could indicate problems with PERT's predictive validity.

In her 2018 dissertation, Alisa Murphy Žujović sought to test PERT's predictive validity as it relates to entry-level math courses at Hillsborough Community College (HCC). Her study was quantitative, using hierarchical linear models in regression analyses to predict final course grade in HCC's four possible entry-level math courses: Developmental Math 1, Developmental Math 2, Intermediate Algebra, and College Algebra. She used student level predictors race, gender, PERT math score, full-time versus part-time enrollment, whether the student was first-time in college, and age for her models and course level predictors such as part-time versus full-time instructor, years of instructor experience, day class versus night class, and delivery mode (online versus in-person). Žujović found that PERT was a significant predictor in final grade for entry-level math class but did not detect any differential prediction, meaning race and gender were not significant predictors in her models. However, she notes that her study only examined one college, HCC, and so is only evidence that HCC uses PERT in a "fair" way; statewide analyses need to be undertaken. The theoretical framework appears to only be the theory behind

multilevel models, which the author notes as being a powerful tool to compare multiple nested groups and recommended by researchers Ma, Ma, and Bradley (Žujović, 2018). While specific and enlightening, Žujović's study examined the predictive validity of PERT for math courses and not looking for signs of overall Test Bias like the current study. Nonetheless, she agrees with this researcher that the available literature on PERT is lacking and that the state of Florida needs to begin more in-depth analysis into PERT. Where this study differs is that a more established procedure will be implemented to determine if a specific definition of bias is evident in PERT overall, not just as it relates to math courses.

Summary

In this chapter, bias in testing and its necessity have been reviewed as well as a preview of Florida's efforts in higher education. The history of the SAT and its offshoots began the chapter, which lead into a discussion of differential prediction. Several theories in differential prediction, sometimes called selection bias, were introduced and the regression model became the focus. Cleary's model, the archetype of the regression approach to bias detection, was detailed along with its shortcomings and Meade and Fetzer's updated version. Finally, the movement of the Florida legislature that began PERT was discussed and issues about its effectiveness and PERT's validity were discussed. This is a good historical pathway to understand the origins of standardized testing, the rise of the field of differential prediction, and the collision course of the Meade and Fetzer Updated Cleary Model with PERT.

It is important to note that Cleary's model was not perfect in its first iteration. As noted by Terris, the flaws in her model can lead to claims of bias against the majority of the population when the test is truly unbiased. Worse, tests actually biased against minorities could be

determined to be unbiased. (Terris, 1997). This is why the Meade and Fetzer update is so important. Trying to equate two regression lines when one has large variation in its predictor values and the other does not will almost certainly lead to a declaration of bias by the original method. Accounting for these variations is a vital first step in being able to determine the relative equality of two regression lines, followed by applying the model to PERT data.

Chapter 3 will detail the steps behind the Meade and Fetzer Updated Cleary Model and how it was applied to a sample of PERT scores and first-year GPAs. Chapter 4 will examine the results of this application and answer whether or not PERT exhibits Test Bias as defined by Meade and Fetzer. Finally, Chapter 5 will summarize the project, closing this study but opening a new door into examinations of Florida's common placement test. While the first two chapters were focused on the background of the problem and contextualizing it, the rest of this paper will be focused on the future and any implications arising if Test Bias is discovered.

CHAPTER III: Methodology

The purpose of this study is to examine Florida's PERT placement test for Test Bias according to Meade and Fetzer's definition. To this end, Meade and Fetzer's updated Cleary method will be applied to PERT scores predicting first-year GPA and examining if the regression lines for each race and gender systematically under or over predict success based on a student's race or gender. Chapter 3 will describe the population under study by describing the sample obtained from Polk State College and how the sample was collected. Meade and Fetzer's updated Cleary method will be described and any changes made by the researcher justified. Finally, a summary will review key points from this chapter and preview what to expect in the chapter detailing the results. Again, as will be reiterated multiple times throughout this study, any detection of Test Bias is **NOT** a final condemnation of PERT but a call for more research to be performed. These methods only indicate that there is a systematic error in the test and the causes of such error must be determined through more robust methods.

Setting & Participants

Data was collected from Polk State College, an institution in the Florida College System, located in Central Florida along the I-4 Corridor between Tampa and Orlando. PERT scores and first-year GPA were collected for every student that sat the PERT exam between 1 August 2014 and 1 August 2017 along with their race and gender, which were self-identified as part of the PERT process. Absolutely no data that could be used to identify individual students was collected (such as name, age, birth date, etc.) in accordance with agreements made between the researcher and the Institutional Review Boards of both Shawnee State University and Polk State College. The possible participant pool from Polk State College came from a student body that is

43.6% male and 56.4% female while the race demographic breaks down to 45.6% white, 26.2% Hispanic, 22.5% black, and 5.7% all other races. Ages of students range from 16 to 85. Student demographic data came from Polk State College's *Annual Equity Update Report* (Polk State College, 2018).

The sample collected totaled 17,832 students with PERT scores ranging from 50 to 150 ($\bar{x}$=106.0, sd=12.48) and FYGPA ranging from 0.00 (F) to 4.00 (A) ($\bar{x}$=2.74, sd=1.15). All students are over the age of 18 due to laws surrounding who is required to take PERT and the time period chosen. Because demographics are self-reported for PERT, 887 students chose not to identify their race, gender, or both demographics. These students had to be removed from the sample, leaving a total of 16,945 students. This resulted in a sample with the following demographic breakdown:

Table 2: Demographic Breakdown of Sample

Demographic	White	Hispanic	Black	Underrepresented	Male	Female
n	8,779	3,913	2,957	1,296	6,594	10,351
%	51.8%	23.1%	17.5%	7.6%	38.9%	61.1%

Results are expected to generalize to the population of Florida citizens attending a Florida public community college. However, it is recognized that the demographics of each region in Florida are markedly different and results may not generalize as well to areas of South and North Florida. A priori power analysis through power analysis software G*Power estimates that a sample size of 12,810 is needed for the procedure. Further power analysis indicates that the collected sample size of 16,945 should result in a power between 75% and 90%. Andy Field in Discovering Statistics using R states that ".80 is an acceptable power to aim for" so this is an

acceptable range for power.

Procedure

Data was collected through Polk State College's Department of Institutional Research, Effectiveness, and Accreditation. Permission was obtained from Shawnee State University's Institutional Review Board to conduct the study and from Polk State College's Institutional Review Board to collect the sample (approvals located in the appendix). A request was made to Institutional Research for an Excel spreadsheet of all students that took the PERT between the dates 1 August 2014 and 1 August 2017 that included the students' scores, race, gender, first term of enrollment at Polk State College, and GPA after two terms (or one if only one was completed). Students that took the PERT at Polk State that started classes but withdrew and never returned were recorded as having a 0.00 GPA. Confidentiality has been maintained by keeping this spreadsheet on a removable drive that is locked in the researcher's personal safe (except when it is taken out to analyze the data) and will be deleted immediately upon completion of this thesis. Additionally, it was requested that the Department of Institutional Research not include any data beyond the data points requested to minimize the chance of a breach of confidentiality.

Data Processing & Analysis

In order to research Test Bias in PERT, a sample of 17,832 students was requested from Polk State College. This data was first cleaned in Excel: any students that did not report their race or gender were removed as these are the primary variables under study. This resulted in a final sample size of 16,945. The goal under the Cleary method and its Meade-Fetzer update is to create regression equations for each subgroup and determine if they are equivalent. Data was

analyzed using R statistical software according to the Meade-Fetzer update, described in the next few paragraphs. Each student had their subsection scores arithmetically averaged to create a Composite variable of their PERT scores. This is similar to the process used by ACT (American College Testing, 2021) and is preferable to using a primary subscore and the other two as covariates since the Cleary method was designed and validated using one composited SAT score (Cleary, 1968). However, unlike SAT, students can opt to take all three, just two, or only one of the PERT subsections, so they could not be added like the SAT composite scores are. Hence, the choice to composite like ACT.

First, differences in predictor values must be assessed for later interpretation. For gender, which has only two levels, a simple t-test for independent samples can determine if there is a statistically significant difference in mean PERT Composite scores by gender. Race, however, has four levels in this study. An ANOVA to determine if there are significant differences between each race category is the most efficient, though it will only determine if significant mean differences exist. Tukey t-tests can be used post hoc to determine individual paired differences. In both cases, Meade and Fetzer recommend computing a Cohen's d effect size in the common way:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{sd_{pooled}} \qquad \text{[Eqn 2.1]}$$

where

$$sd_{poooled} = \sqrt{\frac{1}{2}(s_1^2 + s_2^2)} \qquad \text{[Eqn 2.2]}$$

(Lakens, 2013). This will be the d_x used in the final interpretation below and will have to be computed separately for each race category determined to be statistically significantly different

from baseline. This is different from normal, as ANOVA usually requires an Eta-squared effect size calculation. However, using ANOVA is supplemented with Tukey multiple comparisons t-test and it is truly the results of the Tukey procedure that will be used in the final calculation, so this Cohen's d is acceptable.

Second, differences in criterion values must be assessed. Unlike predictor differences, the Meade-Fetzer update calls for regressing the criterion using the social group as the predictor. This means a regression line predicting FYGPA from gender and a regression line predicting FYGPA from race are created. Statistically significant regression coefficients indicate the race or gender associated with the coefficient has a statistically significant difference from the baseline group. Once again, Cohen's d effect sizes are computed in the common way [Eqn 2.1]. This will be the d_y used in the final interpretation below. Also like above, d_y will have to be computed for each race category determined to be statistically significantly different from baseline.

Finally, criterion is regressed using predictor and group membership as the predictor variables. Regression lines are created predicting FYGPA from Composite, gender, and the Composite-gender interaction and predicting FYGPA from Composite, race, and the Composite-race interaction. Statistically significant interaction effects indicate significant differing slopes for regression lines based on group membership while statistically significant race or gender coefficients indicate significant differing intercepts. The case of differing slopes is an immediate conclusion of Test Bias as it shows that belonging to a particular social group changes predicted success solely by belonging to that social group. If the slopes are not shown to be different, then difference in intercepts is assessed according to results from the other tests (shown in chart below).

Table 3: Interpretation of Test Results (modified from Meade and Fetzer, 2009)

Slope Differences	Intercept Differences	Predictor Differences	Criterion Differences	Conclusion
Significant	-	-	-	Test Bias
Not Significant	Not Significant	Not Significant	Not Significant	No Test Bias Detected
			Significant	No Test Bias Detected
		Significant	Not Significant	No Test Bias Detected
			Significant	No Test Bias Detected
	Significant	Not Significant	Not Significant	***Test Bias***
			Significant	No Test Bias Detected
		Significant	Not Significant	***Test Bias***
			Significant	***Test Bias if*** $\boldsymbol{d_y < \widehat{d}_y}$

To compare against observed differences d_y, expected difference $\hat{d}_y$ is computed according to the following formula:

$$\hat{d}_y = \frac{\frac{2r^*_{yx}d_x}{\sqrt{d_x^2+4}}}{\sqrt{1-\left(\frac{r^*_{yx}d_x}{\sqrt{d_x^2+4}}\right)^2}} \qquad \text{[Eqn 2.3]}$$

where d_x is the observed standardized predictor difference (computed in the first step) and the observed correlation between the criterion and predictor corrected for correlation between predictors is given by

$$r^*_{yx} = \frac{r_{yx}}{\sqrt{r_{xx}}} \qquad \text{[Eqn 2.4]}$$

where r_{xx} is the correlation between predictors. This is Meade and Fetzer's own derivation and can be viewed in the appendix to their 2009 paper *Test Bias, Differential Prediction, and a Revised Approach for Determining the Suitability of a Predictor in a Selection Context*. For gender, this expected difference can be computed once but for race it will have to be computed for each group that is statistically significantly different from baseline. This is due to the original Cleary method and its Meade-Fetzer update only testing differences between one pair of social groups whereas this study needs to test six different pairings.

Summary

This chapter summarized the major details of the study. First, the population from which the sample was drawn was detailed, including demographics and geography. Next, the steps taken to collect the sample and ensure confidentiality were discussed. Finally, the Meade-Fetzer update to the Cleary method was detailed, including how the data was cleaned and analyzed. The results of applying this method to the collected sample will be discussed in Chapter 4.

Next, Chapter 4 will discuss the results from the method applied to collected PERT data. The chapter will be organized according to the stated research questions from Chapter 1: Test Bias for Gender, Test Bias for Race, and Test Bias for Gender-Race Interaction. As a reminder from Chapter 1, the Gender-Race Interaction will only be tested for Test Bias if one of the Race categories is within 0.001 of the significance level. This is to save time, as significant indicators of Test Bias for Race mean that Test Bias is exhibited for both males and females of that race while being barely above or below the significance level means there could be Test Bias for, as an example, female Hispanic students that does not exist for Male Hispanic students. Such an occurrence is worth investigating, and a cursory observation according to the methodology outlined in this chapter is well within the scope of this study.

Finally, Chapter 5 will summarize the results of this study in (mostly) plain English. Whether or not Test Bias was discovered will be stated as well as what implications that may have for the PERT in the future. Next steps will be recommended based on the results of this study, which are likely to be internal methods such as differential item functioning and item response theory. Again, the researcher must clearly state that any determination of Test Bias according to Meade and Fetzer's updated Cleary method is **NOT** a condemnation of the PERT but merely a

call for more research to be conducted. This will be repeated a final time in Chapter 5 when recommended future research is discussed.

CHAPTER IV: Results

This chapter contains the results of the statistical tests discussed in Chapter 3. In order to make a determination of Test Bias, the Meade-Fetzer update was applied to the dataset for Gender and Race individually and for their interaction. The descriptive statistics for each are presented first, followed by the results for each step of the procedure for Gender and for Race in that order. If either Gender or one of the Race categories is within .001 of the significance level then that Race-Gender interaction effect will be tested as well. Finally, this chapter will conclude with a brief summary and preview of Chapter 5, which will summarize this study. For convenience, Table 3 from Chapter 3, which is used to interpret the statistical tests and conclude the existence of Test Bias, is reproduced at the end of this chapter after the summary section. As has been said before in this paper and will be said again, any detection of Test Bias is NOT a condemnation of the Florida PERT but a call to study the test with more robust methods.

Descriptive Statistics of the Sample

Of the 16,945 students in the sample, there were 6,594 males (38.9%) and 10,351 females (61.1%). The Composite score for males ranged from 50 to 150 ($\bar{x} = 106.9, sd = 12.46$) as well as for females ($\bar{x} = 105.3, sd = 12.36$). First-year GPA ranged from 0.00 to 4.00 for males ($\bar{x} = 2.60, sd = 1.161$) and for females ($\bar{x} = 2.81, sd = 1.141$).

In terms of the racial demographic, there were 8,779 White students (51.8%), 2,957 Black students (17.5%), 3,913 Hispanic students (23.1%), and 1,296 students from Underrepresented race groups (7.7%). Composite scores ranged from 50 to 150 for White ($\bar{x} = 108.6, sd = 11.40$), Black ($\bar{x} = 99.1, sd = 12.71$), and Hispanic students ($\bar{x} = 104.5, sd = 12.10$) but ranged from 65.7 to 150 for students of Underrepresented race

groups ($\bar{x} = 107.7, sd = 12.90$). First-year GPA ranged from 0.00 to 4.00 for White ($\bar{x} = 2.82, sd = 1.112$), Black ($\bar{x} = 2.43, sd = 1.231$), Hispanic ($\bar{x} = 2.71, sd = 1.156$), and Underrepresented race ($\bar{x} = 2.85, sd = 1.121$) students.

The interaction between Race and Gender creates a new variable. Because there are 8 levels to this new variable, the descriptive statistics are summarized in Table 4 below.

Table 4: Descriptive Statistics for the Race-Gender Interaction Variable

		Composite			FYGPA		
Level	n (%)	Mean	SD	Range	Mean	SD	Range
White Male	3547 (20.9%)	109.5	11.46	50.0-150	2.69	1.139	0.00-4.00
White Female	5232 (30.9%)	107.9	11.32	50.0-150	2.91	1.084	0.00-4.00
Black Male	1068 (6.3%)	99.6	12.85	54.7-150	2.30	1.210	0.00-4.00
Black Female	1889 (11.1%)	98.9	12.62	50.0-150	2.50	1.236	0.00-4.00
Hispanic Male	1469 (8.7%)	105.7	11.98	53.3-150	2.56	1.161	0.00-4.00
Hispanic Female	2444 (14.4%)	103.8	12.12	50.0-150	2.80	1.143	0.00-4.00
Underrepresented Male	510 (3.0%)	107.3	13.13	66.0-150	2.77	1.094	0.00-4.00
Underrepresented Female	786 (4.6%)	107.9	12.74	65.7-149	2.90	1.136	0.00-4.00

Percentages may not total to 100% due to rounding

Does PERT Exhibit Test Bias for Gender?

For the analysis, "Male" was set as the baseline for Gender. This was not chosen for any specific reason, but by serendipity the first case in the sample was a White Male student. The programming behind the R software takes the first case to establish baselines for factors if not instructed otherwise and no reason to change the baseline was determined.

Step 1: The Predictor Test

To begin, an independent samples t-test was conducted to determine if significant mean differences occurred between Male and Female Composite scores. Levene's Test supported the assumption of equal variances ($F(1) = .082, p = .775$) for the t-test, and the Composite score appeared to be from a normally distributed population. A significant difference was indeed found ($t(16{,}943) = -8.070, p < .001$). A Cohen's d_x effect size was computed as

$$d_x = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{2}(s_1^2 + s_2^2)}} = \frac{106.9 - 105.3}{\sqrt{\frac{1}{2}(155.29 + 152.83)}} \approx 0.127 \qquad \text{[Eqn 4.1]}$$

Step 2: The Criterion Test

Next, FYGPA was regressed by Gender to determine if the regression coefficient was significant. Analysis in R showed that the regression coefficient corresponding to the student being Female was statistically significant ($\beta_1 = .21, t(1) = 11.62, p < .001$). This can be interpreted to mean that, on average, Female students are predicted to have a FYGPA 0.21 units higher than Male students. A Cohen's d_y effect size was computed as

$$d_x = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{2}(s_1^2 + s_2^2)}} = \frac{2.6 - 2.8}{\sqrt{\frac{1}{2}(1.35 + 1.30)}} \approx -0.183 \qquad \text{[Eqn 4.2]}$$

Step 3: The Slope and Intercept Test

Finally, FYGPA was regressed on Composite, Gender, and the interaction between Composite and Gender. A student's Composite score was found to be a statistically significant ($t(1) = 26.93, p < .001$) predictor of FYGPA but neither their Gender ($t(1) = .01, p = .991$)

nor the Gender-Composite interaction $(t(1) = 1.74, p = 0.80)$ was found to be statistically significant.

Conclusion

After running the analyses, their statistical significance was checked against Table 3 from Chapter 3. Because the Gender-Composite interaction effect was not found to be significant in Step 3, that indicates that drawing separate regression lines for Male and Female students will not have significantly different slopes. Similarly, Gender not being significant in Step 3 indicates the intercepts will not be significantly different. Despite statistically significant differences in the predictor (Composite) and criterion (FYGPA), it can be concluded that the PERT test does not exhibit Meade and Fetzer's definition of Test Bias based on Gender.

Does PERT Exhibit Test Bias for Race?

For the analysis, "White" was set as the baseline race group. Similar to the test for Gender above, this was not done intentionally but by serendipity the first case happens to be a White Male student and R uses the first case to set baselines.

Step 1: The Predictor Test

To begin, it was expected to use ANOVA to determine if there were statistically significant differences between race groups. However, Levene's Test was significant $(F(3) = 30.35, p < .001)$ meaning the assumption of homogeneity of variance could not reasonably be assumed. Instead, Composite was regressed along Race and regression analyses used, which found statistically significant coefficients for Black $(\mathrm{t}(3) = -37.15, \mathrm{p} < .001)$,

Hispanic $(\mathrm{t}(3) = -17.63, \mathrm{p} < .001)$, and Underrepresented $(\mathrm{t}(3) = -2.52, \mathrm{p} < .05)$ students. Bonferroni-adjusted pairwise t-tests concluded that there was a statistically significant difference between each pairing of racial groups except between White and Underrepresented race students (see Table 5 below).

Table 5: Bonferroni-Adjusted Pairwise T-Tests, Composite by Race

Race Group Pairing	Absolute Value of Mean Difference	p-value	Cohen's d_x
White – Black	9.42	$< .001$	0.780
White – Hispanic	4.04	$< .001$	0.344
White – Underrepresented	0.89	$.11$	0.073
Black – Hispanic	5.38	$< .001$	0.434
Black – Underrepresented	8.52	$< .001$	0.666
Hispanic – Underrepresented	3.15	$< .001$	0.252

Step 2: The Criterion Test

Next, FYGPA was regressed by race group. Analysis found that the coefficient corresponding to Underrepresented students was not significant ($\beta_3 = 0.03, t(3) = 0.84, p = .399$) while the coefficients corresponding to Black ($\beta_1 = -0.39, \mathrm{t}(3) = -16.03, \mathrm{p} < .001$) and Hispanic ($\beta_2 = -0.11, \mathrm{t}(3) = -4.94, \mathrm{p} < .001$) students were statistically significant. Criterion differences and Cohen's d_y are summarized in Table 6 below.

Table 6: Pairwise Differences, FYGPA by Race

Race Group Pairing	Absolute Value of Mean Difference	Cohen's d_y
White – Black	0.39	0.333
White – Hispanic	0.11	0.096
White – Underrepresented	0.03	−0.026
Black – Hispanic	0.28	−0.236
Black – Underrepresented	0.42	−0.356
Hispanic – Underrepresented	0.14	−0.121

Step 3: The Slope and Intercept Test

Finally, FYGPA was regressed along Race, Composite, and the Race-Composite interaction effect. Like with Gender, Composite was found to be a statistically significant ($t(1) = 29.41, p < .001$) predictor of FYGPA. The Race variable was found not to be significant for Black ($t(3) = 1.21, p = .227$) or Underrepresented ($t(3) = -1.04, p = .298$) students but it was found to be significant for Hispanic students ($t(3) = 2.10, p < .05$). Similarly, the Race-Composite interaction effect was found to be not significant for Black ($t(3) = -1.83, p = .068$) and Underrepresented ($t(3) = 1.25, p = .211$) students but it was found to be significant for Hispanic students ($t(3) = -2.07, p < .05$).

Conclusion

After analysis, results were compared to Table 2 from Chapter 3. The Race-Composite interaction effect being significant for Hispanic students in Step 3 means an immediate conclusion of Test Bias can be made, since that result means separate regression lines drawn for Hispanic students and other race groups will have significantly different slopes. In both Black and Underrepresented students, the coefficient for Race-Composite interaction and the coefficient for Race being non-significant mean that, despite predictor and criterion differences for Black students, there is no evidence of Meade and Fetzer's definition of Test Bias with regard to race. However, there is evidence of Meade and Fetzer's definition of Test Bias with regard to race for Hispanic students.

Does PERT Exhibit Test Bias for Race-Gender Interaction?

Because none of the Race variables nor the Gender variable were within .001 of the significance level $\alpha = .05$, it is not necessary to perform the analysis for Race-Gender interactions. There was lack of evidence of Test Bias for Gender with high p-values as well as a determination of Test Bias for Hispanic students with a low p-value while Black and Underrepresented race students did not exhibit Test Bias with high p-values. Having p-values far enough away from the significance level of .05 is convincing enough evidence that interaction between gender and race is not masking any potential Test Bias in PERT. The Test Bias that exists is for Hispanic males and Hispanic females and the lack of Test Bias is lacking for males and females of Black and Underrepresented race with no reason to suspect it does also exist when the race groups are further subdivided by gender.

Summary

This chapter presented the numerical results of the statistical tests detailed in Chapter 3. First, descriptive statistics were used to describe the variables collected. Next, the results of each step in the Meade-Fetzer Cleary update were described for the research questions. Here, it was learned that Meade and Fetzer's definition of Test Bias was not evident between the genders Male and Female nor between White, Black, and Underrepresented race groups, but Test Bias was evident for Hispanic students, both Male and Female. Finally, it was determined that testing the Race-Gender interaction was unnecessary since the Test Bias that was uncovered (and the Test Bias that was not found) was not ambiguous enough to warrant testing divisions of the groups against each other. Again, it must be stressed that this discovery of Test Bias does NOT

condemn the Florida PERT as an unfair test, but instead demands the test be explored more deeply and completely with other available methods.

Chapter 5 will conclude this thesis with a summary of the contents and a discussion of related topics. The method, its application to the dataset, and a discussion of the results obtained will start that chapter. Then the connections back to the literature will be discussed. Ideas for future research in this area will be discussed as well as implications of this present study.

Reproduced: Table 3 – Interpretation of Test Results (modified from Meade and Fetzer, 2009)

<table>
<tr><th>Slope Differences</th><th>Intercept Differences</th><th>Predictor Differences</th><th>Criterion Differences</th><th>Conclusion</th></tr>
<tr><td>Significant</td><td>-</td><td>-</td><td>-</td><td>Test Bias</td></tr>
<tr><td rowspan="8">Not Significant</td><td rowspan="4">Not Significant</td><td rowspan="2">Not Significant</td><td>Not Significant</td><td>No Test Bias Detected</td></tr>
<tr><td>Significant</td><td>No Test Bias Detected</td></tr>
<tr><td rowspan="2">Significant</td><td>Not Significant</td><td>No Test Bias Detected</td></tr>
<tr><td>Significant</td><td>No Test Bias Detected</td></tr>
<tr><td rowspan="4">Significant</td><td rowspan="2">Not Significant</td><td>Not Significant</td><td>Test Bias</td></tr>
<tr><td>Significant</td><td>No Test Bias Detected</td></tr>
<tr><td rowspan="2">Significant</td><td>Not Significant</td><td>Test Bias</td></tr>
<tr><td>Significant</td><td>Test Bias if $d_y < \hat{d}_y$</td></tr>
</table>

CHAPTER V: Summary

The purpose of this quantitative, exploratory post hoc study was to determine if Florida's Postsecondary Education Readiness Test (PERT) showed signs of Test Bias as defined by Adam Meade and Michael Fetzer. Through their update to the Cleary regression method, a widely accepted regression line system to detect bias in assessments, this study examined what happened when PERT scores were used to predict first-year college GPA. Specifically, this study sought to answer the following questions:

1. Does the Florida PERT exhibit Meade and Fetzer's definition of Test Bias with respect to Gender?

2. Does the Florida PERT exhibit Meade and Fetzer's definition of Test Bias with respect to Race?

3. Does the Florida PERT exhibit Meade and Fetzer's definition of Test Bias with respect to the Race-Gender Interaction?

This chapter will discuss the major findings of this study as well as implications of these findings and how the study is situated in the literature. Also discussed are limitations of this study and suggestions for future research, including suggestions for future researchers to overcome these limitations when replicating this study. A brief summary will close the chapter and, finally, the study.

Summary and Interpretation of Findings

Meade and Fetzer define Test Bias as a "systematic error in how an assessment measures performance for a particular group" (Meade, 2009). The main focus of this study was to determine if the Florida PERT exam exhibited this definition of Test Bias with respect to the Male-Female gender dichotomy and/or the race demographic. Unfortunately, this specific definition of Test Bias was indeed detected for Hispanic students of both Male and Female gender. However, it is important to note that there was no evidence of Test Bias against either Black or Underrepresented Race students and this was determined without ambiguity regarding gender. That is, there is no reason to suspect that PERT exhibits Test Bias for Black Female students and not Black Male students, to name one example. Also, no evidence was found to suspect that one gender was favored over the other in general. However, the finding that Hispanic students' success is consistently under predicted is concerning since that is the largest minority group in Florida (FL OEDR, 2021). This does not, however, mean the PERT is without its merits.

This study validated Florida in at least one way: PERT scores are indeed significant predictors of college success. Alisa Žujović's 2018 dissertation, the only other readily available literature specifically about PERT as of January 2021, found that PERT Math subsection scores were significant predictors of success in a student's first math course (Žujović, 2018). That a Composited PERT score was generalized to predict overall first-year success, quantified as first-year college GPA, and was still determined to be a significant predictor indicates that the team behind the PERT ascribed to best practices as well as possible. However, Žujović did not find race or gender to be significant predictors in conjunction with PERT Math score while this study *did* find race to be significant. A necessary context is that Žujović studied scores from Hillsborough

Community College (HCC) which is located in and serves Hillsborough County, Florida, the county immediately west of Polk County, Florida, which Polk State College serves. Since the dataset analyzed in this study came from Polk State College, that Test Bias was found with regard to Hispanic students in Polk while not found in Hillsborough suggests that there may be confounding variables that need to be addressed.

Two additional concerns from the literature can be assessed from this study. Kobrin et al stated in their meta-analysis that standardized tests are better predictors of success for White and Asian students than for Black or Hispanic students (Kobrin, 2007). This study, in contradiction, has shown that Female, Male, Black, White, and Asian (Underrepresented) students' success is predicted by the PERT standardized test with equal regression lines as outlined in the Gulliksen-Wilkes Method (Gulliksen, 1950). However, it does agree that Hispanic students are not predicted equally with the other race groups, a problem hard to pin down since Hispanic alternates between being classified as a race or as an ethnicity depending on who is doing the research. Still, this assuages the concern from Nancy Cole (Cole, 1972) that regression models favor minority students over White students by showing the race groups (except Hispanic) are on equal footing in their first year because of where PERT placed them. This attests to just how significant Meade and Fetzer's update to the Cleary method truly was, as accounting for within-group variation reduced the factors that lead to minority groups being predicted to perform higher than White students in regression models. This also works with a position on the SAT from Young and Kobrin that actual minority grades were lower than predicted so the SAT could not be biased (Young, 2001). By their logic, if actual minority grades were higher than predicted, the SAT *could* be biased against minorities. For PERT, actual Hispanic grades were higher than predicted Hispanic grades, so it follows from their logic that PERT could be biased against Hispanic students.

Mokher and Leeds measured student persistence in college to determine that the efforts of Florida's Initiative had no long-term substantial impact (Mokher, 2019). Because this study found no evidence of Test Bias for Black, White, and Underrepresented race students, it is possible that Mokher and Leeds have the whole story already. However, since Florida has a large Hispanic population and this study found evidence of PERT under-predicting Hispanic student success, it is possible that Hispanic students being placed below their ability level contributed to demoralization and decreased persistence levels for that demographic group. Mathematically, it is obvious that such a situation could decrease overall persistence rates, so it is quite possible Mokher and Leeds' 1.8% persistence rate is actually artificially decreased through no fault of their own. Taking the results of this study into account, examining persistence rates of appropriately placed students could either validate Florida's current efforts or provide a more clear pathway for the state to focus on improvement.

Limitations and Recommendations

It goes without saying, of course, that there are some limitations to address as well as recommendations for overcoming those limitations. The most obvious is the threat to generalization. Why did this study find Test Bias for Hispanic students when no issue with race was detected in Žujović's dissertation conducted right next door? It could be because 17.7% of Polk county's population identifies as Hispanic while 24.9% of Hillsborough's does (FL OEDR, 2021). For the same time period, Polk State College reported 26.2% of its student body was Hispanic (Polk State College, 2018) while HCC reported 30% of the overall student body and 36% of first-time-in-college students identified as Hispanic (HCC, 2018). This indicates that the population of the county is not necessarily reflected in the population of the student body but does

not indicate any reason why, such as language barriers or scholarships to universities straight from high school for this group. A good suggestion for future researchers is to perform purposeful sampling to draw more accurate conclusions from deliberate, representative samples that account for different cultural and demographic makeups.

The difference in student population between the two colleges despite being neighbors also speaks to the possibility of a confounding variable, something the original Cleary method is accused of not taking into consideration (Terris, 1997). Student choice could be the biggest, as school reputation could draw students from one county to a school in a different county. However, it is impossible to know if students chose to attend HCC over Polk State or vice versa without asking them directly or including a variable to account for the students' county of residence. Such an action could help, but school choice is more likely a function of socioeconomic status. This study was unable to categorize students by socioeconomic status due to its scope, so future researchers may find gathering that information in addition to what was gathered for this study will be beneficial and in line with previous studies on standardized testing (Atkinson, 2009; Freedle, 2003; Mattern, 2009). Of course, this may not matter at all, so a first step is to conduct a study following this study's methodology for each of the 28 institutions in the Florida College System to see if the same situation appears to arise in every Florida college demographic combination.

The preceding limitations and recommendations are mostly academic in nature, but lawmakers and citizens may wonder what to do with this study in the immediate future. Vote, the simplest action that can be done. An overview of how the PERT has been used since its inception as well as what impacts that has had on Florida citizens is necessary if Florida is sincere in its

desire to create new, better educational protocols and systems. Citizens should vote for legislators intent on making that happen and legislators should vote on resolutions and proposals that do make that happen. But the first step is to validate the findings of this study by conducting similar (or hopefully better) studies at other Florida institutions. Then internal methods such as differential item functioning and confirmatory factor analysis should be undertaken to determine if the test itself really is the biased indicator. Better data scientists may even decide upon stratifying the scores by placement level and using categorical prediction instead of a continuous Composite score like performed here, though there did not appear to be supporting literature for such a method as of January 2021.

By far, the best thing that can be done is for colleges to examine a prospective student's entire academic portfolio instead of relying solely on standardized test scores. This is already the practice at Polk State College, the school from which the sample used in this study was drawn, and is supported by the literature (Belfield, 2012; Brothen, 2003; Soares, 2011; Woods, 2018). A complete, holistic picture of the student provides a more balanced view than one instance of test scores. This is especially pertinent since, like the SAT and ACT before it, students can retake PERT multiple times and there are study guides and tutoring services available for those that can afford them. However, this is just one more instance in which socioeconomic status must be accounted for, as students whose parents have higher disposable income are more able to participate in extracurricular activities or focus on their academic work just as they are more able to afford the extra study materials to score higher on the PERT.

Conclusion

It has been said ad nauseum throughout this study and now it will be said once more: any detection of Test Bias by this study is **NOT** a condemnation of the PERT but instead a call for further research. Just the very nature of how ubiquitous this test has become in Florida high schools and colleges should have been enough to ensure mounds of professionally conducted research on this test were available to the public; they were not. The original company, McCann, suggested that PERT's placement cut scores be updated once enough data had been gathered; they were not. Given the past century of standardized testing and the controversies that came with it, every available tool should have been used to show the public the PERT is the best placement test and that other states should follow suit; they were not. But there is no indication this was Florida's fault nor done maliciously, but rather that it seemed to be working and "if it ain't broke don't fix it" (Chery, 2020).

The specific definition of Test Bias that was discovered is not damning. PERT's formation was a multi-year intensive collaborative effort between many Florida educators and McCann and the test is still excellent at what it was designed to do…just not for Hispanic students. But McCann even noted that placement cut scores should be updated every few years just like the AERA/APA/NCME *Standards* say. Minute changes to the content of the exam may remove this unintended bias. Finally adjusting the cut scores could remove this unintended bias. Shifting to a holistic model of placement could remove this unintended bias. The author of this study makes no claim to know the perfect route Florida should take but reaffirms that the only purpose motivating the research was to start the conversation. It is curious that nothing further has been done to study PERT since its inception and implementation, but this can only be said about publicly available

studies. McCann or private researchers may very well have done extensive research on the test but not made it easy to locate through accident or intent. But the purpose, again, was to provide impetus for more study and it is hoped that has been achieved. Again, and for the final time, the detection of Meade and Fetzer's definition of Test Bias for Hispanic students is **NOT** a final condemnation of the PERT but merely a call to action, a call for more research to be performed.

.

REFERENCES

American College Testing (2021). Multiple ACT Scores. https://www.act.org/content/act/en/products-and-services/the-act-postsecondary-professionals/scores/multi-scores.html

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (1999). *Standards for educational and psychological testing*.

Atkinson, R.C., & Geiser, S. (2009). *Reflections on a Century of College Admissions Tests*. Research & Occasional Paper Series, Center for Studies in Higher Education. University of California, Berkeley. http://cshe.berkeley.edu

Belfield, C.R., and Crosta, P.M. (2012). *Predicting Success in College: The Importance of Placement Tests and High School Transcripts*. Community College Research Center, Teachers College, Columbia University. https://ccrc.tc.columbia.edu/media/k2/attachments/predicting-success-placement-tests-transcripts.pdf

Brothen, T., & Wambach, C. (2003). Is There a Role for Academic Achievement Tests in Multicultural Developmental Education?. Center for Research on Developmental Education and Urban Literacy, University of Minnesota General College. https://www.cehd.umn.edu/crdeul/pdf/monograph/4-b.pdf

Chery, S. (2020). *Florida bill could make college, university president searches private*. UNF Spinnaker. https://unfspinnaker.com/81595/news/florida-bill-could-make-college-university-president-searches-private/

Cleary, T.A. (1968), Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges. Journal of Educational Measurement, 5: 115-124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x

Cole, N.S. (1972). Bias in Selection. Research and Development Division, American College Testing. https://www.jstor.org/stable/1433996?seq=1

Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R*. SAGE Publications Ltd.

FL Dept. of Education (2010). Florida's Postsecondary Education Readiness Test. *Zoom* [Newsletter]. Florida College System. http://www.fldoe.org/core/fileparse.php/5592/urlt/0078245-zoom2010-03.pdf

FL OEDR: Office of Economic and Demographic Research (2021). County Profiles. http://edr.state.fl.us/content/area-profiles/county/index.cfm

FL Senate Bill 1720 (2013), https://www.flsenate.gov/Session/Bill/2013/1720

FL Senate Bill 1908 (2008), https://www.flsenate.gov/Session/Bill/2008/1908

FLA. STAT., Common placement testing for public postsecondary education, Title XLVIII Florida Statutes §1008.30 (2020). http://www.leg.state.fl.us/statutes/index.cfm?App_mode=Display_Statute&Search_String=&URL=1000-1099/1008/Sections/1008.30.html

Freedle, R. (2006). How and Why Standardized Tests Systematically Underestimate African-Americans' True Verbal Ability and What to Do about It: Towards the Promotion of Two New Theories with Practical Applications. *St. John's Law Review*, *80*(1), 183–226.

Freedle, R. O. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review, 73*(1), 1 – 43. https://doi.org/10.17763/haer.73.1.8465k88616hn4757

Gulliksen, H., & Wilks, S. (1950). Regression tests for several samples. *Psychometrika*, *15*(2), 91.

Hillsborough Community College (2018). *2018-2019 Annual Equity Update Full Report*. https://www.hccfl.edu/sites/default/files/docs/2021-05/2018-2019%20Equity%20Report_20210507-508.pdf

Kobrin, J.L., Sathy, V., & Shaw, E.J. (2007). A Historical View of Subgroup Performance Differences on the SAT Reasoning Test. College Board. https://files.eric.ed.gov/fulltext/ED562569.pdf

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4. DOI:10.3389/fpsyg.2013.00863

Leonard, D.K. and Jiang, J. (1995). Gender Bias in the College Predictions of the SAT. Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995). https://eric.ed.gov/?id=ED383736

Loewen, J.W. et. al. (1988). Gender Bias in SAT Items. Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988). https://eric.ed.gov/?id=ED294915

Mattern, K. D., & Packman, S. (2009). *Predictive Validity of ACCUPLACER Scores for Course Placement: A Meta-Analysis* (No. 2009–2). CollegeBoard. https://files.eric.ed.gov/fulltext/ED561046.pdf

Meade, A. W., & Fetzer, M. (2008). A New Approach to Assessing Test Bias. Paper presented at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, April 2009, San Francisco, CA

Meade, A.W., & Fetzer, M., (2009). Test Bias, Differential Prediction, and a Revised Approach for Determining the Suitability of a Predictor in a Selection Context. *Organizational Research Methods*. 12(4), 738-761. https://doi.org/10.1177/1094428109331487

Meade, A. W., & Tonidandel, S. (2010). Not seeing clearly with Cleary: What test bias analyses do and do not tell us. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*(2), 192–205. https://doi.org/10.1111/j.1754-9434.2010.01223.x

Medhanie, A. G., Dupuis, D. N., LeBeau, B., Harwell, M. R., & Post, T. R. (2012). The Role of the ACCUPLACER Mathematics Placement Test on a Student's First College Mathematics Course. *Educational and Psychological Measurement*, *72*(2), 332–351. http://rave.ohiolink.edu/ejournals/article/313171537

Mokher, C.G., & Leeds, D.M. (2019). Can a College Readiness Intervention Impact Longer-Term College Success? Evidence from Florida's Statewide Initiative. *The Journal of Higher Education*, 90(4), 585–619. https://doi.org/10.1080/00221546.2018.1525986

Nettles, M. T. (2019). History of Testing in the United States: Higher Education. *Annals of the American Academy of Political and Social Science*, *683*`(1), 38–55. https://doi-org.proxy01.shawnee.edu/10.1177/0002716219847139

Polk State College (2018), *Annual Equity Update Report*, https://www.polk.edu/wp-content/uploads/Polk-State-College-2017-2018-Annual-Equity-Update-Report.pdf

Robert L. Linn. (1973). Fair Test Use in Selection. *Review of Educational Research*, *43*(2), 139–161.

Russell, C.J. (2000). The Cleary Model: "Test Bias" as Defined by the EEOC Uniform Guidelines on Employment Selection Procedures. https://www.ou.edu/russell/whitepapers/Cleary_model.pdf

Santelices, M.V., & Wilson, M. (2010). Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning. *Harvard Educational Review*, 80(1), 106-133.

Soares, J. (2011). For Tests that are Predictively Powerful and without Social Prejudice. *Journal of Research and Practice in Assessment*, adapted from *SAT Wars: The Case for Test-Optional College Admissions*, Teachers College Press. https://eric.ed.gov/?id=EJ1062726

Terris, W. (1997). The Traditional Regression Model for Measuring Test Bias Is Incorrect and Biased against Minorities. *Journal of Business and Psychology*, *12*(1), 25.

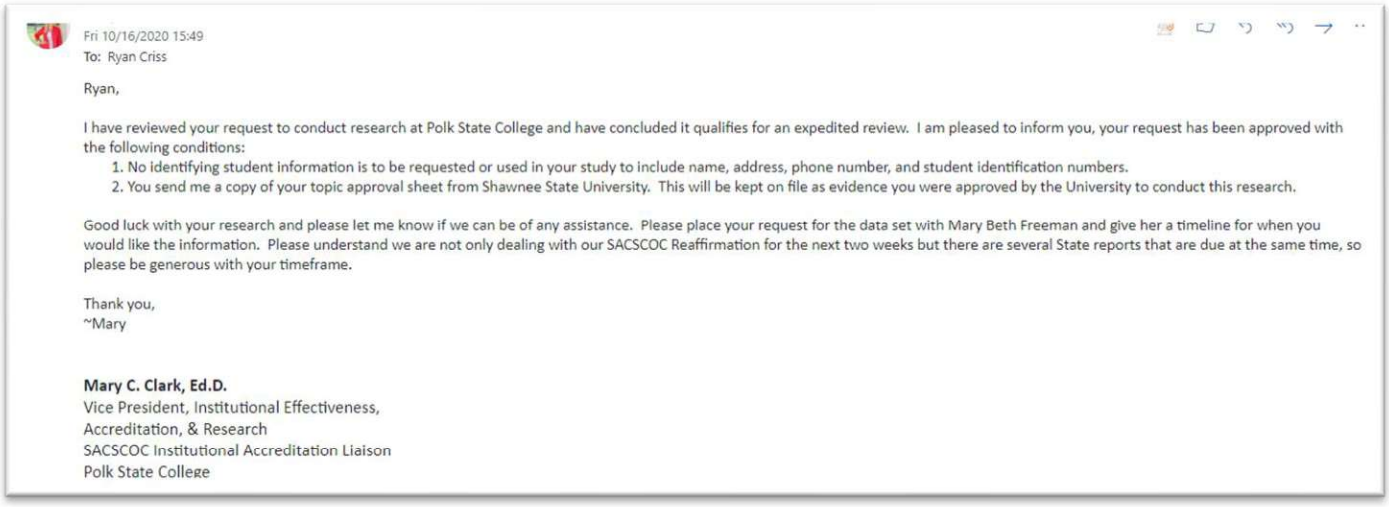
Young, J.W., & Kobrin, J.L. (2001). *Differential Validity, Differential Prediction, and College Admission Testing: A Comprehensive Review and Analysis*. College Board. https://files.eric.ed.gov/fulltext/ED562661.pdf

Woods, C.S., Park, T., Hu, S., & Jones, T.B. (2018). How High School Coursework Predicts Introductory College-Level Success. *Community College Review*. 46(2) 176-196. DOI: https://doi.org/10.1177%2F0091552118759419

Zujovic, A.M. (2018). Predictive Validity of Florida's Postsecondary Education Readiness Test. Graduate Theses and Dissertations. http://scholarcommons.usf.edu/etd/7253

Appendix A: IRB Approvals

Approval for this study was requested from the Institutional Review Boards of both Shawnee State University and Polk State College. The IRB chair at Polk State College is Dr. Mary Clark and the IRB chair for Shawnee State University is Dr. Tim Hamilton. Their approval emails are collected here.



Fri 10/16/2020 15:49
To: Ryan Criss

Ryan,

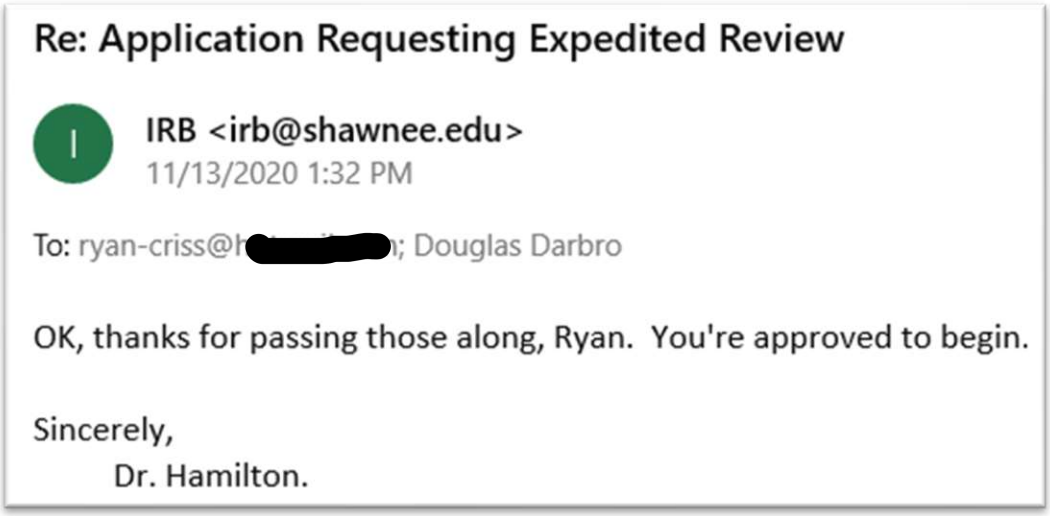
I have reviewed your request to conduct research at Polk State College and have concluded it qualifies for an expedited review. I am pleased to inform you, your request has been approved with the following conditions:

1. No identifying student information is to be requested or used in your study to include name, address, phone number, and student identification numbers.
2. You send me a copy of your topic approval sheet from Shawnee State University. This will be kept on file as evidence you were approved by the University to conduct this research.

Good luck with your research and please let me know if we can be of any assistance. Please place your request for the data set with Mary Beth Freeman and give her a timeline for when you would like the information. Please understand we are not only dealing with our SACSCOC Reaffirmation for the next two weeks but there are several State reports that are due at the same time, so please be generous with your timeframe.

Thank you,
~Mary

Mary C. Clark, Ed.D.
Vice President, Institutional Effectiveness,
Accreditation, & Research
SACSCOC Institutional Accreditation Liaison
Polk State College



Re: Application Requesting Expedited Review

IRB <irb@shawnee.edu>
11/13/2020 1:32 PM

To: ryan-criss@[redacted]; Douglas Darbro

OK, thanks for passing those along, Ryan. You're approved to begin.

Sincerely,
Dr. Hamilton.

CERTIFICATE

OF COMPLETION

PHRP Online Training, Inc. certifies that

Ryan Criss

has successfully completed the web-based course "Protecting Human Research Participants Online Training."

Date Completed: **2020-08-28** Certification Number: **2856141**

PHRP
Protecting Human
Research Participants
Online Training

Appendix B: PHRP Certification

BIBLIOGRAPHY

Ryan Clay Criss

Candidate for the Degree of

Master of Science Mathematics

Thesis: EVALUATING THE FLORIDA POSTSECONDARY EDUCATION READINESS TEST FOR BIAS: AN ADAPTATION OF THE MEADE-FETZER UPDATED CLEARY METHOD

Major Field: Mathematics

Biographical:

Personal Data:

Education: Bachelor of Science in Mathematics, University of Central Florida

Completed the requirements for the Master of Science in Mathematics, Portsmouth, Ohio in 2021.

Douglas G. Darbro, Ph.D.

ADVISER'S APPROVAL: Dr. Douglas Darbro