ABSTRACT OF CAPSTONE

Wendy Zagray Warren

The Graduate School

Morehead State University

March 5, 2015

JUSTICE FOR ALL? COSTS AND CONSEQUENCES OF STANDARDIZED TESTING REQUIREMENTS IN TEACHER EDUCATION REFORM POLICY

Abstract of capstone

A capstone submitted in partial fulfillment of the
requirements for the degree of Doctor of Education in the
College of Education
at Morehead State University

By

Wendy Zagray Warren

Berea, Kentucky

Committee Chair: Dr. David Barnett, Professor

Morehead, Kentucky

March 5, 2015

ABSTRACT OF CAPSTONE

JUSTICE FOR ALL? COSTS AND CONSEQUENCES OF STANDARDIZED TESTING REQUIREMENTS IN TEACHER EDUCATION REFORM POLICY

U.S. Teacher Education Policy has undergone rapid transformation in the U.S. in the last two decades (Darling-Hammond, Wei, Johnson, 2009). One area of change is in the now widespread use of standardized tests to evaluate pre-service candidates. Test results are increasingly used in high-stakes ways. Much has been written about the intent of these policies, which is to raise standards of teacher quality (CCSSO Task Force, 2012). The purpose of this study was to evaluate the impact and outcomes of these new policies, and particularly the high-stakes use of large-scale standardize test scores as if they might act as predictors of high-quality teaching. Methods of research reached across disciplinary boundaries in a historiographic study of the development and use of standardized testing in the U.S. educational system, selected psychometric properties of these tests, an evaluation of assumptions underlying considerations of valid use of the Praxis Series in teacher education, and an application of Critical Race Theory in order to analyze the outcomes of the use of such tests. The statistical models and the adaptation of the tests for educational use occurred during the U.S. Eugenics Movement (Au, 2013; Jensen, 2002; Norton, 1978). Based on known and predicted outcomes of any large-scale standardized test, the impact of the use of these tests as gatekeepers in teacher education is that the demographics of the pool of teacher educators will remain white and from the middle

and upper classes of U.S. society (Bennett, et al., 2006; Grant, 2004; Nettles, et al., 2011). This will have long-term implications for the success of students of color in U.S. schools (Goldhaber and Hansen, 2009) and will limit the perspectives of all students (Loewen, 2007), maintaining conditions which will allow for the continuation of a white norm in U.S. society (powell, 2012).

KEYWORDS: critical race theory, standardized testing, teacher preparation, Praxis Exams, privilege and oppression

_____
Candidate Signature

_____
Date

JUSTICE FOR ALL? COSTS AND CONSEQUENCES OF STANDARDIZED TESTING REQUIREMENTS IN TEACHER EDUCATION REFORM POLICY

By

Wendy Zagray Warren

Approved by

_____
Rocky Wallace
Committee Member Date

_____
Bobby Ann Starnes
Committee Member Date

_____
David Barnett
Committee Chair Date

_____
Chris Miller
Director of EdD Date

_____
Chris Miller
Department Chair Date

RULES FOR THE USE OF CAPSTONES

Unpublished capstones submitted for the Doctor's degree and deposited in the Morehead State University Library are as a rule open for inspection, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but quotations or summaries of parts may be published only with the permission of the author, and with the usual scholarly acknowledgements.

Extensive copying or publication of the capstone in whole or in part also requires the consent of the Dean of the Graduate School of Morehead State University.

A library that borrows this dissertation for use by its patrons is expected to secure the signature of each user.

Name Date

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

CAPSTONE

Wendy Zagray Warren

The Graduate School

Morehead State University

March 5, 2015

JUSTICE FOR ALL? COSTS AND CONSEQUENCES OF STANDARDIZED TESTING REQUIREMENTS IN TEACHER EDUCATION REFORM POLICY

Capstone

A capstone submitted in partial fulfillment of the
requirements for the degree of Doctor of Education in the
College of Education
at Morehead State University

By

Wendy Zagray Warren

Berea, Kentucky

Committee Chair: Dr. David Barnett, Professor

Morehead, Kentucky

March 5, 2015

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

DEDICATION

This project is dedicated to my students, past present and future, and especially to the students in the Teacher Education Program at Berea College. May your resilience, love for learning and hard work allow you to find places of leadership in our nation's schools. The future of our country depends on it.

I further dedicate this work to the fellow-educators in my family, my mother and brother, aunt and uncle; to my husband and to the memory of my father, both sources of unconditional love and ongoing support for my social justice work.

ACKNOWLEDGEMENTS

I am grateful to my Education Studies students at Berea College, whose experiences raised my awareness of the issues presented in this Capstone, and to the Education Studies faculty at Berea College, especially Dr. Bobby Ann Starnes, who has been a mentor to me since the beginning of my teaching career, and Dr. Kathryn Akural and Lisa Rosenbarker, with whom I had many conversations during the course of my research. I am thankful for the professors with whom I worked during my doctoral journey at Morehead State University, and especially the Chair of my Capstone Committee, Dr. David Barnett, who was a never-ending source of encouragement for my work.

I also want to express gratitude for my husband, Bob, and his unwavering love and support even as we moved across the country in order for me to shift the focus of my career, and for his enthusiasm about my work on this project, even though it limited our time together. I'm also grateful to my mother, Nayda Colomb, and my brother, Jim Zagray, who checked in often by phone to see how the work was going and to remind me to also set aside time for rest and for play. Now that the project is nearing its end, I look forward to more time for both.

TABLE OF CONTENTS

Executive Summary and Introduction

The Core of the Capstone

As a prime example of qualitative research, this Capstone evolved as an inquiry project, finding new directions as new questions emerged. What began as a project I thought would directly serve Berea College, my current educational home, has grown into a piece of writing intended to have a much broader impact. What began as a literature review intended to build understandings about various theoretical and conceptual frameworks guiding programs of teacher education led to an analysis of the effects of recent policy changes in teacher education in Kentucky. As the project continued, policy changes continued to be announced at the state and then the national level, raising ever more questions that guided the project's direction. As I researched, learned and wrote, I began making connections between pieces of information I had not seen connected in the literature. I began with a plan to focus on policy changes, assess their impact on the Berea College Education Studies Program and the College as a whole, and then propose possibilities for response at the college level. The project I thought I would pursue would be to design a non-certification Education Studies option to propose to the College faculty.

Instead, as I researched, I grew increasingly interested in one aspect of the policy changes in particular: the increasingly high-stakes use of standardized test scores as if they were predictors of a candidate's potential for teaching success or failure. Questions arose as I saw the futures of my students negatively impacted by

the difficulties many of them had in passing the required Praxis Exams. Some of these students had met, and sometimes exceeded, all other requirements for certification, yet they failed to pass the Praxis exam, barring them from teacher certification in Kentucky.

I began to research the psychometrics used in the creation of the tests themselves, as well as the history of their use in education in the U.S. My research became increasingly interdisciplinary, and the project morphed into historiographic research, utilizing methods of critical race theory. Researching across disciplinary barriers, I began to make connections from the history of the educational use of standardized testing in the U.S. to the demographic hierarchies created by test scores that mirrored the social and economic hierarchies of American society at large. I began to identify evidence of racism and classism that ultimately led me to my guiding questions: What is the impact, known and projected, of the high-stakes use of standardized test scores as gatekeepers in teacher education programs? What are the benefits of the use of standardized testing? What are the costs? Are the benefits worth the costs?

Context and Intended Audience

Context. Coming to understand the importance of these questions at Berea College in particular requires some background in the history of the College and the demographics of the attending students. In 1855, John G. Fee, Along with a group of "ardent abolitionists and radical reformers," (Strategic Planning Committee, 2011, p. 1) founded Berea College on the premise that, based on nonsectarian Christian values,

the college should be interracial and coeducational, a concept almost unheard of in that place and time. Berea's continued mission is defined in its Great Commitments, adopted in 1969 and revised in 1993 (Berea College, 2013).

Berea College commits itself

- To provide an educational opportunity primarily for students from Appalachia, black and white, who have great promise and limited economic resources
- To provide an education of high quality with a liberal arts foundation and outlook.
- To stimulate understanding of the Christian faith and its many expressions and to emphasize the Christian ethic and the motive of service to others.
- To provide for all students through the labor program experiences for learning and serving in community, and to demonstrate that labor, mental and manual, has dignity as well as utility.
- To assert the kinship of all people and to provide interracial education with a particular emphasis on understanding and equality among blacks and whites.
- To create a democratic community dedicated to education and equality for women and men.

- To maintain a residential campus and to encourage in all members of the community a way of life characterized by plain living, pride in labor well done, zest for learning, high personal standards, and concern for the welfare of others.
- To serve the Appalachian region primarily through education but also by other appropriate services.

(Strategic Planning Committee, 2011, p. 7).

Berea College is unique in its policy, based on these Great Commitments, of providing full tuition scholarships to each of its students. Berea is also a labor college, so in addition to attending classes full time, students work twelve hours a week in the labor program, assisting in every aspect of the operations of the college itself. Money earned through this labor helps offset costs of textbooks, room and board. "Its insistence upon limiting the Berea educational opportunity only to those students and families who have economic need and could not otherwise afford a Berea education makes Berea College literally unique in American higher education" (Strategic Planning Committee, 2011, p. 41).

The official admissions policy further delineates Berea's unique commitment to its students. It states:

> Berea College should seek to recruit students mostly from Southern Appalachia, black and white, men and women, (a) who have limited economic resources; (b) whose "great promise" is defined by significant potential for

> academic success and leadership; (c) who will be attracted to Berea's Great Commitments and its clearly articulated emphasis on learning, labor, and service as worthy educational personal goals; and (d) who, along with students from other areas of the U.S. and abroad, will compose a diverse cultural and ethnic mix that will create a 21st-century learning environment. The College seeks to inspire, educate, and graduate service-oriented leaders for Appalachia and beyond. The total student body should number 1,600. (Strategic Planning Committee, 2011, p. vii)

In a strategic action area titled "Admissions and Student Success" (Strategic Planning Committee, 2011, p. vi), Berea further commits itself to:

> …improve its capacity to help the students it seeks to serve by (a) studying the national literature and conducting studies of its particular population of students to better understand the academic, personal, and attitudinal characteristics of Berea students; (b) systematically identifying the diverse strengths and weaknesses that students bring with them to Berea, building on the strengths and addressing the weaknesses; (c) assessing the effectiveness of Berea's current curriculum, teaching, advising, academic support, students services, and residential programs in addressing student preparedness; (d) creating the necessary curricular, academic support, faculty/staff development, and residential/student-life structures and programs to better support students' academic and personal success; and (e) monitoring the progress of this initiative. (Strategic Planning Committee, 2011, p. vii)

Further strategic planning initiatives relevant to this project include a commitment to sustainability in all its forms and reasserting Berea's commitment to interracial education, which states, in part, "Our purpose is not simply to create greater numerical diversity but to engage white and black Bereans more fully in what it means to live together and to learn from and about each other" (Strategic Planning Committee, 2011, p.viii).

According to the Berea College 2013-2014 Fact Book, the faculty to student ratio at Berea College is 11:1. In the fall of 2013, there were 1,581 students seeking degrees at Berea College. Demographic breakdowns of Berea College students in the fall of 2013 include: 57% female and 43% male; 71% from within the defined Appalachian territory, 22% out-of-territory; 8% international; 67% White, 19% Black or African American, 10% international and unknown, 4% other minorities. Of those, 4% consider themselves of Hispanic, Latino or Spanish origin. In 2012-13, 63% of Berea graduates were first-generation college graduates.

Racial demographics of full-time faculty members in the fall of 2013 were: 85% White, 7.5% African American, 3% Asian, 1.5% two or more races, 2.3% chose not to respond. Of those, 4% considered themselves of Hispanic, Latino or Spanish origin (Berea College Office of Institutional Research and Assessment, 2012-13).

According to the U.S. Census Bureau (2014), the city of Berea, Kentucky was estimated to have a population of 14, 374 in 2013. In 2010 the population was reported as 90.7% White, 4% Black or African American, 2.7% Hispanic or Latino, 2.6% mixed race, 1.2% Asian, and .5% American Indian U.S. Census Bureau (2014).

Berea College's curriculum operates within a liberal arts tradition and also offers several professional curriculum programs, including Education Studies. Berea currently offers B.A. and B.S. degrees in 28 fields housed in 26 traditional academic and interdisciplinary programs, as well as a General Education Program. Multiple disciplinary and interdisciplinary commitments have given rise to majors such as Women's Studies, African and African American Studies and Sustainability and Environmental Studies (SENS). Interdisciplinary minors are offered in Appalachian Studies, Peace and Social Justice, and a new Forestry minor has recently been created. The 2013-14 Berea College Fact book states that Berea College "considers the preparation of teachers one of its major areas of focus," (p. 82) and that "many programs at Berea College contribute to the education of teachers" (p. 82).

Currently, Berea students have an option of pursuing both certification and non-certification majors within the Education Studies program. Certification is offered in Elementary Education, Middle Grades Science and Math, and Secondary Certification in a number of content areas.

Intended Audience. The initial audience for my project was to be Berea College administrators and faculty. My intention was to provide an overview of the rapid recent policy changes in K-12 education, with a special emphasis on the impact of new policies as they extend their reach into higher education by way of teacher education programs. That purpose remains, even as this background has expanded to include a deep understanding of the role played by high-stakes uses of Praxis Exam scores in newly implemented state requirements for teacher education programs.

These changes have been difficult to keep up with even as a practitioner in the field. For this reason, I feel a special responsibility to share the information with administrators and professors outside the field of education who have likely not had the opportunity to stay current with the rapid pace of policy change. As a K-12 educator of over 25 years, my research revealed the degree to which my limited understandings about standardized testing and its uses prevented me from making connections I now see as vitally important. I feel a professional obligation to share these insights with my colleagues, especially at Berea College, where the demographics of the student population increases the likelihood that test scores will exclude them from the possibility of becoming teachers. This information is crucial to making thoughtful decisions about what role the College will choose to play, if any, in lobbying for policy change, and also decisions about how the College might best support its Education Studies students.

The audience I held in mind as I was writing, however, grew far beyond the faculty of Berea College. Every person in the United States is a stakeholder in the system of public education. I wanted to write in ways that were understandable to educators and non-educators alike, carefully building a background of understanding that would inform people in ways that might help them make personal choices about possibilities for response. The information contained in this project seems especially important for members of disenfranchised communities and their allies, particularly since current educational policy and the use of standardized test scores in particular has been billed as if it will help, rather than harm, their children's futures.

I also had an audience of educators in mind. Teachers are, for the most part, so busy, and these policy changes are coming at such a rapid-fire pace that, like me, they have likely not had time to do this type of in-depth research. As a researcher and writer, I see part of my role as helping people to connect the dots from history to statistics to psychometrics to policy, perhaps in ways that are new to them. It seems especially important to help arm K-12 educators with this information so their voices can be heard when policy decisions are being made and so they can help inform parents and community members.

Project Implementation and Potential Impact.

Implementation. This project idea has grown over the course of a year and a half from a seed of an idea to a piece of writing that I hope is almost ready for publication. My vision is to identify sections of the manuscript that might be appropriate for a variety of journals, from academic to popular, in order to reach many audiences. I may also send the manuscript to a few publishers to see if there is any potential to expand this manuscript into a book. Because of increasing controversy over the Common Core Standards and especially the use of standardized testing, there might be an audience for a book-length publication. I would also like to try to condense the information into smaller pieces, perhaps summarizing each section, in order to work it into a presentation format. There are a few conferences around the country for which I think this topic might be appropriate.

In the meantime, I will send copies of my project to Robert Brown, Executive Director of Kentucky's Education Professional Standards Board, Dr. Lyle Roeloffs,

President of Berea College, Dr. Chad Berry, Dean of Faculty, Dr. Linda Strong Leeks, Associate Dean of Faculty, Meta Mendel-Reyes, Chair of Division 6, and Dr. Yolanda Carter, Chair of Berea's Education Studies program. Hopefully in this way, it might be passed on to other faculty as well. Perhaps I can even arrange to give a presentation of my findings on Berea's campus for interested faculty and students.

Potential Impact. It's hard to imagine the potential impact this project might have, but I like to dream big. Almost everyone I talk with, both in and out of educational circles, assure me that high-stakes standardized testing is here to stay. I see, however, a ground swell beginning and continuing to grow. Interestingly, opposition is coming from across the political spectrum. I have begun to think of the racialized impact of standardized test scores as part of the "Black Lives Matter" movement. It's humbling to think that the information included in this project might have the potential to strengthen that movement for social and economic justice.

I would also hope that readers in leadership positions, who have the power to establish policy and guide program development, might be guided in their decision-making by the information in this study. What if decisions were made to end the high-stakes use of standardized testing in teacher education, due, even in part, to continued research prompted by questions I raise?

Reflections

I had never before realized the power and the potential of research. This, perhaps, is because I had a limited understanding of what research could be, not knowing the many types and possibilities of research. I now realize that classroom

teachers often conduct research. They might not be aware of their methodologies nor be able to name them, but research is undertaken nevertheless. Results are shared in professional learning communities and teacher work rooms around the country. I didn't know it, but I have actually been researching for many years.

The requirement to engage in a formal, long-term research project intimidated me at first. I wasn't confident I could do anything that "counts" as research. I'd done a lot of writing and knew I had ideas to share, but I wasn't sure I knew how to use the research of others to form the warp through which my own ideas would weave. I now view this requirement as an opportunity. I find I enjoy the work of research as much as I enjoy teaching. My notions of what constitutes educational research have certainly broadened. I would like to find ways to push the edges of that envelope even further. In conducting research for this Capstone, I have come across an idea for a new project, based on a New Zealand research project called Te Kotahitanga. I am fascinated to learn that the project employed Maori research methods as well as European-style research. I would like to know more about what that means. The resulting project has great potential for increasing culturally relevant instruction in schools around the world, and Christine Sleeter (2011) suggests that these methodologies are ready to be applied elsewhere.

The difficulties I experienced in completing this project had more to do with self-doubt than anything else. As I gathered ideas from many disparate sources, it was a great challenge to weave them together in ways that would lead readers gently from one idea to another. I wanted reading this piece to feel like going up a flight of stairs.

I hoped that each paragraph would raise questions in a readers' mind that would be answered in the next paragraph, until we reached a place where the questions became too big for answers. To find out if I accomplished this goal, the piece will require a variety of readers, from inside and outside the field of education. I may find some willing readers, but the piece is so long I may, instead, need to wait until I break the manuscript into articles, which can be more easily read in one sitting

My hope is that the manuscript that follows will flow naturally from the literature review into an in-depth look at standardized testing, each section fitting together in ways that make them feel necessary parts of a larger whole.

Literature Review

In Pursuit of Best Practices in Teacher Education

Linda Darling-Hammond and Ruth Cung Wei, with Christy Marie Johnson (2009), begin their chapter "Teacher Preparation and Teacher Learning: A Changing Policy Landscape" in the *Handbook of Educational Policy Research* by writing, "The last two decades have witnessed a remarkable amount of policy directed at teacher education—and an intense debate about whether and how various approaches to preparing and supporting teachers make a difference" (p. 613). This has certainly been true in Kentucky, where new regulations of the last few years are greatly impacting Teacher Preparation Programs at colleges and universities throughout the Commonwealth.

A few key questions underlying these policy initiatives include points of debate. How can teacher quality be defined and measured? What are the impacts of teacher education on teacher quality? These questions lead to others: What *are* best practices in teacher education? Who ultimately determines an answer to that question and on what basis?

Darling-Hammond et al.'s (2009) chapter briefly summarizes research beginning in the 1960's that suggest many factors seemingly linked to teacher effectiveness, including "general academic and verbal ability; subject matter knowledge; knowledge about teaching and learning; teaching experience;…and traits like adaptability and flexibility" (2009, p. 614). The authors go on to say that relatively little research has been done that would suggest which of these qualities has

the greatest impact, because each characteristic is typically studied independently, even though, in practice, they are intertwined and operate simultaneously.

Darling-Hammond et al. (2009) further cite a 2001 U.S. Department of Education research review of 57 studies published in peer-reviewed journals that, as one would expect, indicate that teacher education does have an impact on teacher quality. In this review, the authors confirm that, by any measure, determining teacher effectiveness is quite complex. It naturally follows, then, that determining how to *assess* those complexities and determining how teacher education programs can *best support* building the skills, knowledge, and dispositions necessary to become effective teachers is exponentially more complex.

Defining core practices in teacher education: complex and ambiguous. Darling-Hammond et al. (2009) cite Darling-Hammond's 2006 study of seven teacher education programs "that graduate extraordinarily well-prepared candidates—as judged by observations of their practice, administrators who hire them, and their own sense of preparedness and self-efficacy as teachers" (p. 618). Darling-Hammond identified several things the programs had in common. Perhaps not surprisingly, they included:

> …a strong, shared vision of good teaching and well-defined standards of practice guiding coursework, clinical placements, and performance assessments; a common core curriculum grounded in substantial knowledge of development and learning in cultural contexts, as well as subjects matter pedagogy, taught in the context of practice, using case methods and other

> pedagogies that connect theory and practice; extended clinical experiences (at least 30 weeks), interwoven with coursework and carefully mentored; and strong partnerships between universities and schools (p. 618).

Although Darling-Hammond goes on to say that few studies focus on specifics about impacts of teacher education practices on teacher effectiveness, perhaps this list of shared traits might be a starting point (1990).

Cherry Collins (2004), of Australia's Deakin University, poses similar questions in the article "Envisaging a New Education Studies Major: What are the Core Educational Knowledges to be Addressed in Pre-Service Teacher Education?" Collins argues that neither of the two primary approaches in the last generation's teacher education practices, "the 'subject matter approach' and the 'social scientific' approach" (p. 228), are sufficient in the 21st century. While the subject matter approach "fails to offer a common expertise to teachers as one profession" (p. 228), the notion that education could be treated solely as a science, with the underlying presumption that "children are likely to respond in the same way to a scientifically verified 'best teaching' approach must now be regarded by thoughtful educators as wishful thinking" (p. 229).

Education Studies Orientations

Traditional approaches to teacher education. In "Teacher Preparation: Structural and Conceptual Alternatives," Sharon Feiman-Nemser (1990) states that what most people consider a traditional teacher education program is a four-year program comprised of two years of general education and two years of professional

courses. For elementary education majors, the sequence of courses generally include a sequence includes an introductory course, educational psychology, separate methods courses for each content area and student teaching. Traditional secondary programs often require adolescent psychology, one general methods course, many courses in their content major, a methods course specific to the student's major and then student teaching. Feiman-Nemser cites Lawrence Cremin's (1978) claim that these models were developed by James Earl Russell and colleagues and were adopted around 1900. Feiman-Nemser says that in Russell's view, teacher education curriculum should include coursework in "general culture, special scholarship, professional knowledge, and technical skill" (p. 11). Feiman-Nemser adds that "[o]rganizationally and conceptually, general education and professional education are separate and distinct" (p. 12).

In a *Phi Delta Kappan* article titled "Traditional Teacher Education Still Matters," Nick Jacobs (2013), a teacher educated at a liberal arts college using a traditional liberal education approach he describes as "often scorned and demeaned" (p. 21), writes of the value he finds in his professional education courses. He argues that beyond just learning skill sets often associated with teaching, teachers need a deep understanding of the theory that backs their practice as a basis for the thousands of quick decisions and adjustments required during the course of a teaching day. Without the opportunity to develop a theory-based teaching philosophy, teachers lack the necessary internalized principals to guide their actions. Jacobs writes that this knowledge of theory coupled with practical classroom experience, especially his

semester of student teaching, gave him the background he needed to be able to stay in the classroom beyond the first year. Clearly, Jacobs' article was written in response to various "fast-track" alternative routes to teacher certification, where theory is sometimes left behind in lieu of quick skill development.

A call for flexibility in teacher preparation. In the same issue of *Phi Delta Kappan*, James Shuls and Gary Ritter (2013) explore the often established dichotomy between college-based teacher education programs and alternative routes to certification. In their minds, establishing this relationship as a dichotomy is unnecessary. They cite the 2013 Measures of Effective Teaching Project (MET), which they define as a rigorous study funded by the Gates Foundation, as further evidence that the process of identifying uniform qualities of teacher effectiveness is anything but a straightforward, simple endeavor. This research corroborates the earlier findings of Darling-Hammond, et al. (2009). Shuls and Ritter agree that stablishing exactly which practices are best for *preparing* teachers is an equally complex endeavor. They call for flexibility in approach, raising the possibility that the "best" approaches for preparing elementary teachers might be different than for secondary teachers, for example, or that the preparation of math teachers might need to be different than for science teachers. The title of their article , "Teacher Preparation: Not an Either-Or," makes this call for flexibility transparent.

As Feiman-Nemser (1990) discusses the difficulty of defining qualities that make teacher education programs "effective," she underscores the fact that program differences have evolved from the philosophies and values of various institutions. She

calls for the use of "conceptual clarity," (p. 39), rather than external pressures, in defining a quality program at a given institution. Feiman-Nemser agrees with Shuls and Ritter (2013) that context matters. Programs differ because there are many more goals than can be focused on at one time, and these goals are intricately interwoven. Therefore, programs have had the flexibility to choose to focus on the philosophies which lie at the heart of their conceptual frameworks. The conceptual frameworks, in turn, are based on the values of the institution in which a teacher education program is housed. Australian professor and researcher Cherry Collins (2004) writes, "…all practice, indeed all structure, is theory-saturated. Research and debate…over the past generation has established…that all human perception and practice is theory-immersed" (p. 237).

Learning to teach involves a process of personal and professional transformation, Feiman-Nemser (1990) claims, requiring reflexive evaluation on the part of each teacher candidate into their own education as well as their personal identities. This kind of transformative reflection is even more essential if candidates are ever going to be able to imagine the new educational paradigms called for by policy makers who claim that educational innovation is needed. Feiman-Nemser cites a general lack of research focused on teacher education. She cautions that no "fully developed framework to guide program development" (p. 39) exists.

Structural and conceptual frameworks for teacher education. Feiman-Nemser (1990) identifies various *structural* models for teacher education, including traditional four-year programs, extended fifth-year programs, Master of Arts in

Teaching (MAT) programs and professional model programs as well as the alternative routes to teacher certification available in many states. Structural models, she writes, "reflect political and economic considerations more than clear thinking about what teachers need to know or how they can be helped to learn that" (p. 1).

Feiman-Nemser (1990) synthesizes her research into five existing *conceptual* orientations for teacher education programs: academic, practical, technological, personal and critical/social, which are philosophically based. Conceptual orientations are not determined by structure. Although they often overlap, each orientation prioritizes different areas. The determination as to which orientation is emphasized has been based on a philosophical stance chosen by each teacher education program. Feiman-Nemser explains that this great variety of orientations exists because of the complexities involved in teaching and learning, and, as "in any complex human endeavor, there are always more goals to strive for than one can achieve at the same time" (p. 38). Therefore, based on their conceptual frameworks and on the missions and values of the institutions they serve, faculty and administrators have had the freedom to establish priorities for their own individual programs.

The academic orientation Feiman-Nemser identifies prioritizes teaching about how people learn and come to understand. In this model, as in the liberal arts tradition, teacher education candidates' content learning often occurs outside the Education Studies Program. While no one would claim that content knowledge is *un*important, Feiman-Nemser cites research that shows that content matter knowledge alone is not enough for teacher education candidates to be effective classroom

teachers. To best impact student learning, content knowledge must be paired with knowledge about how to best *teach* a given content. In other words, even a stellar command of content does not imply that a candidate has developed the skills or dispositions necessary to engage students *with* that content in ways that help them learn. The focus of the academic orientation is that candidates must also learn and practice the skills and dispositions needed to *teach* content within a teacher education program. To illustrate what such a pairing might look like, Feiman-Nemser gives an example of "The Academic Learning Program" model (p. 25), made up of a sequence of core courses paired with field experiences. This model emphasizes conceptual foundations of knowledge, and some Academic Learning Programs also include team-teaching partnerships between a subject area expert and a subject area educator.

A second orientation Feiman-Nemser (1990) identifies is the practical orientation, or the craft-apprenticeship model. This approach prioritizes the "craft, technique, and artistry that skillful practitioners reveal in their work" (p. 26). This model prioritizes classroom immersion as the best way for candidates to learn. The philosophical basis for this orientation is an acknowledgement of the fluid, ambiguous circumstances that are part of any field involving human interactions. While this approach acknowledges that the work of teachers involves complexities that can be learned only in context, Collins' critique is that it often frames teaching as a craft which can be taught and assessed through a checklist of skills. In Collins' view, this attempt to define such complex practice in such relatively simple terms is, by itself, inadequate. The prevalence of the craft-apprenticeship approach in England

has brought about suggestions that university-based knowledge components of teacher education be reduced or even abolished. Collins contends that this model limits possibilities for pre-service teachers to imagine ways of teaching and learning beyond what they see in practice, which, in turn, limits possibilities for educational innovation. The Teachers for Rural Alaska Program is one example of a graduate program that uses the apprenticeship model. A summer planning class is led by master teachers who later act as mentors. Courses of study include close examination and discussion of case studies, using the master teachers' experience as a guide.

The technological orientation is a direct instruction training model, based on the philosophical stance that teaching can be studied as a science. This model is guided by research into specific teaching behaviors that yield specific gains in pre-identified student achievement as measured by standardized test scores. In this orientation, teaching is taught systematically, based on the philosophical stance that if a certain set of plans is followed in sequence, and those plans include effective instructional strategies shown to motivate students, all children will learn. Examples cited include Competency Based Teacher Education (CBTE), which began in the late sixties and early seventies, and PROTEACH, a Master's degree program adopted in 1983 and still in use at the University of Florida in Gainesville. Based on a state-wide assessment instrument called the Florida Performance Measurement System (FPMS), PROTEACH breaks teacher education into six areas, including: instructional planning, student conduct, instructional organization, subject matter presentation, communication and testing.

A fourth conceptual orientation identified by Feiman-Nemser's (1990) research is the personal orientation, which is, in essence, a reflective, student-centered approach. A teacher candidate is viewed as both a teacher and a person, whose uniqueness requires differentiation rather than a standardized model of education. Developmental and inquiry-based approaches are highlights of this orientation, which emphasizes that teacher candidates should develop their own teaching styles through reflection on their practice as they work closely with teaching mentors. Examples of the personalized approach are the Personalized Teacher Education Program (PET) at the University of Texas and the advisement program at Bank Street College.

Lastly, the critical social orientation arises from the philosophical standpoint that education is a vehicle for working toward social equity and a more democratic nation. In this approach, Education Studies professors model systems based on democratic values, and focus on issues of educational policy as well as pedagogy. Feiman-Nemser (1990) writes that research shows this approach to often be mostly theory-based, rarely helping teachers navigate the realities of every day teaching practice in schools. Courses and field experiences are designed to "promote critical analysis and critical pedagogy" (p. 36). Examples of such models include New College experiment in the 1930's, a part of Teachers College at Columbia University and the student teaching component at the University of Wisconsin at Madison.

Considering complexities. Collins (2004) suggests a new organizing principle might supersede the limitations of each of the orientations Feiman-Nemser (1990) outlines. Collins suggests introducing multiple lenses through which teacher

candidates are asked to view their growing knowledge and practice. This would allow them to expand their "conceptual repertoire" (p. 232) to include the kind of unconventional, complex thinking Collins suggests is needed in teacher education. These lenses through which teachers can learn to view their practice include:

> ethical lenses, research-based lenses, cultural lenses, rich theories of human development, insights into diversity and inclusivity, ways of thinking about children and adolescents, experience in a wide variety of learning sites, knowledge of the variety of human learning practices and debates about them, and so on (p. 232).

The number and scope of the lenses Collins suggests again sheds light on the complexities involved in attempts to define all the important things teachers should know and be able to do.

These complexities are further highlighted in a 1993 report by the National Center for Research on Teacher Education, based on research conducted between 1986 and 1990. In "An Annotated Bibliography: Findings on Learning to Teach," (1993) the studies were summarized in the form of an annotated bibliography, organized around six *myths* about teacher education revealed through this body of research.

> Myth #1: Majoring in an academic subject satisfies the requirement for subject matter knowledge needed for teaching.
>
> Myth #2: Giving teachers information about the cultures of various groups enables teachers to teach children from these groups.

> Myth # 3: Mentor teacher programs encourage thoughtful teaching among novice teachers.
>
> Myth #4: We can produce good teachers if we start with people who are smart and who have subject matter degrees, and then give them classroom management survival skills.
>
> Myth #5: Different program structures in teacher education will lead to different knowledge and skills in teachers.
>
> Myth #6: Short-term in-service workshops are an effective device to improve teaching practice.
>
> (pp. 1-4)

Clearly, any attempt to assess the complexities of teaching, and to evaluate the teacher education programs that teach to these complexities, will require a flexible and complex assessment process.

A case study: Re-conceptualizing a teacher education program. Cherry Collins' article about the process undertaken by Deakin University in Victoria, Australia to re-conceptualize and redesign their teacher certification program serves as a useful example of establishing a philosophical base around which a program can be built. The faculty began by gathering information from many teacher education institutions. They studied current education policy and attempted to look into future policy initiatives. They explored various perspectives about what was "most" important in the research literature, nationally and internationally. "The point of all of

this," Collins writes, "was to provide a basis for discussion of what the Education Major could become" (p. 233).

Through this process, the faculty identified five qualities to use as the basis for six semester long core courses. These qualities reflected a shared philosophy similar to that defined in a program's conceptual framework. They decided graduates should:

> 1.) Be inclusive in their teaching practice 2.) Be aware of students as active meaning makers 3.) Be committed to teaching for deep understanding and clear thinking 4.) Be skilled at quality professional relationships and 5.) Be committed to life-long learning as reflective, professional practitioners
>
> (Collins, 2004, p. 234)

This set of shared beliefs then acted as a guide as they re-imagined their program from the ground up.

Making the change. Based on these common values, the faculty decided to create a series of carefully designed questions around which to base their six semester-long inquiry-based units, rather than following the traditional model of thinking about each course in terms of academic disciplines. The questions framing the units are:

> Unit 1: What are some of the most useful ways in which teachers might think about children and adolescents as persons?
>
> Unit 2: What do we know about how children and adolescents learn?
>
> Unit 3: If you now have some ways of thinking about who children and adolescents are and of the variety of ways they can learn and the factors

> involved, what then might be effective ways to support learning and to create good learning environments?
>
> Unit 4: What is taught in schools and what should schools be teaching?
>
> Unit 5: …explores the idea of teaching as a profession requiring expertise in human relationships and teaches the knowledges which such expertise requires.
>
> Unit 6: …is called *Transition to beginning teaching*. It is double-stranded. One pragmatic strand is to ensure that graduates are familiar with the institutional and professional responsibilities and administrative tasks required of teachers in schools. A second strand consists of a research and reflection project of some value to a school…built around an issue of inclusivity in teaching practice.
>
> (pp 235-236)

This is an example of a process undertaken by one teacher education program to define their shared philosophy and then open their thinking to new ways in which this philosophy might be enacted. At Deakin University, faculty considered both the structural model and the conceptual orientation of their program. The analysis of their experience might act as a guide for teacher education programs around the world who are searching for ways to re-define and re-structure their offerings.

Capstone Project

Teacher Education in the United States: A Changing Policy Climate

A move to standardization. Based on a history of academic freedom that has long characterized American education at the university-level, teacher education programs were asked to conceive and write individual conceptual frameworks based on contextualized, institutional values which guided their practice. At the K-12 level, however, U.S. education policy is on a course of standardization, as evidenced by the adoption of Common Core standards in most states and in the high-stakes use of standardized assessments for accountability purposes. Based on a close read of the new standards adopted in 2013 by the Council for Accreditation of Educator Preparation (CAEP), this emphasis on standardization is trickling up to college and university level teacher preparation programs. This marks a sea-change in U.S. teacher education.

Accrediting bodies have required each teacher education program to define its individual philosophical stance in a written conceptual framework used to guide program development and review (NCATE, 2010-2013). Until quite recently, there have been two accrediting bodies in the U.S. In 2013, those two accreditation agencies merged to form CAEP. In the name of program accountability, CAEP has developed a new set of standards which will have a uniform, national impact on teacher preparation programs. Enforced through accreditation evaluations, this set of standards will now guide standardized, systematic program evaluation of all teacher education programs, regardless of context or philosophy.

Standardizing a philosophical stance. In pursuit of quality and accountability, the standards developed by CAEP seek to align teacher education programs nationwide. As detailed previously, research has shown teaching to be so complex and multi-faceted that choices must be made. The CAEP standards, too, are based on a selected philosophical stance, which emphasize certain philosophical orientations over others. Because these chosen philosophies are sometimes presented and accepted as universal truths as policy is rolled out, it is important to be able to identify and name these stances and their underlying assumptions, so institutions can consider their alignment with their own missions and values.

One philosophical stance clearly favored in current educational policy is what Feiman-Nemser (1990) identifies as a technological stance, in which education is regarded as a science, guided by research conducted to identify specific teaching behaviors that yield gains in pre-identified student achievement as measured by standardized test scores. Authors and educators Paul Gorski (2013) and Christine Sleeter (2014) identify this stance as part of a larger neoliberal framework. As Feiman-Nemser (1990) points out, in each decision made, some types of knowledge and skills are privileged over others, and each is based on a series of assumptions. The nature of assumptions is that they often become such a part of the landscape they are taken for granted. People sometimes forget that alternatives continue to exist.

As proposals affecting teacher education move forward at a break-neck pace, rapidly becoming codified and enforced by CAEP, the sole accrediting body in the U.S., it seems we are at a critical juncture for naming and examining these

assumptions. Each institution will then need to study the policies resulting from these assumptions and the impact of the resulting policies on teacher education programs and the students they serve. Ultimately, institutions will need to analyze the underlying philosophies alongside their statements of mission and vision, as well as their strategic plans. If this analysis identifies conflicting philosophies, institutions may find themselves at a crossroads.

Neoliberal philosophy. Several author/educators are writing as fast as they can in order to help institutions more clearly see these philosophical stances and their underlying assumptions. In the introduction to *Power, Teaching and Teacher Education*, author Christine Sleeter (2013) summarizes neoliberal ideology and how it came to form the basis of the new U.S. education agenda. This is important to understand in any exploration to try to find the roots of the new CAEP standards. Neoliberal philosophy, Sleeter explains, has the core principals of "individual liberty, private property and market competition" (p. 3). In *Reaching and Teaching Students in Poverty* (2013), Paul Gorski describes neoliberalism as an ideology that applies free market strategies and corporate style reforms to all areas of life, including community resources like public schools. Words like standardization, accountability, and data-driven decision-making have clear ties to corporate America. As any teacher in today's classrooms knows, these words have quickly become part of everyday language in educational circles. Gorski says that the pervasiveness of this mindset is demonstrated when even educators begin citing test scores as *the* definition of student success.

Sleeter also claims that neoliberal practice has paved the way for corporations to expand globally in many areas. By way of example, she names groups like the American Legislative Exchange Commission (ALEC), which she says now influences everything from tax codes and federal policy, increasing its power exponentially. To illustrate this reach into the educational realm, New Orleans community activist Karran Harper Royal says that the New Orleans Public Schools are "operating the ALEC agenda lock, stock and barrel" (Karp, S. & Sokolower, J., 2014).

Sleeter (2013) describes a large philosophical divide between education for democracy and what she calls education for corporatocracy, which she identifies as the philosophical base of the self-proclaimed reform movement. This divide is based on philosophical differences in assumptions about *what* is important for students to learn, *how* students learn, and *why* student learning is important (Sleeter, 2013). As Feiman-Nemser (1990) points out, teacher education is so complex that every program has previously had the freedom to make choices in setting its own priorities. In the new educational policy, Sleeter supplies strong evidence that multi-culturalism, an emphasis on valuing and teaching the multiple perspectives that exist among American citizens, a value essential to democratic thought and practice, is being de-emphasized in lieu of a sharp focus on test-driven accountability. Rather than education for democracy, Sleeter writes, "Neoliberalism has framed schools like businesses designed to turn out workers for the new global economy, and as venues for profiteering…" (p. 146).

Sleeter (2013) identifies three ways she sees neoliberalism influencing the track of teacher education, as well. Teacher education, she observes, is being pushed "(1) away from social justice teacher preparation and toward preparing teachers as technicians to raise student test scores; (2) away from being linked to teacher professional knowledge and teacher quality; and (3) toward becoming shorter or by-passed altogether" (p. 146).

A focus on technical/scientific philosophical orientation. Feiman-Nemser (1990) identifies assumptions that underlie the technical/scientific philosophical orientation. This perspective calls for research to guide the identification of effective teaching behaviors, so the teaching of those behaviors can be standardized. The goal is to find the right combination of strategies that lead to specific gains in student achievement, as measured by standardized test scores. In this orientation, teaching is straightforward and systematic. Little consideration is given to context of any kind, including classroom relationships. The view is that if teachers follow a certain set of research based plans, students will learn. According to James Popham (2014), this philosophical standpoint led to the model known as "programmed instruction" (p. 62), based on the theories of B.F. Skinner. Popham goes on to explain that this philosophy laid the foundation from which criterion-referenced testing grew. Criterion-referenced tests are designed to measure student learning against pre-identified objectives, such as the Common Core standards, based on the assumption that the tests used will accurately measure student learning. What follows, then, is that a clear set of teaching behaviors and curriculum materials, identified through

research and followed with fidelity, would lead every student to attain learning to that standard.

A question arises, however, about what research methodologies might be used to determine this combination of materials and behaviors. In fact, Sleeter (2013) points out ways in which neoliberalism has attempted to redefine research to "manage dissent while building consensus for its expansion" (p. 151). This redefinition has narrowed what "counts" as educational research to quasi-experimental quantitative models which use standardized test results as the sole measure of academic success (Sleeter, 2013). This philosophical stance simply discounts the multiple types of research which have always existed and those that continue to evolve. Researchers didn't make this change; policy makers did. Somehow, the newly named federal Institute for Education Science has been allowed to determine what counts as research *for* the academic world. This stance narrows the kinds of research questions that can be asked and certainly influences the "answers" that are obtained. "The only question left on the table," writes Sleeter, is "[w]hat teaching strategies have been found to raise student test scores, using experimental or quasi-experimental research?" (p. 152).

This narrow re-definition of what research *is,* along with the unstated assumption that research can determine *the way* new knowledge is gained and how it can be measured, has been carried right into classrooms and teacher education programs, sometimes without question. Sleeter writes, "raising standards has become synonymous with standardizing curriculum" (p. 28). It is crucial to understand that

this stance leaves many other ways of knowing, including those of indigenous peoples and other marginalized groups, out of the equation completely. Further, teachers and teacher educators who question this framing of a single stance are viewed as impediments to progress in this new educational era. In a brilliant TED talk, author Chimamanda Adichie (2009) reminds us of the dangers of a single story. This privileging of a single way of knowing should raise many questions about who and what is being left behind as the gates narrow to allow only one story line to move forward. This, after all, is what standardization means. It raises important questions about the relationship between standardization and assimilation.

Education for Democracy: A Philosophy Left Behind?

The Common Core standards do not in themselves call for any kind of scripted curricula. They also do not preclude the use of culturally relevant teaching and materials that include multicultural perspectives. They are, however, based on specific ideas about the purposes and goals of U.S. public education as determined by the groups who created and continue to promote the standards, the Chief Council of State School Officers (CCSSO) and the National Governors Association (NGA) in partnership with Achieve, ACT and the College Board (National Alliance of Black School Educators, 2014). Achieve is an organization founded at the 1996 National Education Summit by "leading governors and business leaders" (Achieve, Inc. 2014). ACT and the College Board are both companies that create standardized tests. Interestingly, the CCSSO website itself, which used to identify the involvement of each of these entities in the creation of the standards, now lists only the NGA Center

and the CCSSO (Berry, 2014). The standards created by these organizations, called the Common Core Standards, call for students to leave high school "college and career ready" (Achieve, 2014). As Sleeter (2013) points out, this is based on a philosophical view that the purpose of education is to prepare workers and socialize them in ways that meet the needs of employers.

Juxtapose this to the purpose of education as defined by retired Supreme Court Justice Sandra Day O'Conner in the introduction to Sam Chaltain's (2010) book, *American Schools: The Art of Creating a Democratic Learning Community.* O'Conner writes, "…the primary purpose of public schools in America has been to help produce citizens who have the knowledge, the skills and the values needed to sustain our centuries-old experiment in democracy" (pp. xii-xiii).

This American democracy is made up of a diversity of citizens, all of whom need educational opportunities that allow them to find the power of their own voices in order to exercise their rights and their responsibilities, if this democratic experiment, as Justice O'Conner refers to it, is to succeed. The choice made by Partnership for Assessment of Readiness for College and Careers (PARCC) and the CCSSO to emphasize college and career readiness alone seem based on a different set of assumptions about the purpose of education than those stated by Justice O'Conner. Indeed, rather than a community-based outlook that considers the social good of the whole, a focus on college and career readiness seems closer to the assumptions Sleeter (2013) identifies as underlying corporatocracy, where "…education is a resource for national global competition and for private gain…and it prepares

workers and socializes them to connect their own self-interest and future with those of their employees" (pp. 18-19).

A missing "C"? This emphasis on college and career readiness, then, further demonstrates how policy decisions are based on particular philosophical choices. These choices should raise questions about the assumptions upon which current educational policy is based. Might the addition of a third goal, calling for students to also be "community ready" enlarge the purpose of education to extend outside the workplace? What impact will this omission have on the kinds of learning that occur within the schoolhouse walls? Indeed, what impact might it have on the American experiment in democracy, an experiment that has not yet lived up to its ideals? (Starnes, 2006).

A white norm in American society. To examine the possible impacts of a missing "C" for community in the college and career ready outcomes prioritized by the creators of the Common Core Standards requires a brief, racialized, historic overview. In *Racing to Justice: Transforming Our Conceptions of Self and Other to Build an Inclusive Society,* author john powell (2012) asks readers to question the "truth" of the key values Christine Sleeter (2013) identified as those of neoliberalism. These values of our larger society have gone almost unquestioned by European Americans since the time of colonization. Powell says the conception of the "Western self" (xviii) is constructed on what he calls the "Enlightenment ideals" (xxiv) of "radical individualism: rationality, objectivity, private property, market capitalism, and race" (p. xviii). Powell writes, "This notion of self is at the core of the American

dream of liberty and opportunity for all, pure meritocracy, but also of exclusion and domination" (p.xviii.)

It is only in naming these assumptions, as powell does, that they can be examined. Powell places this notion of the supremacy of the individual in a historic and racial context and shows how the assumed supremacy of this view has been used to marginalize non-white groups, many of whom organized their societies in ways that placed greater value on the collective good, rather than on the individual. Powell refers to a white norm as a way of naming this central set of assumptions that still guide our government, legal systems and public institutions. The establishment of this norm created clear boundaries of what has been determined to be "civilized" and "uncivilized," terms used in multiple ways throughout history to raise some to power and subjugate others. In order to avoid questioning the inequitable impacts of policies based on these assumptions, powell writes, citing several sources within this quote, "Other complementary ideologies have been employed as needed to provide scientific (for example eugenic and polygenic effects), and, more recently, cultural (as in 'the culture of poverty') explanations for the inequalities of Western society" (p. 169).

A white norm in American education. When examined through a racialized lens, it is clear that the white norm is the pervasive ideology in schools, as well. Powell (2012) cites the use of an ideology of false neutrality in order to maintain the status quo. When something is considered "normal," to people for whom that something is, indeed, a cultural norm, it is sometimes difficult to recognize that the notion of normal is dependent on a context. Rather than acknowledge this context,

however (i.e. ask the question: *normal for whom*?), sometimes people regard "normal" as a neutral, or objective stance.

In addition to recognizing and naming what are often assumed to be societal norms, powell emphasizes the need to expose the existing power relationships on which these norms are based. Left unexamined, polices are implemented and decisions are made in ways that result in what authors David Barnett, Carol Christian, Richard Hughes and Rocky Wallace (2010) identify as "privileged thinking" in schools. Bobby Starnes (2006), in an article titled "Montana's Indian Education for All: Toward an Education Worthy of American Ideals," outlines the many negative outcomes, for students of color and for white students, that result from these systems of privileged thinking.

Alternative value systems do, in fact, exist, especially in a country with a population as diverse as the United States. Rather than defaulting to a single norm, determined by those who hold the power, powell (2012) calls for "an alternative vision, a beloved community where being connected to the other is seen as the foundation of a healthy self, not its destruction…" (p. xix). This possibility of interconnectedness and beloved community, powell continues, must be "reflected in social structures and institutions" (powell, p. xix)…institutions like schools. In agreement with Starnes (2006) and Sleeter (2013), Carjuzza, Jetty, Munson and Veltkamp (2010) point out in "Montana's Indian Education for All: Applying Multicultural Education Theory," it is in the scholarship around multicultural education where many of these alternative visions can be found. This, however, is the

very research and pedagogical methodology Sleeter finds pushed to the margins in the neoliberal stance of current education reform policy.

Standardization of curricula. As mentioned previously, standardization of curricula has been one result of educational policies and research based on the technical/scientific ideological stance. Currently, this policy relies on standardized test scores as the sole measure of academic impact. Standardization necessitates removing learning from any kind of local context. Rather than a localized idea of what worldviews are "normal," standardized curricula is based on curricular developers' ideas of a nationalized norm that john powell (2013) identifies as a white norm.

Based on this technical/scientific orientation and its claim that effective teaching and curricula can be standardized, scripted programs have been developed and are still being rolled out in schools around the country, especially in schools determined (according to standardized testing data) to be low-achieving. When such a program is put into place, teachers are instructed to follow the program with fidelity. This means that teachers are not permitted to vary the program according to the context of geographic location, community, cultural context or student need. As one example, in my experience I have seen the same scripted reading program mandated for use in schools on the Blackfeet Nation in Montana and in a school district in the heart of Appalachian Kentucky. In those districts where this program has been adopted, people are hired to move between classrooms with checklists, making sure

that teachers are on the "right" page of the program on the "right" day, in order to keep the learning uniform and on the programmed schedule.

Elizabeth Dutro (2009) writes, "…the language of curriculum…constructs a particular view of the world and speaks from a particular perspective that necessarily values some perspectives and knowledge more than others" (p. 90). Dutro's research goes on to demonstrate ways in which curricula designed for nationwide distribution "necessarily operate from assumptions about students and what they do, can, and should know…regardless of race, class, gender, or region" (p. 91) and how this negatively impacts students who don't meet such presumptions of what is "normal." James Hoffman (2000) agrees with Dutro when he writes:

> Schools are institutions that serve multiple functions but a singular goal: to prepare the young to assume a contributing place in society. Schools are not neutral in their stance toward the nature of that society…They enculturate the young toward the values, beliefs, skills, and understandings that will preserve existing structures. But schools can also, under the best of circumstances, challenge us to examine our own society, reflect on its strengths and weaknesses, and set our sights on improvements. This is what a democracy demands if it is to thrive, not just survive (p. 616).

After all, education for democracy would ask teachers and students to learn to identify and name the very norms that keep systems of privileged thinking in place. This kind of critical literacy requires critical thinking of the highest order, just as the Common Core Standards call for.

Not surprisingly, research on the positive impacts of ethnic studies programs demonstrate that standardization based on a nationalized norm leaves many students behind. Sleeter (2013) writes,

> Ironically, when we use student-centered rather than textbook-centered teaching, embed preparation for college in rich thematic units that have meaning to one's own students, and engage students in critically questioning society and learning to act for justice, then students from communities that had not been achieving well in school blossom in ways that show up even on standardized tests. Doesn't this make more sense than the current approach that consists of marching everyone lock-step through the same pre-packaged curriculum materials? (p. 75)

Multicultural education matters. Because Culturally Relevant Teaching (CRT) and multicultural education have never played a prominent role in U.S. educational policy, these practices are only carried out by some individual teachers in their own classrooms by their own choice. This, of course, makes extensive research on the impact of these practices difficult. Christine Sleeter (2011) traveled all the way to New Zealand (NZ) to study the lasting impact of CRT in the context of a research study called Te Kotahitanga. At the time Sleeter wrote *Professional Development for Culturally Responsive and Relationship-Based Pedagogy*, the study of the impact of teacher professional development in CRT on student achievement was in its sixth year. Sleeter found that in New Zealand, researchers are documenting that changes in teacher attitudes and classroom behaviors are resulting in increased achievement

outcomes for Maori students (indigenous people of NZ) and Pakeha (non-Maori) students alike. Significantly, the research methodology was grounded in a mix of Maori-defined research methods as well as traditional European-based methods. Rather than relying on data gleaned from standardized testing, outcomes are triangulated from multiple sources of data. The research results are based solely on changes that came about as a result of teacher professional development in CRT. No curricular changes were made to better represent Maori perspectives in the curriculum, nor were recruiting efforts in place for greater representation of Maori teachers during the time of the study. In *Power, Teaching and Teacher Education*, Sleeter (2013) wonders what the compounded impacts of such accompanying changes might be.

Sleeter's (2013) review of similar kinds of research in the U.S. focuses on scholarship in the area of Ethnic Studies, which, like research on CRT, is scarce. Sleeter was able to find a substantial body of research on the outcomes of individual programs. She cites research which documents the positive impact of such programs. These studies highlight the frustration of students of color with the Euro-centric curricula in most U.S. schools, which leads to their disengagement with the subject matter. Ethnic studies programs intentionally add perspectives and knowledge of other racial and ethnic groups to curricula. In fact, the term "ethnic studies" itself makes visible the assumption of a white norm. "Ethnic" in this case, represents "not white." Clearly, the "ethnic studies" curricula are different than the "normal" curricula, in ways that make a white norm visible.

In the body of research Sleeter (2013) reviews, all but one study shows positive achievement outcomes for students who are members of the ethnic group that was the focus of the study. One specific example Sleeter gives is the Mexican-American Raza Studies program, founded in Tucson, Arizona in 1998. Sleeter cites reports that clearly demonstrat the program's great success for the Mexican-American students the district serves. Despite this documentation of increased academic achievement, in a 2012 ruling aimed specifically at this program, the Tucson school district banned ethnic studies entirely, along with specific books used in the Raza Studies program (Bigelow, 2012). This, of course, raises a question: Is European curriculum *not* an ethnic study? This very question reveals the assumption of a Euro-centric norm.

This example helps clarify Christine Sleeter's (2013) point when she writes, "White adults generally do not recognize the extent to which traditional mainstream curricula marginalizes perspectives of communities of color and teaches students of color to distrust or not take school knowledge seriously," (p. 80) For this reason, she further cautions against adding even a well-designed ethnic studies curriculum without accompanying professional development for teachers. When ethnic studies curricula are thoughtfully and carefully carried out, however, research finds its academic and social benefits are clear, for students of color as well as for white students.

Schools and the overwhelming presence of whiteness. The demographic divide between students and their teachers is growing ever wider, making it even

more important to identify the assumptions of a white norm underlying education policy, and all the more important to carefully consider the *impact* of any given policy. Sleeter (2013) finds that, compared to rapidly increasing diversification of the student population in terms of race and ethnicity, there is an ongoing lack of diversity in the teaching force. In *Profile of Teachers in the U.S. 2011*, Feistritzer reports that 84% of American teachers are white, while between 1980 and 2008, the racial and ethnic makeup of U.S. citizens has gone from 80% white to 66% white (National Center for Education Statistics, 2010).

The title alone of Sleeter's 2001 review of data-based research, "Preparing Teachers for Culturally Diverse Schools: Research and the Overwhelming Presence of Whiteness" (p. 94) says much about the impact of the divide between the rapidly changing demographics of students in the United States and the relatively stable demographics of teachers and administrators. The significance of this gap is highlighted in the research focused on culturally relevant pedagogy, which shows a relationship between the cultural gap and the achievement gap (Ladson-Billings; 1995, Sleeter, 2001). As a result, Sleeter (2013) defines continued efforts to diversify teacher education "a demographic urgency" (p. 175).

Teacher diversity matters. A report by the Center for American Progress (Ahmad, F.Z. & Boser, U., 2014) makes recommendations for ways to diversity the teacher workforce based on extensive research highlighting the positive effects for students when the cultural or ethnic background of their teacher is similar to their own. As it now stands, most white students continue to enjoy this advantage, because

over 80% of the teacher population is white (Ahmad, F.Z. & Boser, U., 2014). This report cites a 2010 literature review that delineates practices demonstrated by teachers of color that benefit their students. They include: "having high expectations of students of color; providing culturally relevant teaching; developing trusting relationships with students; confronting issues of racism through teaching; and serving as advocates and cultural brokers" (p. 6). Sleeter (2013) cites similar studies, and emphasizes the importance of student access to the multiple worldviews present when there is a diverse group of teachers at a school. As American society and schools become ever more segregated, as Kozol (2005) shows they have in recent years, intentional diversification of the teachers within a school would give all students the opportunity to learn from both people who "look like them" and people who don't. Both are critical, for students of all ethnicities, if we are ever to reach the level of interconnectedness john powell (2013) suggests we strive for.

The "achievement gap" and deficit thinking. A key tenant of the "Effective Teacher Profile" (Sleeter, 2011, p. 40), developed as a part of the Te Kotahitanga research project in NZ, is that effective teachers "positively and vehemently reject deficit thinking as a means of explaining Maori students' educational achievement levels (and professional development programs need to ensure that this happens)" (p. 40). In fact, many of the teaching practices of teachers of color cited in the report of the Center for American Progress (2014) are listed as part of the NZ Effective Teacher Profile, which suggests possibilities for research into the broader applicability of this work. Citing NZ researchers Shields, Bishop and Mazawi (2005),

Sleeter (2013) points out that sometimes teachers blame academic underachievement on their students. Sleeter, Gorski (2013) and others call this deficit thinking. All too often, deficit thinking leads to practices that perpetuate negative outcomes for students.

Paul Gorski (2013) has written extensively about the harm deficit ideologies can manifest. In *Reaching and Teaching Students in Poverty*, Gorski gives many examples to show that the deficit view of students dominates current U.S. educational policy. In my own experience, it has become increasingly common to hear teachers and policy makers to refer to "gap children," referring to the so-called achievement gap. Framed in this way, this group of children poses a "problem" to be "solved." This is a prime example of deficit thinking. Without questioning assumptions about what the perceived "gap" is, how it has been defined, and what its alleged causes are, teachers, and now teacher education programs, are frantically being asked to find ways to "close the gap." Gorski makes a strong case that the "achievement gap" is actually an "opportunity gap" (p. 83), caused by lack of access to goods and services available to some and not to others. In addition, he claims this deficit view "misdirects a lot of well-intentioned efforts to create equitable schools by leading us to believe we can solve the problem of unequal educational outcomes by 'fixing'…people rather than the conditions that are unfair" (p. 109). Indeed, there are legitimate questions about whether this perceived gap may actually be *created* by the very standardized testing that claims to demonstrate its reality.

Since the time of U.S. colonization, deficit thinking has been so common it may be hard for some to recognize. Historical and contemporary references to a group of people as if they are the source of some social problem are an indicator of deficit thinking. References to the "Indian problem," for example, are pervasive throughout U.S. federal policy documents (T. Lyman, 1973). Sleeter (2011) points out that deficit thinking sometimes leads to expectations that people assimilate to the dominant viewpoint. By way of example, Sleeter cites Ruby Payne's' expectation that people should want to rise out of a so called "culture" of poverty… defined and named as a culture by Payne herself.

An alternative to deficit thinking in schools is for teachers to come to learn about the lives, worldviews and cultures of the students they teach and of American citizenry in general, in all its diversity. It is important that the definition we imply with the use of the term "we Americans" includes everyone. A starting point is coming to understand the Euro-centricity of U.S. educational curriculum. Working to operate from other perspectives, in addition to what john powell (2013) calls the white norm, is another step. Bobby Starnes' (2006) article "What We Don't Know *Can* Hurt Them: White Teacher, Indian Children" outlines the story of a seasoned teacher who finds herself in a new context and learns the importance of incorporating the Chippewa-Cree perspectives and worldviews of her students and their families into her teaching. Sleeter (2013) gives examples of the many studies that demonstrate how important it is that teachers "learn to teach students whose culture and language differ from their own" (p. 215).

Clearly then, with an overwhelmingly white teaching force, there is a pressing need for professional development work in this area. Sleeter (2013), however, found little research taking place that would set a direction for this professional development. So little, in fact, that she had to turn to New Zealand to find results from a long-term study. This magnifies Sleeter's contention that current educational policy is pushing multicultural education, which is necessary for the achievement of *all* students, even further into the margins. The costs for children who find themselves caught in the so-called "gap" are high.

Ironically, the children caught in this "gap" are the very children new educational policy claims to benefit. The argument is that by using standardized test scores to point out this gap, researchers, teachers and schools will find ways to "fix" it. This so-called "fix" has been underway since the implementation of No Child Left Behind, yet according to these very test scores, little has changed. Close examination of the *impact* of these policy changes, whether it is intended or unintended, may offer explanations.

Intent vs. Impact: Questioning Assumptions behind Policy Shifts in Teacher Education

In Kentucky, one of the front-runners in the K-12 education reform/accountability movement, policy makers are now focused on aligning educational standards and policy in ways that impact higher education as well. This move seems only logical. Premised in a technical/scientific philosophical stance, where research-based instructional strategies are thought to lead to predictable,

positive outcomes, it stands to reason that once these instructional strategies are in place, the entire K-16 educational system could be aligned. This seems a simple solution. The complication lies in the notion that research has pointed the way to the "right" instructional strategies, under the assumption that such a universal conclusion is possible. As Sleeter (2013) points out, what "counts" as research in education is now narrowly defined by the neoliberal stance taken by policy-makers. Thus policy-makers, not researchers, have defined the parameters for the research that defines which instructional strategies are called "research-based" and which are not (Sleeter, 2013). In my experience, educators and administrators who are not researchers and who have little experience reading research, often take this "research-based" branding at face value.

This is an important moment in the history of American education. The reach of these new policies into higher education, and teacher education in particular, will complete a self-perpetuating cycle that will lock the entire system into place. Therefore, in this moment, it seems crucial to not only identify and question the assumptions upon which policy decisions are based, but to also closely examine the known and potential *impact* of these policies on the pool of teacher education candidates. By impacting the candidate pool, these policies, in turn, will impact the students who are America's next generation of citizens, and therefore, arguably, the future course of our nation.

Policy shifts in teacher education: Overview and intent. Recent changes in current teacher education policy stem from the philosophies and assumptions that

guided the creation of the No Child Left Behind Act and later Race to the Top. In each of these movements, accountability-over-all has been a top policy priority in K-12 education, and policy makers determined exactly how that accountability would be defined. This marks a sea-change in U.S. education, where national standards and national systems of accountability have never before existed. Policy makers view this lack to be part of the problem.

The solution, in the view of supporters of these policies, has led to the creation of a system of national accountability based on common standards and common measures to assess whether or not the standards are being met. The intent behind these policies, as stated in a 2012 report by the Chief Council of State School Officers (CCSSO) report "Our Responsibility, Our Promise," is to set a new, high bar for students across the country and then hold teacher education programs accountable for preparing teachers to teach in ways that help students reach those standards.

This all sounds simple and straightforward. It is important, however, to ask some key questions:

*How did the developers of the Common Core standards, (the Chief Council of State School Officers (CCSSO) and the National Governors Association (NGA), Achieve, ACT and the College Board), *define* academic success for individual students, for schools and for school districts? How did they decide such success should be *measured* in each of those cases?

*Once that process was completed, how did the Council for the Accreditation of Educator Preparation (CAEP), the newly created sole teacher preparation

accrediting body, along with state agencies responsible for monitoring and assessing teacher preparation programs, *define* the levels of teacher quality they deemed necessary to help students reach those standards of success? How do they plan to assess those qualities?

*And now that this process is moving into higher education in ways that impact teacher preparation programs, how are CAEP and state policy makers *defining* how teacher preparation programs can best support pre-service teachers in meeting their definition of quality? How do they plan to *assess* this program quality?

Answers to these extremely complex questions have apparently been reached and are now established as policy at state and national levels. Clearly, these answers are necessarily based on chosen philosophical stances about the purposes of education in the U.S. The creators of the Common Core Standards have determined that the purpose of schooling is that students should be college and career ready (Common Core Standards Initiative, 2014). The standards were written to that end, beginning with a determination about what college and career ready should look like and working the standards backward all the way to kindergarten (Common Core Standards Initiative, 2014). As the standards have been introduced in classrooms across the nation, questions have been raised by longtime educators about whether or how the large body of research on child development was taken into consideration in the creation of the standards. (Strauss, January 29, 2013).

Once standards to define academic benchmarks were in place, the next logical question was how student learning would be assessed relative to these new

benchmark standards. This required the creation of standardized criterion-referenced tests. By 2014, two nationally normed standardized tests had been created and field-tested by two different consortiums. In many states, scores from these tests will be used not only to measure and compare the academic success of students, but also to evaluate teachers, schools and districts (Layton, L., 2014). As states have announced their plans for this use of the tests, it is increasingly being called into question (Layton, L., 2014).

As Harvard professor and testing expert Daniel Koretz (2008) warns, creating such seemingly simple solutions to such a complex task as measuring the academic success of individuals, let alone the "success" or "failure" of teachers and entire schools, and then as the next planned step for evaluating teacher education programs in Kentucky (Walters-Parker, K, 2014), is a dangerous practice. In *Measuring Up: What Educational Testing Really Tells Us*, Koretz writes, "…test scores usually do not provide a direct and complete measure of educational achievement" (p. 9). Further, Kortez says, "these tests can only measure a small subset of the goals of education…[and] even in assessing the goals that can be measured well, tests are generally very small samples of behavior that we use to make *estimates* of students' mastery of very large domains of knowledge and skill" (p. 9, emphasis added). Testing experts, known as psychometricians, know this. Apparently, in their search for straightforward answers, policy makers, using their chosen paradigm of treating education as a science, are choosing to ignore the clear limitations of what can and can't be assumed based on test score data. As a result, policy based on the use of

standardized testing results as the basis of accountability is now moving into the realm of teacher education as well (Walters-Parker, 2014).

A close look at policy development. The 2012 Council of Chief State School Officers (CCSSO) report on teacher education titled "Our Responsibility: Our Promise," was written by the Task Force on Educator Preparation and Entry into the Profession. This task force was comprised of a group of nine individuals from the CCSSO, two from the National Governors Association (NGA) and three from the National Association of State Boards of Education (NASBE) in consultation with twelve members of an "Expert Advisory Group" (p. 35). In the report's Executive Summary, the task force defined a "learner ready teacher" (p. iii) as someone who is

> …ready on day one of his or her career to **model and develop** in students the knowledge and skills they need to succeed today including the ability to think critically and creatively, to apply content to solving real world problems, to be literate across the curriculum, to collaborate and work in teams, and to take ownership of their own continuous learning. More specifically, learner-ready teachers have **deep knowledge of their content and how to teach it,** they **understand the differing needs** of their students, **hold them to high expectations**, and personalize **learning** to ensure each learner is challenged; they **care about, motivate, and actively engage students in learning**; they **collect, interpret and use student assessment data** to monitor progress and adjust instruction; they systematically **reflect, continuously improve, and collaboratively problem solve**; and they **demonstrate leadership and**

> **shared responsibility** for the learning of all students. (CCSSO, 2012, pp. iii-iv, emphasis in original)

This list looks much like a compilation of the complex definitions of teacher quality cited in this document, evidence that few would disagree with such a comprehensive statement which clearly demonstrates the complexities involved in teaching. This statement from the Executive Summary can be seen, then, as a statement of intent. The goal is for teachers to be able to develop the knowledge, skills and dispositions to allow them to accomplish each aspect of teaching delineated by this list.

The next complex question is how teachers might be educated in order to meet each of these goals, a question that has long been the focus of teacher education programs. The writers of the 2012 CSSO report, however, determined that the new Common Core Standards set such a high bar for students that newly certified teachers entering the field aren't sufficiently prepared to meet these new requirements. The report simply states: "current policies and practices for entry into the education profession are not sufficient to respond to this new challenge…" (p.1). This report does not cite the research upon which this statement is based. In fact, 2010 was the earliest any state adopted the Common Core Standards. This leaves one to wonder how teacher education programs would have even had time to respond, let alone how research studies could have been conducted to determine that teacher education policy was insufficient.

Because the 2012 CCSSO report defined this problem, however, the rest of the report makes recommendations for states to take action. The report acknowledges there are limitations on what state and federal regulations are able to do to hold teacher education programs accountable for every aspect of preparing learner-ready teachers, as defined in the report. Many of those qualities, the report acknowledges, are not easily measured for purposes of accountability. Further, the report continues, the jurisdiction of state or federal regulatory bodies is limited. Therefore, the Task Force "focused on areas where chiefs have responsibility," (p. 6) and "the recommendations focus on what chiefs and their agencies and partners have authority to exercise" (CSSO, 2012, p.6). The impact of the admittedly limited policies set in the CAEP standards, however, and the accompanying accountability measures and consequences, is far-reaching. If these are the standards by which preparation programs will be held accountable by accrediting agencies, these are the standards that will become the focus of every Teacher Education Program in the country, thus standardizing programs in very specific ways.

In making its recommendations, the CCSSO (2012) targeted only those aspects of the complexity of teacher and program evaluation they felt were within the jurisdiction of states to regulate. It then became the role of CAEP to determine answers to questions like: What evidence will be used to evaluate whether or not a teacher education program is preparing teachers to be learner-ready? What factors will define program success and failure? What will the consequences be when a program is found lacking? A close read of the CAEP standards (2013) makes it clear

that the standards rest squarely on the recommendations of the 2012 CCSSO report, so that policy is consistent across state agencies and with the now national accreditation standards.

In order to remedy the problem with teacher education defined in the CCSSO report (2012), the authors spell out recommendations for "state actions" (p. iv) that they ask the full CCSSO to commit to working toward in their own states. These actions include recommendations in three areas: teacher licensure, teacher educator program approval, and data collection, analysis and reporting. Clearly, these are the areas over which the authors determined states had jurisdiction.

The section of the report focused on program approval is very specific in the action it expects from states. These recommendations call for unprecedented levels of state control over teacher education programs, housed, as they are, in colleges and universities where academic freedom has long been a core value. Recommendation 5, for example, says:

> States will hold preparation programs accountable by exercising the state's authority to determine which programs should operate and recommend candidates for licensure in the state, including establishing a clear and fair performance rating system to guide continuous improvement. States will act to close programs that continually receive the lowest rating and will provide incentives for programs whose ratings include exemplary performance.
>
> (CCSSO, 2012, p. v)

Recommendation 6 says,

> States will adopt and implement rigorous program approval standards to assure that educator preparation programs recruit candidates based on supply and demand data, have highly selective admissions and exit criteria including mastery of content, provide high quality clinical practice throughout a candidate's preparation that includes experiences with the responsibilities of a school year from beginning to end, and that produce quality candidates capable of positively impacting student achievement.
>
> (CCSSO, 2012, p. v)

Kentucky's response. By all accounts, Kentucky has taken its charge by the CCSSO very seriously. Perhaps this isn't surprising, given that Terry Holliday, Kentucky's Commissioner of Education, was Vice Chair of the CCSSO Task Force and Kentuckian Jim Cibulka, President of the newly formed Council of Accreditation of Educator Preparation (CAEP), was a member of the report's expert advisory group (CCSSO, 2012).

In 2013, "Design of an Education Professional Standards Board (EPSB) Preparation and Accountability System for Teacher Training Programs," a report by Terry Hibpshman, was published by Kentucky Education Professional Standards Board (EPSB). According to Hibpshman, the focus of education policy in the U.S. changed following the publication of two key reports in the 1980's. One was published by the National Commission on Excellence in Education in 1983, which claimed the U.S. school system was mediocre at best. The other, by the Carnegie Forum on Education and the Economy, published in 1986, focused on what it

determined to be problems in U.S. Teacher Preparation Programs. Hibpshman identifies this as the point where the key question in education shifted from whether schools had the resources they needed to do their jobs to whether "the product being produced was adequate to serve the presumed goal of education as an economic engine" (p.2). As Gorski (2013) and Sleeter (2013) point out, with this shift in language and focus, educational goals were suddenly being defined in corporate terms and for corporate ends, which aligns with what each author identifies as a neoliberal agenda. This makes it clear that a philosophical stance has been chosen.

And if the "product being produced" (Hibpshman, 2013, p. 2) is determined to be inadequate, as it is in the eyes of policy makers advocating for accountability, something must be determined to be the cause. Citing a study by Sanders & Horn (1994), which Hibpshman says was "echoed by many" (p. 8), teacher quality was determined to be "the single largest factor in student achievement" (p.8). Since the 1990's, then, evaluation of teacher performance has been at the center of national attention. Citing his own unpublished manuscript and a 2012 study by Goldhaber & Hansen, Hibpshman modifies Sanders and Horn's statement with the addition of this important caveat: "teacher performance is the greatest education factor *amenable to administrative management*" (p. 8, emphasis added). In other words, like the recommendations in the 2012 CCSSO report, regulating teacher performance standards was viewed as being within the jurisdiction of the state. It was also deemed possible to establish policy that at least had the appearance of positively impacting teacher performance, which Hibpshman said would be a politically expedient move

(2004). This shift in focus to "product" outcomes provided the rationale for establishing new far-reaching accountability measures for teacher preparation programs in an attempt to "manage the quality of the teacher workforce" (p. 8).

Although acknowledging that in any large and complex accountability system results are inherently ambiguous, the stated purpose of Hibpshman's 2013 report is to draw inferences from educational and industrial models to inform the establishment of evaluation procedures for Kentucky's teacher education programs, which might, in turn, inform the nation. Based on models Hibpshman cites in studies by Baker (2005), Newmann, King & Rigdon (1997) and O'Day, (2002), Hibpshman identifies basic elements common among many kinds of accountability systems. They include "at least some set of performance goals, together with measurements for evaluating whether the goals have been accomplished, and some set of rewards and sanctions for performance" (p. 4). Recent policies enacted or under consideration in Kentucky demonstrate that these are the very elements guiding policy decisions in Kentucky's efforts to hold teacher education programs accountable for the quality of the teachers they educate.

Interestingly, Daniel Pink's (2009) analysis of psychological research on motivation, which he says is very well documented and frequently ignored, is in direct conflict with the recommendation for a system of rewards and sanctions in Hibpshman's report. Pink's research finds that for simple tasks rewards and sanctions can be effective, but for more complex tasks they actually act as *de*-motivators. Instead, Pink's analysis of the research indicates that systems which allow people and

programs autonomy to establish their own sense of purpose and work toward mastering goals they have aligned to that purpose lead to far greater innovation than does the use of punishment and rewards.

Kentucky has been working on the development of a system of accountability across all levels of the educational spectrum (K-16) since 2011. The Kentucky Department of Education (KDE) has developed a new Professional Growth and Effectiveness System (PGES), being implemented in the 2014-15 school year, for the purpose of gauging the effectiveness of classroom teachers (Kentucky Department of Education, 2014). This system includes various measures of teacher performance, one of which is an extrapolated Student Growth Percentile (SGP). SGP is the rate of change in each student's test scores from year to year as compared to peers who score similarly on standardized tests (Allred, Draut, Ellis & Liguorni, 2014). SGP is reported as a percentile and the system will require two test scores a year for each student in each subject assessed (Allred, et al., 2014). The stated theoretical premise basis for this kind of a measure is:

> When students with "like" scores are placed in an academic peer group and then compared one year later, we *assume* teacher and school actions happened between the two tests to cause a student to stay even with or out- perform the academic peer group. The actions may include instruction, curriculum, on-going assessments, etc.
>
> (Allred, et al., 2014, slide 22, emphasis added)

Based on the assumption that teacher effectiveness is a central factor in this equation, the state has established cut scores which label learner growth as low, expected, or high, based on a bell curve distribution. In this way, SGP will be used to determine teacher effectiveness (Allred, et al., 2014). The important underlying assumptions that teacher effectiveness can be evaluated through the use of test scores extrapolated in this way, *and* that teacher effectiveness really *is* the main factor determining student academic growth, have not been well researched, according to a policy brief posted on a New Jersey Education Policy Forum website (Baker & Oluwole, 2013). The brief concludes that many states are turning to the SGP model as a way to measure teacher effectiveness because much has been written recently to invalidate the Value Added Measures (VAM), a proposed alternative. Baker and Oluwole write:

> …there has been far *less* research on using student growth percentiles for determining teacher effectiveness. The reason for this vacuum is not that student growth percentiles are simply immune to problems of value-added models, but that researchers have until recently chosen not to evaluate their validity for this purpose – estimating teacher effectiveness – *because they are not designed to infer teacher effectiveness* (2013, section 2, emphasis added)

Despite this lack of research and possible misuse of test data, Hibpshman (2013) says the teacher effectiveness data generated by the PGES system is important to the EPSB because it will *also* play a role in evaluating teacher preparation programs. This begs the question, of course, about the research base for this decision,

especially given Baker and Oluwole's (2013) claim that the SGP system was not designed even to measure teacher effectiveness.

In Kentucky, an elaborate computerized tracking system, a "workforce dashboard" (EPSB Data Dashboard, 2014; Hibpshman, 2013) is being put into place in order to follow the progress of pre-service students through their teacher education programs, and then continue tracking them into their teaching careers. The Teacher Effectiveness Scores, based at least in part on SGP as measured by standardized tests, will then be tracked back to the programs where the teacher was educated (Walters-Parker, 2014), and used as part of the evaluation data for that program. In fact according to Hibpshman (2013), through its use as an assessment tool, the PGES system will also provide a vehicle for the state to play a role in defining the focus of teacher education programs.

Acknowledging the need for any Kentucky teacher preparation accountability system to take into consideration new standards set by the Council for Accreditation of Educator Programs (CAEP) in 2013, Hibpshman (2013) outlines three areas of accountability by which Kentucky should measure teacher preparation programs: "program management," "program processes" and "measures of the effect of programs on local educational systems, either at the district/school level, or through the performance of individual teachers" (p. 12).

These three areas are then broken down into recommended accountability measures outlined in tables at the end of the report. In the first of three tables (pp. 21-27), Hibpshman identifies twenty-five different principles for which educator

preparation programs should be held accountable. The table shows how each principle aligns with five goals set by EPSB and the ways in which each principle might be measured. Table II (pp. 27-29) breaks each of EPSB's five goals into twenty-eight actionable strategies, and educator preparation programs will be required to provide evidence that they are utilizing each strategy. Table III (pp. 29-32) includes details about *measures* for each of the twenty-five principals and *consequences* for programs not meeting the target set for each measure. Hibpshman recommends that the sanctions should escalate if, after additional support, programs continue to miss the targets set by the state. These sanctions would include "[i]dentification of the program as a poorly–performing program" (p. 16), placing "[l]imitations on the program's privileges (e.g., limitations on the level, content areas, or geographic locations permitted to the program)" (p. 16) and, finally, "[d]ecertification" (p. 16). Using the structure already fully outlined by Hibpshman in this 2013 report, next steps for EPSB staff, he says, should include setting appropriate target values for each accountability measure and further refining "the consequences model" (p. 17).

Impact of the Hibpshman report. Not all of the accountability measures and consequences outlined in the 2013 Hibpshman report were yet in place, but in a PowerPoint presentation given at a CAEP Conference in Nashville (March, 2014) Kim Walters-Parker, Director of the Division of Educator Preparation for Kentucky's EPSB, directed the audience to this report in order to learn more about plans for Kentucky's "accountability suite" (slide 19), implying that her division is closely

following the report's recommendations, and that more regulations and consequences would be forthcoming.

The pace of changes in requirements resulting from so-called "reform" policies and "accountability" measures have been difficult and costly for Kentucky teacher education programs, and the institutions that support them, to keep up with. Previously, programs had been required to align their priorities with conceptual frameworks they wrote, allowing them to define program goals and align them with the mission and goals of their supporting institutions. According to Bobby Ann Starnes, former Education Studies Program Chair at Berea College, the purpose of the required writing of conceptual frameworks was to show assessors the philosophical framework underlying the program goals which were used as a base to establish program priorities (2012, personal communication). Over the course of just a few years, by 2014 the reach of the state had extended into almost every aspect of teacher education. As Sleeter (2013) points out, this attempt to standardize policies and programs is based on a philosophical stance policy makers have chosen and now put forth as if it is correct, claiming it is based on what they have determined should "count" as research. In the complex system of teacher education, with many important areas of emphasis, state policy now determines priorities which used to be determined by programs themselves.

Over the course of about five years, the state's primary role in teacher education had changed from setting and enforcing requirements for teacher

certification, with teacher education programs being afforded autonomy in preparing candidates to meet those standards, to Kentucky state regulations determining:

* *who may enter* a teacher education program: Students are now required to pass a Praxis I exam in three areas of basic skills: reading, writing and math. Cut scores for each exam are determined at the state level (Hibpshman, 2004). GPA requirements were also raised (from 2.5 to 2.75) for admission to a teacher education program (16 KAR 5:020);

**how* candidates should be *taught*: Students are required to complete and enter 200 field hours into a state data-base, which breaks the hours down into specific required experiences that must be completed and verified before student teaching (16.KAR 5:040). Students must also pass the Praxis II exam for certification (16 KAR 5:040), which by implication determines at least some content in required courses (Hibpshman, 2013).

*how students should be *evaluated*: Kentucky adapted the "Framework for Teaching" (Danielson, 2011) to be used for teacher evaluation, which will also impact pre-service teacher evaluation. Kentucky has set standardized testing requirements and cut scores before a candidate can become certified (16 KAR 5:040), and has regulated the length and type of student teaching experience a candidate must have as well as requirements that teacher education programs "train" cooperating teachers in methods of co-teaching, through an EPSB-approved program (16 KAR 5:040).

Evaluation as an inferential process. The intent behind Kentucky's development of a shared accountability model that crosses state agencies, including

EPSB, is to ensure high teacher quality for all Kentucky students (Hibpshman, 2013). As more data are collected and measurement systems are field-tested, Kentucky's model will continue to be studied and refined (Walters-Parker, 2014). Hibpshman (2013) acknowledges that no accountability system can measure many of the goals set by institutions, and further, that no complex accountability system is without limitations. In fact, Hibpshman says that assessing teacher preparation program performance is, by nature, an "inferential process" (p. 18) where results are never certain. Despite this statement, near the end of his report he makes an argument for the implementation of such a system:

> To the extent that we create measures of the goals we hope to achieve, and evaluate the results within a consistent decision-making framework, we are likely to make better decisions about program quality than we would without such a framework. If we monitor the performance of the accountability measures and make adjustments when necessary, we can deal effectively with the problems of uncertainty and nonmonotonicity" (p. 18).

At great expense and much effort on the part of many, time will tell whether Hibpshman's hypothesis that the new system will increase the likelihood that the state will be able to make better determinations about the quality of teacher preparation programs. In the meantime, with sanctions in place, some programs will be lauded, some will be closed, and overall, programs are likely to look more and more alike, which, of course, is the impact, and often the intent, of any kind of standardization. Some will view such standardization as a good thing…as if it is a way to effectively

ensure program quality. Others will raise questions about what might be lost as programs become more standardized.

As is true with all practice (Collins, 2004), each decision made on the state level is based on a philosophical stance. When a state or federal agency selects a philosophy on which to base policy, it necessarily impacts the practice of teacher education programs. Therefore, this standardization of expectations reduces program autonomy in the areas of both theory and practice. It is also important to acknowledge the impact of policies on outcomes for students. Policy changes affect students' lives and career choices. Regardless of intent, when policies establish barriers for students desiring to enter a program or profession, it is important to study the *impact* of these policies, which will reverberate across the nation. Only then can a cost benefit analysis occur. It is in this analysis, and the choices made as a result, that the core values of an institution, a state and a nation come to the fore.

Down a Rabbit Hole: A Close Examination of the Impact of a Single Policy Shift: Praxis Exam Requirements for Teacher Certification Candidates in Kentucky

Often in U.S. history, policy decisions have impacted groups of citizens differently. Policies intended to "raise standards" have historically had adverse impacts on marginalized populations by further restricting their access to opportunities, regardless of the intent of those policies (Gorski, 2013). This is one reason Gorski (2013) and others refuse the term "achievement gap" and choose instead to name what he says the test score gap actually represents…an opportunity gap. As author john powell (2012) points out in *Racing to Justice: Transforming Our*

Conceptions of Self and Other to Build an Inclusive Society, policies set at federal, state, or institutional levels often have outcomes that keep a white norm in place. Following a policy from intent to impact is, of course, complicated, because policies operate within complex systems which are rooted in the history of the United States as a nation. Inarguably, these historic roots are racist. While blatantly racist laws and policies have largely been taken off the books, the legacy of those laws from a not too distant past continues to reverberate. Further, policies and laws that result in disparate outcomes for different groups of people remain and are widely accepted (powell, 2012).

Because each of the many new policies in teacher education has such a complex history, it might be useful to engage in a close study of a single policy which impacts teacher education in Kentucky and many states across the nation: the requirement that teacher education candidates pass one or a series of standardized exams known as the Praxis Series, created by the Educational Testing Service (ETS).

Intent. The public and even people within a system impacted by a policy often hear about the policy's *intent*, but rarely about its *impact,* actual or projected. In the case of educational policy in this era of accountability, as it is commonly known, the focus is almost solely on learning *outcomes* for students and teachers, and on finding ways to "prove" these outcomes. In other words, what matters is that educators can demonstrate the impact of their teaching.

A policy itself, however, is clearly intended to fix a perceived problem. If based on solid research, an outcome may be predicted, but it can almost never be

known. Sometimes policy makers and the media focus so intently on a perceived problem and the *intent* of policies put in place as an attempt to solve it, that the *outcomes*, known or projected, are virtually ignored. In some cases, even a system's stakeholders become so caught up in the act of implementing these policies that they, too, fail to analyze the outcomes.

In the case of policies requiring that students in teacher education programs pass Praxis exams, the problem identified is that teacher quality in the U.S. is too low. Linda Darling-Hammond, et al. (2009) said this public perception began in the mid-80's with reports by the Carnegie Task Force, the Holmes Group and a "collection of analysts, policy makers, and practitioners of teacher education" (p. 613). As a result, raising teacher quality has become a major focus of school reform efforts in the U.S. (Jennings, 2012).

Concerns about teacher quality remained a focus of the 2013 CSSO report, which simply stated that new teachers don't have the knowledge or skills needed to support students in meeting the goals of the Common Core standards. The report extrapolated this problem to its perceived origins in teacher education programs. It concluded that policies of the time were inadequate to ensure that teachers would be prepared in ways that would allow them to help students meet the standards of the Common Core, which at the time of the report had only recently been adopted in many states (CCSSO, 2012). Rather than allowing programs of teacher education time to define how courses might be adapted to include content based on the new standards, the CCSSO instead asked states to take on a regulatory role, intended to

ensure teachers would have the levels of content knowledge needed to teach the new standards (CCSSO, 2013). CAEP, the national accrediting body formed in 2013, has also incorporated requirements intended to raise standards of teacher quality into its standards (CAEP Accreditation Standards, 2013).

Yet ensuring that teachers are "classroom ready" as they exit teacher education programs is not a straightforward task. Rather, charging states with setting policies to try to ensure this outcome is clearly an exceedingly complex task. Perhaps, then, it is no surprise that many states have turned to Praxis Exams as a way to measure candidate readiness. After all, the *Praxis* Technical Manual (2010) states:

> Praxis tests assess a test taker's knowledge of important content and skills required to be licensed to teach. States adopt the *Praxis* tests as one measure of helping to ensure that teachers have achieved a specified level of mastery of academic skills, subject area knowledge, and pedagogical knowledge before they grant a teaching license (p. 9).

For states newly charged with finding ways to document program improvement or the maintenance of high standards among teacher candidates, the use of Praxis tests, with results that seem to assure that candidates have gained the necessary knowledge and skills, seems, at first glance, an important piece of a complex puzzle.

How is Praxis information used in Kentucky? As a result of the charge by the CCSSO, regulations put into place in Kentucky to document and improve teacher quality included the use of *Praxis* exams at two points in the education and certification process. Praxis I is a test intended to assess the reading, writing and math

skills that have been "identified as important for a career in education" (Murphy, C., 2013, slide 12). The Praxis I test is now used as a way to determine who will or will not be allowed to enter an accredited teacher education program anywhere in the Commonwealth, at both public and private institutions (16 KAR 5:020, 2014).

In an ETS document titled "Proper Use of Praxis Exams" (2000), ETS claims Praxis I may be validly used to identify "rising juniors" (p. 3) who have "sufficient reading, writing and mathematics skills to enter a teacher preparation program" (p. 3). Presumably, then, colleges and students have two years to make up any core content knowledge that students may not have learned during their high school educations. In 2014, a new version of Praxis I was released, focused on the Common Core Standards (Murphy, 2013). This will make the next ten to twelve years especially challenging for teacher preparation programs and their institutions, as students entering college will not have received the twelve years of education in the Common Core Standards upon which the Praxis I test is based. Regulations in Kentucky now state that students may not take any course intended to support preparation for teacher certification without first being accepted into the institution's teacher education program, which requires having passed Praxis I (16 KAR 5:020, 2014). One outcome of this requirement is that students who are committed to wanting to become teachers but don't pass Praxis I the first time must delay their educations until they have successfully met the cut score on the exam. This delay, in many cases, is a costly prospect. Students who can't afford the extra tuition, the extra time, or the cost of taking the Praxis Exams multiple times, may be forced to choose another major. This

is one example of a policy with a disproportionate impact on students from different socioeconomic groups. Yet the inequity of this policy has rarely been questioned.

Once students who have been allowed into a teacher education program have successfully completed it, they must again pass a standardized test, Praxis II, in order to become certified as a Kentucky teacher (16 KAR 2:010, 2014). Again the cost of these tests is born by the test-taker. The booklet *Proper Use of Praxis Exams* (ETS, 2000) states that Praxis II may validly be used for teacher licensure decisions. It goes on to say that each state is responsible for verifying that the content of the test is appropriate for the way the state uses test results, and states also bear responsibility for establishing appropriate cut scores for each test. The ETS Proper Use document goes on to state: "Proper use is a joint responsibility of ETS…and of states, agencies, associations and institutions of higher education, as the users of assessments" (p. 6). The assumption of test validity is predicated on verification by the state and institutions that the test covers content material that students have been exposed to during their college career (p. 4). The state has no way of ensuring this, except to assume programs will adapt by incorporating test content into their courses. This seems a desirable expectation if the test content reliably contains the information teacher candidates need in order to become successful teachers. Indeed, this is the effect Criterion Referenced Tests are designed for.

What are Praxis Tests? Praxis Exams are Minimum Competency Tests, designed to compare student performance to a set of expectations, in contrast to norm-based tests like the SAT. Minimum Competency Tests are also referred to as

Criterion Referenced Tests (CRT) (Kortez, 2008). The use of CRTs has grown in recent years, because the tests are seen as a way to shape educational practice in ways intended to improve instruction, an idea that has now become a cornerstone of U.S. education policy (Koretz, 2008).

In a 2014 *Educational Leadership* article about criterion referenced assessment, W. James Popham, who was one of the first researchers to focus on criterion referencing, claims that a common misperception is that there are two kinds of tests—norm referenced and criterion referenced. In fact, he says, the contrast is not in the type of test, but in the interpretation of test scores, and the kinds of inferences made based on those interpretations. Scores can be interpreted in a norm referenced way, according to a bell curve, where student scores are compared relative to the scores of other test-takers, a norm group, or to scores of past groups over time. Many familiar tests such as the SAT, ACT, and GRE are norm-referenced tests.

In contrast a student's scores might be interpreted in reference to a set standard (or criterion), which eliminates the need for comparisons between students or with a norm group. Popham describes this criterion referencing as an absolute rather than a relative measure. Some tests are deliberately developed for one type of interpretation or the other, but, as is the case of Praxis Exams, it is possible for test scores to be interpreted and used in both ways (Praxis Technical Manual, 2010).

According to Popham (2014), criterion referenced testing originated with a 1963 article by Robert Glaser, who was a student of behaviorist B.F. Skinner. Programmed instruction grew out of Skinner's behaviorist theories, which in turn

gave rise to criterion referenced assessment. Popham explains that in its development, the purpose of CRT was to allow educators to target instruction toward established criteria, and use ongoing test results to refine that instruction until the greatest possible number of students reached the standards measured by the CRT. Its intent, in fact, was to *encourage* instructors to teach to a test which was built to accurately measure each standard. CRT, therefore, originated from an educational theory, behaviorism, which is based on a clearly defined philosophical stance.

Popham (2014) explains there was an initial lack of clarity among practitioners, some of whom interpreted that CRT criterion should be a pre-determined level of performance rather than a set of standards. By the late 1970's, however, Popham says that most assessment specialists had agreed that the criterion defined should be specific knowledge or skills rather than a specific level of performance. In Popham's view, CRTs are most effective when they are based on the exact knowledge or skills students should learn. Instructors should then teach directly to those targets, and use the test to determine whether the targets were reached. This, of course, requires well established, agreed upon, measureable criterion. In the current accountability climate, however, Popham regrets that rather than following this model, many testing specialists have returned to the idea that the criterion should be a certain overall level of performance, which Popham says is much less helpful in guiding instruction. In the Glossary of Standardized Testing Terms, ETS explains that "[c]riterion referencing is often defined in terms of proficiency levels. The test score

required to attain each proficiency level is specified in advance" (criterion referencing, ETS, 2014).

ETS description: Praxis I and II. The Praxis Technical Manual (2010) says the Praxis Series "reflects what practitioners in that field across the United States believe to be important for new teachers. The knowledge and skills measured by the tests are informed by this national perspective, as well as by the content standards recognized by that field" (p. 9). The Praxis Study Companion (2013) clearly states that Praxis tests do not measure "potential for teaching success" (p. 35) nor teaching ability, because "teaching combines many complex skills that are typically measured in other ways, including classroom observation, videotaped practice or Portfolios not included in the Praxis test" (p. 35). These statements appear to be in conflict, and the issue is further confounded in the following statements about purpose and statements about uses for Praxis exams.

Praxis I tests are "designed to measure basic competency in reading, writing, and mathematics" (*Praxis* Technical Manual, 2010, p. 10). ETS reports the primary use of Praxis I data as "a way of evaluating test takers for entrance into teacher education programs" (p. 10).

The professional portion of the Praxis II Exam is called Principals of Learning and Teaching (PLT). The PLT is designed to "address teaching pedagogy at varying grade levels by using a case-study approach combined with multiple-choice (MC) and constructed-response (CR) items" (p. 10). This test has recently been revamped to include more multiple choice items and fewer constructed response items in order to

increase reliability (Murphy, C. 2013). Praxis II also includes a content test to "cover general or specific content knowledge in a wide range of subjects across elementary or middle school (or both) grade levels" (p. 10). In Kentucky, secondary and middle school candidates are required to pass the PLT and the content test in their major subject area(s); elementary teachers must pass each of four content tests in addition to the PLT (KY Test Requirements, 2014). According to the Praxis Technical Manual (2010), "[t]he test provides states with a standardized mechanism to assess whether prospective teachers have demonstrated knowledge believed to be important for safe and effective entry-level practice" (p. 10).

Hibpshman (2004) explains that Kentucky regulators have three checkpoints for a teacher candidate's academic proficiency: at program entry, before certification and then again during the initial year of teaching, which is considered an internship year. Each of these stages is "designed to eliminate unsuitable persons from the teaching profession" (p. 6). At the first two stages, Praxis scores play a large role in determining a candidate's "unsuitability." Hibpshman points out that Kentucky does allow a candidate to take the Praxis multiple times (provided they can afford the time and expense) in an attempt to meet the cut score. A candidate who hasn't passed Praxis II can also be hired by a district on a conditional certificate, provided the district creates a plan to support the teacher in "remediating the prospective teacher's deficiencies" (p. 6), which would presumably allow them to pass the test.

Praxis Exams: Construction, Purpose and Use. The Glossary in the ETS document *Understanding Your Praxis Scores 2014-15* defines validity as "[t]he

extent to which test scores actually reflect what they are intended to measure. *The Praxis Series* tests are intended to measure the knowledge, skills, or abilities that groups of experts determine to be important for a beginning teacher" (p. 2). Therefore, according to the Praxis Technical Manual (2010), "The *Praxis* tests provide states with the appropriate tools to make decisions about applicants for a teaching license" (p. 11).

The Princeton Review website (2014) reminds educators to carefully consider what tests are being used, for what purposes, and ways in which any selected tests impact teaching and learning, especially when the stakes are high. In a research memorandum for the Educational Testing Service (ETS), creators of the Praxis Series, Drew Gitomer and Andrew Latham (2000) write, "the temptation to generalize about issues facing teacher education, and their potential solutions, are often simplified to the point of being misleading…We also argue that the academic ability of teachers is not adequately characterized by broad generalizations…" (Abstract). This statement begs a question about the use of Praxis scores as if they were a true base line for pre-determining a candidate's suitability for entry to the teaching profession, or whether the results might, instead, be considered a "broad generalization" about a candidate's knowledge and skills. In other words, *can* a Praxis test accurately measure the knowledge, skills and abilities important for a beginning teacher?

Test validity in modern theory. To answer this question, it is important to look more closely at the construction of Praxis exams. Harvard professor and psychometrician Daniel Koretz (2008) writes:

> There seems to be a widespread faith in the wizardry of psychometrics, a tacit belief that no matter what policymakers and educators want a test to do, we can somehow figure out how to make it work…[however] test design and construction entail a long series of trade-offs and compromises (p. 327-8).

Hibpshman (2004) describes the difficulties of designing a test that could measure the quality of a teacher's performance. He writes:

> Validity is often inaccurately described as the ability of a test to predict performance on some criterion, but in fact it is often very difficult to establish what, if anything, a test score predicts. This one of the fundamental problems in testing for employment purposes, where the nature of adequate performance is always at least a bit murky. Prediction of performance on a criterion is one type of inference that establishes validity, but is by no means always either a necessary or sufficient condition. (p. 7)

Hibpshman (2004) then explains that there are differences between modern (post 1970) and classical test theory in the "types of inferences necessary to demonstrate the value of a test" (p. 7). Modern theory is so much more complex, Hibpshman explains, that most people outside the psychometric community, "particularly federal regulators and the courts," (p. 7), tend to frame their ideas about test construction in terms of classical theory. In other words, based on their

knowledge of classical theory, the assumptions many people make about how a test is designed and its ability to measure what it is "supposed" to measure may no longer be true. Hibpshman explains, "A test is never really valid in any abstract sense" (p. 8), and continues: "[v]alidity in modern terms represents a chain of inference establishing that a test provides useful information for a particular purpose" (p. 8).

Since people's lives are impacted by the results of these tests, it seems wise to examine at least a few of the key places where inferences must be made in the "long series of trade-offs and compromises" (Koretz, 2008, p. 327-8) involved in the construction and implementation of Praxis exams.

Test Content Selection. Koretz (2008) implores practitioners and policy makers to keep in the forefront of their minds that test content is meant to act only as a proxy, a small representation of a much broader range of knowledge and skills, too broad and deep to be measured by any single instrument. While anyone who has ever taken a standardized test can attest to this, it is an important piece of information that Kortez says is widely ignored when tests scores are used to make high-stakes decisions, as if these scores are irrefutable evidence of the knowledge and/or skills they were designed to measure. While Koretz sees a role for standardized testing in education (he is, after all, a psychometrician), he reminds us that test scores can only ever give limited information about student learning or achievement (Koretz, 2008). Because a deep understanding of this concept is vital to determining how a test might be used, researchers Heather Hill, Kristin Umland, Eriza Litke and Laura Kapitula

(2012) ask educators and policy makers to take a close look at the way content for exams like the Praxis Series is selected.

Every researcher cited in this document agrees that because the skills and knowledge required for teaching are so complex and interwoven, they are exceedingly hard to analyze and define. How could test designers ever possibly come up with test items that act as proxies for such a broad complexity of knowledge and skills? Citing the 1999 S*tandards for Educational and Psychological Testing,* the Praxis Technical Manual (2010) states: "The main source of validity evidence for licensure tests comes from the alignment between what the profession defines as knowledge and/or skills important for safe and effective practice and the content included on the test" (p.15). The Manual goes on to say that the Standards require that a job or practice analysis be performed in order to create and validate test content for any test that leads to professional licensure. Therefore, a close look at the process of a job analysis might lead to a more specific understanding of this process of test content selection.

***Job analysis*.** The Praxis Technical Manual (2010) references Knapp and Knapp (1995), the creators of a practice analysis process, who explain, "the foundation upon which to build a viable and legally defensible licensure examination" (p. 1) is through "…the systematic collections of data describing the responsibilities required of a professional" (p. 1). The Praxis Technical Manual then defines this systematic data collection process:

Praxis I and Praxis II tests use a job analysis process as follows:

- A review of available professional literature and disciplinary (content) standards to develop a draft domain of knowledge and/or skills
- Meetings with a National Advisory Committee of experts to review and revise the draft domain
- A survey of the profession to confirm the importance of the committee-revised domain (see, for example, Knapp and Knapp, 1995; Raymond, 2001; Tannenbaum and Rosenfeld, 1994).

(p. 16).

The Praxis Technical Manual (2010) then explains that the National Advisory Committees (NAC) must be diverse across a broad spectrum, including: race, ethnicity, gender, practice settings, grade levels, regions of the country, and professional perspectives. The Manual explains that committee members are nominated by professional organizations, superintendents, deans and state departments of education. Committee members can also self-nominate, as there is a call for applicants to the NAC on the Praxis website (Keeping Test Content Current, Praxis, 2014).

At this point, one might envision a room full of experts sitting around a table, discussing and debating the all-important question of what beginning teachers should know and be able to do in a given teaching discipline. This would certainly not be an easy conversation, given the various philosophical stances held by teaching "experts." The webpage titled "Praxis National Advisory Committees (NAC)" specifies that

"[e]ach NAC typically consists of a diverse group of 12-15 licensed practitioners (i.e., teachers or school leaders) and higher education faculty who are involved in teacher preparation for a particular subject area or licensure test" (2014). Knapp and Knapp (1995) state that, in fields with a great amount of variability in "theoretical orientation or professional practice" (p. 7), this number of practitioners is "barely enough" (p. 7) to represent such a broad range of views.

Once selected, the NAC is involved at two steps in the test design process. ETS explains that at each stage, their psychometricians work closely with the committee to ensure the test will meet standard specifications for test design (Keeping Test Content Current, 2014). The committee's first task, as described above, is to review a "draft domain of knowledge and/or skill statements believed to be important for entry level practice" (Praxis Technical Manual, 2010, p. 19), which had been previously compiled by ETS employees. These statements are intended to become part of a job analysis survey of practitioners, in which they are asked to identify what they consider to be "core tasks and core knowledge" (Knapp & Knapp, 1995, p. 9) in their field.

The NAC is to review the statements to determine if they are: 1) important for "safe and effective practice" (Praxis Technical Manual, 2010, p. 19) 2) needed for entry-level teaching and 3) clearly written. Statements that don't meet all three criteria are revised or eliminated from consideration for the job analysis survey. The *Praxis Technical Manual* (2010) explains that the purpose of this survey is "to obtain independent judgments of the importance of the knowledge and/or skills defined by

the committee" (p. 16). Therefore, rather than *defining* a set of knowledge and skills deemed necessary for entry into the profession, the committee actually *reviews* and *revises* a set of statements compiled by ETS employees (Praxis Technical Manual, 2010).

Once this proposed domain of skills and knowledge is approved by the NAC, it is

> …administered as a survey to a large sample of teachers and college faculty for verification of the judged importance of the knowledge and/or skills for entry-level practice. The outcomes of the survey are then used by the NAC to develop test content specifications
>
> (*Praxis Technical Manual*, p. 19)

The second stage of NAC involvement is to re-convene to discuss survey results, once they are in. "[U]nder the guidance of ETS test developers" (*Praxis Technical Manual*, p. 19), the NAC uses the survey information to "construct the test content specifications" (p. 19). It is important to note that the committee is not involved in the development of test *items*. That is done by "content experts, external to ETS" (p. 20), who use these specifications as a guide.

According to the Praxis Technical Manual (2010), test specifications are documents that inform stakeholders of the essential features of the tests. These features include:

- A statement of the purpose of the test and a description of the test takers
- The major categories of knowledge and/or skills covered by the test and a description of the specific knowledge and/or skills that define each category; the proportion that each major category contributes to the overall test; and the length of the test
- The kinds of items on the test
- How the test will comply with *ETS Standards for Fairness and Quality*

(*Praxis Technical Manual*, p. 20)

Hill, Litke, and Kapitulta (2012) call this job analysis process to question. Knapp and Knapp too, imply there has been some controversy around this process, when they state:

> Whether one views the process as soporific or a public spectacle, the *fact* remains that the systematic collection of data describing the responsibilities required of a professional…is the foundation upon which to build a viable and legally defensible licensure examination.
>
> (p. 1, emphasis in the original)

. The process of developing and implementing the survey of practitioners for a job analysis is explained in more detail by Knapp and Knapp (1995). They say that for licensure exams, the 1978 *Guidelines on Employee Selection Procedures* specify

that only "observable work behaviors and tasks and work products" (p. 2) may be included "as opposed to personality and other individual characteristics that are not directly observable" (p. 2). Guidelines for test creation also exist; Hibpshman (2004) refers to the *Standards for Educational and Psychological Tests* as the "'Bible' of test construction" (p. 7). There has, however, never been a Supreme Court case challenging the valid use of professional tests (Knapp & Knapp). The legal rationale for using a job analysis process is to establish the levels of knowledge and skills deemed crucial for responsibly *protecting the public from harm*, which Knapp and Knapp say is the primary purpose for requiring licensed practice in any profession. Because, in a job analysis, decisions about what content material to include on a test are based on professional judgment of practitioners, Knapp and Knapp emphasize the importance of a selection process which ensures the credibility of the professionals chosen for job analysis committees.

Cowan (2007) lists four types of validity in the case of educational testing: content, construct, concurrent and predictive. Hibpshman (2004) references three types of test validity: content, criterion (which he defines similarly to Cowan's predictive validity) and construct validity. Knapp and Knapp (1995) state that for legal purposes, content validity is the most important validation for licensure testing, and they believe the job analysis process helps establish content validity. Hibpshman agrees that all types of validity need not be established for legal purposes. In fact, he writes:

> [t]he rules in 29 CFR [Code of Federal Regulations] seem to imply that all three types of evidence must be adduced in order to establish a test's validity, but in fact this view is neither accepted by the majority of practicing psychometricians nor required by the courts
>
> (p. 8).

At the same time, Knapp & Knapp acknowledge the inherent ambiguity in a job survey analysis when they write, "by describing a profession only in terms of data, people, and things, one may lose the essence of the profession and critical responsibilities and competencies may be overlooked" (p. 4). In fact, the purpose of a survey data analysis is to determine "which responsibilities, skills, or knowledges can be eliminated" from the list (p. 9), so that tests used for licensing only include the knowledge and skills determined to be "most critical to competent entry-level performance." (p. 10). It is also important to keep in mind the primary *purpose* of licensure tests is "to protect the public from harm" (Knapp and Knapp, p. 15).

Establishing cut scores. Another rather ambiguous process used in the series of inferences behind any standardized test score (Koretz, 2008) also relies on the professional judgment of practitioners is the setting of cut scores. A test's cut score "is simply the score that serves to classify the students whose score is below the cut score into one level and the students whose score is at or above the cut score into the next and higher level" (Bejar, 2008, p. 1). The establishment of a cut score implies that candidates scoring above an identified level on a continuum are proficient and those below a given level of test performance are not (Koretz, 2008). In the case of

Praxis tests, the scores represent a minimum "level of knowledge" (Nettles, Scatton, Steinberg & Tyler, 2011, p. 57) the state has decided candidates need in order to enter the teaching field or even to take courses in a teacher education program. Since the primary purpose of state regulation is to protect the public, by implication, candidates lacking this level of knowledge, as demonstrated by scoring below a given state's cut score on a Praxis exam, may harm the students they teach. Interestingly, cut scores vary from state to state (Understanding Your Praxis Scores, 2014-15). Hibpshman (2004) points out that sometimes the appearance of protecting the public is a political necessity. Samuel Livingston and Michael Zieky (1982), in answer to the question "How good is good enough?" (p. 12) add that "[a]ny standard…is based on some type of judgment," (p. 12). Based on their research, Hill, et al. (2012) call for a close examination of this process, just as they did for the job analysis. They write,

> …most methodologies to establish cut-scores do not seek to ascertain the relationship between cut-scores and actual evidence of on-the-job performance. We also note that across states there is wide variance in cut-scores for most Praxis tests and, within states, variance over time as policy makers react to changing teacher labor markets.
>
> (p. 5)

This logically leads to the question of how cut scores are established. This process is called standard setting, and, in an ETS document, Tannenbaum explains it is based on judgment (Tannenbaum, 2011). The process used by ETS for each of the dozens of content specific tests offered in the Praxis Series is the "modified Angoff"

method (p. 2) for multiple choice tests and an "extended Angoff" method (p. 2) for constructed response items. The Angoff method involves convening a group of 10-15 practitioners who meet once to establish cut scores for a specific test (Tannenbaum, 2011). Once a panel is nominated, ETS identifies "panelists who meet the criteria" (Nettles, et al., 2011, p. 58), and that list then goes to the state agency for approval. ETS has established that the "majority" (Nettles et al., p. 57) of the panel should be practicing educators and the panel should represent the diversity of the state in terms of gender, geography, and race and ethnicity (Nettles, et al., 2011). Tannenbaum adds that in certain subject areas it is sometimes difficult for ETS to convene even 10 educators. Such small numbers clearly have large implications in cases of "representative" diversity.

Once the group has been convened, each person takes the test and scores themselves. Next, they "define the knowledge and skills of minimally qualified test-takers," (Nettles, et al., 2011, p. 58) because, as Tannenbaum (2011) explains, "setting a standard is also setting a policy…about the type and amount of knowledge and skills that beginning teachers need to have" (p. 3). Bejar (2008) concurs. He writes:

> In short, standard setting matters: It is not simply a methodological procedure but rather an opportunity to incorporate educational policy into a state's assessment system. Ideally, the standard-setting process elicits educational policy and incorporates it into the test development process…
>
> (p. 2)

In the view of ETS, this process acts to "reconfirm the relevance (validity) of the test content for teachers in the adopting state" (Nettles, et al., p. 57). Presumably, then, after the job analysis process, if this small group of practitioners working to determine cut scores also determines the test content to be relevant, the content is considered valid.

Next, the panel receives "appropriate training" (Nettles, et al., p. 58) and "practice[s] making standard-setting judgments" (p. 58). Livingston and Zieke (1982) explain that the purpose of part of this training is to help panelists define characteristics of a "borderline" practitioner, someone presumably operating at a level between a proficient entry-level teacher and one who is not proficient, in the panelist's judgment. The process is an attempt for panelists to come to agreement on the kinds of knowledge and skills a borderline test-taker should be expected to possess (Livingston & Zieky).

In order to establish cut scores for multiple choice tests (Praxis I and most of Praxis II), the next step:

> necessitates that each panelist review each test item and *judge the percentage of a hypothetical group of 100 minimally qualified test takers who would answer the item correctly*. For each item, panelists record the percentage (e.g., 10%, 20%, …90%) of the 100 hypothetical test takers who they feel would answer the item correctly.
>
> (Praxis Technical Manual, 2010, P. 27, emphasis added)

Livingston and Zieke (1982) add that the easier the question, the higher the panel member is likely to estimate the percentage to be.

Once the percentages are estimated by each panelist for each test item, the estimates are first added and then divided to compute the mean score (Livingston and Zieke, 1982). This average represents the passing score study value (Praxis Technical Manual, 2010).

Six weeks after this cut-score study, a report is sent to the state verifying the participants, explaining the procedure, and sharing the results. ETS includes documentation about the standard error of measurement for the test and makes its recommendations about passing scores (Praxis Technical Manual, 2010). These recommendations are to be "within one and two standard errors of the panel's recommendation" (p. 28). States then decide on the "operational passing score" they will use on that particular test. (Praxis Technical Manual, 2010).

In the single state approach to standards setting, only one panel makes cut score recommendations, which Tannenbaum (2011) sees as problematic. With only one panel making cut score recommendations, Tannenbaum finds reliability to be an issue. Questions can be raised about whether a different panel of educators would set a different standard. Although it is accepted practice, he says there is no way to directly measure reliability with only one panel, so reliability is, instead, estimated using "standard error of judgment" (p. 1).

In recent years, some states, Kentucky among them, have participated in a multi-state approach to standard setting (Educational Professional Standards Board

Cutscore Framework Procedure, 2012). While the method of standard setting in a multi-state approach remains exactly the same, Tannenbaum (2013) cites two advantages. One is that it is easier to find enough qualified panel members in each testing area. Another is that more diverse perspectives are likely to be represented if panelists are chosen from a number of states. An additional advantage is that two panels can be formed, each representative of the participating states. This allows for increased reliability in the findings, as each panel's results can be shared (Tannenbaum, 2013).

Bejar (2008) concludes that "[f]ar from being a purely methodological process" (p. 4), standard setting is important because it establishes education policy. Tannenbaum (2013) agrees that standard setting process helps determine "the type of knowledge and skills that beginning teachers need to have" (p. 3). In other words, the test guides instructional content. This is one of Popham's (2014) stated goals for criterion referenced testing. These researchers and writers presume that teacher education programs will necessarily emphasize content which appears on the test. Because the standards set by the test are to reflect the content being taught in teacher education programs (Praxis Technical Manual, p. 57), one can infer that this multi-state approach to standard setting will, in turn, necessitate standardizing policy and teacher education program content both within and across states. As discussed earlier in this literature review, this move toward standardization, in fact, seems to be the intent of the new CAEP standards.

In light of the multiple roles for cut scores, Bejar (2008) raises an important question: "So, how do we know if the cut scores for a given assessment have been set appropriately?" (p. 1) His answer is that the "'right' cut scores should be both consistent with the intended educational policy *and* psychometrically sound" (pp. 1-2). It is important to bear in mind that the setting of cut scores is not a responsibility taken on by the testing industry itself (Praxis Technical Manual, 2010). ETS is involved during the cut score setting process, instructing the panel how to do its work, and it makes recommendations to states once the panel has completed the cut-score setting process (Praxis Technical Manual, 2010). Each state, however, bears legal responsibility for the cut scores it selects (Hibpshman, 2004). It is perhaps significant that the process for setting cut scores is outlined in the section of the Praxis Technical Manual in a section titled "Test Adoption Process," (p. 23), which is separate from the section that follows, titled "Psychometric Properties" (p. 29). The establishment of cut scores, therefore, is more related to policy and the decisions about the ways in which tests scores are *used* than to the psychometrics involved in the creation of the test itself. Hibpshman (2004) recommends that the best a state can do in establishing cut scores, described as a "complex matter involving legal, technical, political, and public relations considerations" (p. 20), is to "carefully document the chain of inference and rationale used to justify the selected level" (p. 20).

Koretz (2008) cautions policy makers and practitioners to keep in mind that "…the process of setting [cut scores] standards, while arcane and seemingly 'scientific,' is not a way of revealing some underlying truth about categories of

student achievement. The methods used are just a very complicated way of using judgment to decide which score is high enough to warrant the label "proficient" (p. 324). Kathleen Rhoades and George Madaus (2003) agree. They write, "[t]he decisions that underlie the formation of cut scores and passing scores are largely subjective" (p. 28). They cite Glass (1977) when they challenge the idea that cut scores can objectively separate students who know from students who don't. They write that this notion of objectivity is "largely a fantasy—there is no clear distinction and no mathematical or logical support for such an idea in the realm of education testing" (p. 28).

As a psychometrician, Koretz (2008) explains that because of the ways tests are designed and scores are reported, even in a test that is psychometrically sound, "…there are only trivial differences between students just above and just below a standard [cut score], and there can be huge differences among students who fall between two of the standards and who are therefore assigned the same label" (p. 324). Yet, because of a widespread *belief* in the infallibility of testing, policy makers either don't know or choose to ignore the inherent errors and inferences necessary at every step in the process (Rhoades & Madaus, 2003). The *presumption* of precision and impartiality recently assigned to quantitative data by policy makers, not psychometricians, has led to high-stakes decisions being made based upon what Koretz and Hibpshman (2004) concur are a series of far-from-perfect inferences.

Koretz (2008), then, raises yet another question, important because of its application to the ways in which Praxis exams are used in Kentucky. He writes, "if

you classify a student as being in one category (say, not proficient or proficient) based on one test score, how probable is it that you would reclassify the students as being in the other category if you tested her a second time?" (p. 159). This is a question of statistical reliability, which Cowan (2007) defines as "consistency of measurement" (p. 169). ETS says "…any estimate of a test taker's actual capabilities will contain some amount of error. Psychometrically, reliability may be defined as the proportion of the test score variance that is due to the 'true' (i.e., stable or non-random) abilities of the test takers" (Praxis Technical Manual, 2010, p. 39). Because test scores are only an estimate of these capabilities, there is known to be an "'error' component" in any set of test scores (Koretz, 2008). "Here, 'error' is defined as the difference between the observed and true scores" (Praxis Technical Manual, p. 39).

In psychometrics, it is well established that just as test scores are an estimate of a candidate's abilities, reliability, too, can only be estimated, due to the large number of variables, known and unknown. ETS acknowledges that "[s]ince true scores can never be known, the reliability of a set of test scores cannot be assessed directly, but only estimated" (p. 39). The question posed by Koretz about the consistency of results if a test-taker repeats a test leads to what is known as measurement error, which is used to estimate a test's reliability. It seems important for all educators to understand this system of measurement, often assumed to be scientific and therefore treated as if it is infallible enough to be relied upon to make decisions that have great impact on people's lives.

Measurement error. Daniel Koretz (2008) explains that in statistical terms, a clear understanding of reliability requires an understanding of measurement error, a statistic used to estimate the margin of error in a test score due to inconsistencies in the test itself. Put plainly, Koretz writes "…the 'margin of error' is a way of quantifying the degree to which we don't know what the hell we are talking about" (p. 144). That a degree of uncertainty, a margin of error, exists is simply understood in any discipline which relies on the use of statistics (Koretz, 2008). Koretz points out that this is not inaccuracy of any kind; it is, rather, inconsistency. Standard error of measurement (SEM) "quantifies the variability in a set of multiple measures of one person" (Koretz, p. 156).

Because statisticians know that a test score is never a completely true and accurate measure, they use SEM as a way to try to communicate the degree of statistical inaccuracy of a given instrument. SEM is an estimate of how far off a test score might actually be from a "true" score, which is inherently unknown. Some tests report scores within a score range as a way to communicate SEM. Praxis scores, however, are reported to test-takers as single scores. The only range of scores on the score report is simply the range of possible scores, which is a way to communicate the test's scale. The report also includes the average range for all test-takers, in both scaled and raw score formats, for that particular test. (Interpreting Your Praxis Score Report, 2009).

ETS defines Standard Error of Measurement as:

> …an estimate of the standard deviation of the distribution of observed scores around a theoretical true score. The SEM can be interpreted as an index of expected variation if the same test taker could be tested repeatedly on different forms of the same test without benefiting from practice or being hampered by fatigue
>
> Praxis Technical Manual, 2010, p. 39

Daniel Koretz (2010) explains that the type of SEM described by ETS in the quote above is actually one of three types of measurement error. He writes that this focus on variability on multiple forms of the same test gets "by far the most attention in technical psychometrics" (p. 49). This is also the focus of ETS in their definitions of classification accuracy: "the extent to which the decisions made on the basis of a test would agree with the decisions made from all possible forms of the test" (Praxis Technical Manual, 2010, p. 40) and classification consistency, which ETS defines as "the extent to which decisions made on the basis of one form of a test would agree with the decisions made on the basis of a parallel, alternate form of the test" (p. 40).

Another type of measurement error involves fluctuations in student performance if the same test is taken multiple times (Koretz, 2008). Koretz explains that even on a test that is statistically fairly reliable, any single score will fall somewhere within a *range* of accuracy, due to SEM. If a person takes a test many times (Koretz uses 500 times in his example), the measurement error decreases as scores begin to fall around an average, which would presumably indicate a "true" score. If a test taker has only one attempt at a test, the score will fall somewhere

within a range of probability. Statistically, this is to be expected, and this estimated range, representing testing inconsistency, is quantified and reported as SEM. Kortez explains that "…an examinee with any given true score, taking a test once, has a probability of about two-thirds of getting a score within the range from one SEM *below* that score to one SEM *above*, and a probability of one in three of obtaining a score more than one SEM away from the true score" (p. 155).

Further, Koretz (2008) says there is no guarantee that a single score will even fall within that score band, there is just a probability that it will. "This is not a problem specific to educational measurement; it is true of all statistical inference" (Koretz, p. 155). Coming at it another way, Koretz contends we would not, for example, want a single study to determine the effectiveness of a new medication, because of the number of possible variables involved. As testing is repeated, however, and more data accumulates about the effectiveness of a medication, an inference about its effectiveness becomes clearer. Yet in the case of educational testing, students are not taking the same test 500 times in order to solidify an inference about a score that can be made with confidence. A student's scores on the same test would vary each time the test was taken, due to variables ranging from the students' own personal conditions to variations in external conditions. When tests are taken multiple times, both kinds of measurement error, the variation expected between versions of the test and those conditional variables, come into play (Koretz; Au, 2013)). This is likely the reason many states, Kentucky included, allow

candidates to take a test multiple times, if they want to, or, significantly, if they can afford to.

A third type of measurement error Koretz (2008) identifies is inconsistencies in scoring. These come into play largely as human variation on human scored parts of a test, but also include mistakes made in scoring and reporting. A 2003 report by Kathleen Rhoades and George Maddaus, titled *Errors in Standardized Tests: A Systemic Problem,* finds human error to be "present in all phases of the testing process" (p. 28). The many errors identified in their survey over twenty-five years of testing "…offers testimony to counter the implausible demands of educational policy makers for a single, error-free, accurate, and valid test used with large groups…for purposes of sorting, selections, and trend-tracking" (p. 28).

Three approaches are used as attempts to quantify SEM. One is to report a range of uncertainty for the scores in general. For example, ETS reports the SEM for the test "Principles of Learning and Teaching: Grades K-6" as 7.4. That means there is a probability of 2/3 that a students' score would fall within a range of 7.4 points above or below what is called the "observed score" (Harvill, 1991). As Koretz (2008) points out, this also means that when a test is taken a single time, there is a 1/3 chance the score will fall outside that range.

ETS reports the reliability of the same test as 0.69. Koretz explains that this number represents the reliability coefficient, which is reported as a variance between 0-1, where 0 represents *all* error and 1 represents *no* measurement error, or perfect consistency. Unlike SEM, this measurement is not reliant on the scale of the test,

which allows the reliability coefficient to be compared from test to test. However, it is statistically difficult to understand for non-statisticians, so in most cases, Koretz doesn't think it is the best indicator of how much error there is on a given test. Koretz writes, "even when the reliability coefficient is high (.9 or above), as is the case with the SAT, "substantial measurement error remains" (p. 159).

The impact of these statistical errors, of course, is in their use. When a cut score is used to determine scores considered "passing" and "failing," the range of uncertainty in scores becomes of crucial importance. In those cases, Koretz writes, "…even a small margin of error will have serious consequences for students" (p. 157). He explains that if a students' "true" score is near the cut score, there is a "reasonably high probability" (p. 157) the student will be incorrectly accepted or rejected, simply due to measurement error. Of course,"[i]f scores on the test were used as only one piece of information contributing to the decision to admit or reject students, a modest amount of measurement error would have little impact" (p. 157). This is why the College Board recommends SAT scores be used as only one factor in college admission decisions, knowing that the scores are "approximate indicators" (p. 157, quotes in original) of a true score.

When used with a cut score, Koretz (2008) demonstrates the impact of reliability coefficients. His example might be illustrative of the 0.69 reliability score on the Praxis test used as an example above, since this number is consistent regardless of the test scale. Koretz adapts data from a study of CUNY's Testing Program conducted by Stephen P. Klein and Maria Orlando. He created a table that

shows that if the reliability coefficient of a test is .70, if a cut score is set in a way that allows for 50% of the test takers to pass, *26% of the test takers' scores would change with a second testing*. It is important to keep in mind that, when used in this way, the change in status might be from passing to failing or failing to passing for 26% of the candidates. With a cut score set at either a 30% pass rate or a 70% pass rate on a test with that .70 reliability coefficient, 22% of test takers' scores would change on a retake.

This statistical inconsistency is also an inherent part of testing, and is, in fact, referred to by ETS as classification accuracy and classification consistency. This inaccuracy is considered part of the SEM (Praxis Technical Manual, p. 40). In the Praxis Technical Manual, ETS explains:

> The estimated percentages of test takers correctly (classification accuracy) and consistently classified (classification consistency) tend to increase in value as the absolute value of the standardized difference (SSD) between the mean total score and the qualifying score increases. When the mean score of test takers is well above or below the qualifying score, the number of test takers scoring at or near the qualifying score is relatively small. Therefore, with fewer test takers in the region of the qualifying score, the number of test takers that could easily be misclassified decreases and the decision reliability statistics reflect that fact by increasing in value.
>
> (Praxis Technical Manual, 2010, p. 40).

Hibpshman concurs with ETS that cut scores set at either extreme in the distribution of test scores have fewer misclassifications than cut scores set near the middle. This information is, of course, relevant to the setting of cut scores, which, ultimately, is the responsibility of each state.

In Kentucky where, as in many other states, Praxis scores are used to determine whether or not a student can enter a teacher education program *and* then whether the candidate will gain certification, the stakes are high. Because test scores are used in ways that determine an individual's future, stakeholders often assume these test results must surely be *completely* valid and reliable. Instead, Hibpshman (2004) writes:

> …false positives and false negatives are unavoidable, and an effort to minimize one will usually result in an increase in the other. A false negative denies an individual who is otherwise capable of teaching the right to do so; a false positive places an unqualified person in the classroom (p. 10).

In a footnote, Hibpshman reveals that "many experts in both education and psychometrics view false positives as the more serious error in the case of teacher tests…" (p. 10). In other words, it's seen as a bigger problem for unqualified candidates (as determined by their test scores) to be allowed to practice teaching than for completely competent candidates to be barred from entry to their chosen profession based on a series of mistakes in inference, which are expected and accepted as inevitable.

A new issue is also arising around misunderstandings about the capabilities of tests and cut scores. In a report for CAEP, the national accrediting body for teacher education, authors Edward Crowe with Michael Allen and Charles Coble write, "the number of times a given candidate must take a test in order to meet the state cut score is relevant to questions of candidate and program quality" (p. 47). Given the inconsistencies inherent in these multiple aspects of testing, can this correlation be inferred in any valid way? Continuing, the authors take an accusatory tone in the next sentence when they write, "Some states and programs have taken steps to obscure this fact…" (p. 47), referring to the fact that some candidates have taken the tests multiple times. Clearly, based on all the evidence above, the more times a candidate takes a test, the more likely we are to be able to infer what a "true" score might be. Yet these authors, in advising an accrediting body likely comprised of few if any psychometricians, advise that national accreditation policy be set in ways that consider those retakes reflective of poor program quality.

A closer look at how Praxis Exams are **used.** Daniel Koretz (2008) writes, "One widespread unreasonable expectation is that a test created for one purpose will do just fine for many others. But a single test cannot serve all masters. Remember: a test is a small sample of a large domain" (p. 327). ETS has stated that Praxis exams cannot validly be used as a predictor of teaching success or a measure of teaching ability (Praxis Study Companion, 2013). What inferences, however, are implied in the ways in which Praxis Exams are used in Kentucky? How has the validity of these uses been established?

Researchers Goldhaber and Hansen (2009) point out the rationale for the use of exams as gatekeepers, as Kentucky's uses Praxis I to determine eligibility for program entry and Praxis II to determine eligibility for certification. They write: "At the heart of any licensure test requirement is the exclusion of individuals from the pool of potential employees: Individuals not passing the test are believed to be of unacceptably low quality and are thus deemed ineligible to teach" (p.4). By any standard, this is a high-stakes decision, with consequences for both individuals and programs.

The series of assumptions on which this policy is based is that the job analysis process has accurately determined what teachers should know and be able to do for future classroom success, that the multiple choice test created from this information is an accurate measure of these skills and this content, *and* that the cut scores set by the state are capable of making a fair determination about who is likely to be a "high quality teacher" and who is not. On top of those assumptions, we add measurement error to the mix. When all is said and done, Hibpshman (2004), Koretz (2008) and others contend that standardized test scores are, at best, close approximations of the knowledge and skills being tested. Some stakeholders might ask: Are approximations good enough in determining a person's future?

Even if the Praxis *could* accurately measure what it claims to, the "demonstrated knowledge believed to be important for safe and effective entry-level practice" (Praxis Technical Manual, 2010, p. 10), has research established a clear correlation between mastery of this knowledge and teaching ability? How important

is it that a given body of knowledge is in place *at a given time*—the lack of which disqualifies students from even *beginning* certification courses? By whose authority can the state regulate student entrance to university and college courses, even at private institutions?

In "Teaching Subject Matter," a chapter in *Preparing Teachers for a Changing World* (Hammond, L & Bransford, J., 2005), Pamela Grossman, Alan Schoenfeld and Carol Lee write,:

> To argue that teachers need to know the subject matter they teach seems almost tautological, for how can we teach what we do not understand ourselves? Yet the links between content knowledge and teaching performance are not all that easy to document.
>
> (p. 205)

Terry Hibpshman (2004) agrees when he writes: "It is widely believed that academic proficiency and content knowledge are essential in determining who will be a good teacher, but in fact research studies provide at best weak support for this idea" (p. 8).

Grossman, Shoenfeld and Lee (2005) posit a possible reason, an inference based on their research: "The kind of content knowledge that supports good teaching may, in some cases, be different in kind than that generally acquired by individuals who pursue a college major in a content field" (p. 206). They cite several studies that exemplify this view. This seems an important area for further research, especially given that college students are being barred from even attempting the coursework that would lead to a career in education based on the approximation of their content

knowledge demonstrated by Praxis I. It seems difficult to argue a valid use for a test when a "weak at best" (Hibpshman, 2004, p. 8) correlation is established between that which the test is designed to measure and what policy makers hope to determine, which is, presumably, potential for high quality teaching.

One might wonder why a policy decision would be made to use a test as a gatekeeper without a firm base of research in which to ground such a decision. Hibpshman (2004) addresses this when he states that the public perception of the quality of teachers in the U.S. is low, now likely based at least in part on the CCSSO report (2012) stating that teachers have poor academic skills. The truth of this perception, Hibpshman says, "is arguable" (p. 9), but the public perception is "important and inescapable" (p. 9). It seems, then, that the attempt to promote standardized testing as an objective, "scientific" way to monitor the academic and pedagogical skills of teachers is, at least to some extent, political. Dana Wakefield (2003) traces the widespread use of standardized tests in teacher education to the passage of the 1998 Higher Education Act, which established a new requirement that every teacher education program compile and submit an annual Title II report to the federal government. Quantitative data are often much easier to report in ways that meet Title II requirements, and Wakefield notes that ETS had the Praxis exam ready and waiting, in anticipation of this Congressional Act, and it has been in widespread use ever since.

Testing experts and psychometricians, including those at ETS, make it quite clear that standardized test scores should not be used as the sole factor in making

high-stakes decisions (Praxis Technical manual, 2010). This is predicated on the psychometric understanding that "validity in modern terms represents a chain of inference" (Hibpshman, 2004, p. 8). Harvard professor and testing expert Daniel Koretz (2008) says unequivocally, "…one of the best ways to avoid misusing test data: don't treat any single test as providing the "right," authoritative answer. Ever" (p. 320).

While it is true that once a student has met the Kentucky cut scores on the Praxis exams, other data points are taken into consideration to determine if a student should gain program entry or certification. If a student does not pass the Praxis exams, however, no alternative information is accepted as an equivalent demonstration of the knowledge and skills the Praxis is supposed to measure. The impact of these regulations, then, is that Praxis serves as a solitary gatekeeper at *two points* in teacher educator preparation, barring hopeful students at each gate. A student who doesn't make the cut scores on each subsection of each of these two tests (a total of five subsections on Praxis II for elementary candidates, for example), is ineligible for certification. The clear implication is that they do not have the necessary knowledge and skills to become successful teachers.

Inferences: Praxis exams and teacher quality. Researchers Heather Hill, et al. (2012) and found few studies that even attempt to compare results of standardized assessments like the Praxis Exams to instructional quality in classroom practice. Terry Hibpshman (2004) concurs. This, too, calls into question Kentucky's use of these Praxis test scores. Further, Koretz (2008) observes that because of the public

assumption that test score data are a true, unbiased measure of student learning, "many people consider test-based accountability systems to be self-evaluating" (331), leading to "a disturbing lack of good evaluations of these systems, even after more than three decades of high-stakes testing" (p. 331). In other words, as policy makers and politicians increasingly require research-based accountability in educational systems at all levels, are the policies they set firmly based in research? Are they measuring and assessing policy outcomes? Are they calling for regulation of the testing industry itself?

The National Board on Educational Testing and Public Policy raises this question in a 2003 report by Rhoades and Madaus titled "Errors in Standardized Tests: A Systemic Problem." They find that there is no outside monitor of the testing industry at any level, public or private. They describe the testing industry as a "closed system…exempt from independent examinations" (p. 8). Jay Rosner (2012) demonstrates this closed system as he describes paying ETS for the data he needed for the research he wanted to pursue, and their subsequent refusal to allow him access to the additional information he requested once his initial findings were published. Rhoades and Madaus quote a 1991 report by the National Commission on Testing and Public Policy stating that this lack of oversight means there is little to no "consumer protection" (p. 8) for standardized test takers.

Some researchers are raising questions about what the testing industry considers acceptable practice. Hill, et al. (2012), for example, question the job analysis process used to justify the content validity of tests. Because of how teacher

exams are used, their implied purpose is as a predictor of classroom success. Hill, et al., conclude, therefore, that in order to be valid, there must be a correlation between scores on Praxis Exams and teacher success in classroom practice. If exam results are used in ways that prevent entry into the profession or even to beginning coursework in the field, as they are in Kentucky, it seems an important relationship to establish, yet the authors found no indication that the Praxis Series had been validated in any way as a measure of classroom practice. ETS itself clearly states that Praxis is not intended for use in predicting future classroom success (Praxis Study Companion, 2013, p. 35).

If the knowledge and skills measured by Praxis tests are not predictors of future classroom success, why would these scores be used to restrict entry into the field at two points in a candidates' journey to becoming a teacher? ETS itself says Praxis I can be used to determine program entry (Purpose of Standardized Tests, 2014), as long as the state sets the cut scores, so it seems logical to assume that someone has determined that candidates who fail these tests would not be good future teachers. Upon what information is this assumption based? Hill et al., (2012) cite a 2005 review of studies conducted by Wilson and Young which focused on the relationship between scores on the National Teacher Exam (the precursor to the Praxis Series) and teacher quality. None of the six studies found *any* relationship between candidate test scores and multiple measures of teacher success, including student outcomes. Hill et al. also cite studies in 2004 by Angrist and Guryon and in 2007 by Goldhaber which found that some teacher candidates with low test scores

had high levels of student outcomes and vice versa; some teacher candidates with high test scores had poor student outcomes. Hibpshman (2004) seems to agree when he writes, “When teacher basic skills test scores have been used as predictors of teacher performance, few studies have shown any strong relationship, and some studies have shown no relationship at all” (p. 8).

Is it possible to create a test with predictive power? Hill et al. (2012) find that previous studies suggest no or very limited correlates between content knowledge and high quality teaching as measured by multiple choice tests. The authors express concern that the use of Praxis Exams is not predicated on studies that show how these scores “generalize to practice” (p. 514). Studies showing this correspondence are needed to ensure that the variable preventing a potential teacher from practice truly is an indicator of a candidate who is not prepared. They wonder if it would be possible for a multiple choice test to show this correlation which, arguably, should be the goal of any instrument used as a gatekeeper.

For their study, therefore, Hill, et al. (2012) chose a multiple choice assessment they felt was better aligned than Praxis with the kinds of mathematical skills and knowledge needed for effective teaching. Perhaps they were looking for the shift in knowledge and understanding Grossman, et al. referred to as “different in kind” (2005, p. 206) when embedded in classroom practice. The test they chose, the Mathematical Knowledge for Teaching (MKT), was created from a “grounded study of practice” (p. 513) conducted by Ball and Hill in 2008, rather than the job analysis procedure used by ETS.

Hill, et al. (2012) write, “unlike all teacher certification assessments available today, we attempt to discern the validity of teacher scores on this assessment vis-à-vis other key criteria [of teacher quality]” (pp. 493-4). They chose three data points for comparison: scores on the MKT, scores on an “observational instrument that quantifies the mathematical quality of instruction” (p. 494), and value added scores from student assessments on a state test of mathematical knowledge.

Despite the use of a multiple-choice test shown to be correlated to practice, “the predictive power was still imprecise” (p. 513). This finding gave Hill et al. (2012) the confidence to state that “it is not likely that commercially available assessments would do better in terms of identifying teacher candidates who are unprepared to enter teaching” (p. 514). During the course of their study, the researchers identified many important aspects of teaching that could only be discerned through observation. The authors concluded nonetheless that it still might be *possible* to create a multiple-choice exam that could predict teaching success in mathematics, but in order to do so, much more research of this kind would be needed. Hill et al. clearly state, however, that their study demonstrates that the current multiple choice tests, used in ways that infer they are predictors of teacher quality, don’t measure up, and further, that any teacher exam used as a gatekeeper should be grounded in research that demonstrates predictive validity of teaching success.

Koretz (2008), too, believes the end goal should be to establish “an effective system of accountability, one that maximizes real gains and minimizes bogus gains

and other negative side-effects…All that we have seen so far tells us that the simple test-based accountability systems we use now do not meet this standard" (p. 330).

It seems policy makers in Kentucky believe they have created an effective, multi-faceted system of accountability, as described in a 2013 report by Terry Hibpshman. With Praxis Exams standing guard at two gates, however, there is a clearly predictable *adverse impact* (Koretz, 2008, p. 265) on certain groups of students. Given that the set of inferences gained from Praxis scores are based on processes Koretz describes as "arcane and seemingly 'scientific'" (p. 324), and given that few if any studies have linked Praxis exam scores with a prediction of future classroom success, someone somewhere must have decided the benefits are worth the known costs. Yet do we really know what the benefits are? Despite the many imperfections and inferences behind the meaning of Praxis test scores, policy makers are using information from these tests as if they can assess what *appears* to be teacher quality. The focus has remained steady on the *intent* of the policies, which is to ensure high quality teachers in every classroom. It's important to remember that the outcomes of these policies impact people's lives, as well as teacher education programs and the institutions that house them. For this reason, it is important to analyze the *impact* of these policies, before readily accepting that good intentions are "good enough." In the case of the use of Praxis Exam scores, the outcomes are clear.

Impact of Praxis Exams as Gatekeeper in Teacher Education.

Multiple variables and the scientific method. It is well documented and widely acknowledged that white students systematically outperform students of color

on standardized tests commonly used for college admission (Mitchell, Robinson, Plake & Knowles, ed., 2001), and standardized tests in general (Rosner, December 17, 2013, personal correspondence). This gap disparity is true for the Praxis Exams as well (Nettles, Scatton, Steinberg & Tyler, 2011). In fact, an analysis in a report for ETS by Nettles, et al. "…revealed very large score gaps between African American and White teacher candidates on selected Praxis I and selected Praxis II tests" (p. 47). The authors acknowledge that these gaps are comparable and in some cases larger than those found on the SAT and GRE tests. A large gap is defined by statisticians as a mean difference greater than .80, or 8/10 of a standard deviation unit (Nettles et al.). To demonstrate the size of one standard deviation, Koretz explains that on a standardized scale, with a mean score set at 0, a cut score one standard deviation unit above the mean would eliminate all but 16% of the total scores. 8/10 of that, therefore, represents a large difference (Koretz, 2008). To further clarify, Koretz explains that "a mean difference of .80 standard deviation, which would be among the smaller of the score differences commonly found, would place the median African American student…the student who would outscore half of all black students…in only the twenty-first percentile among whites" (p. 99). Koretz goes on to say that most studies of score differences between African Americans and Whites on standardized tests have found mean differences between 0.80 and 1.1 standard deviation. This large difference, of course, has serious implications in the setting of cut scores (Hibpshman, 2004), especially when those scores are used as the sole indicator used to exclude individuals who score below that mark (Koretz, 2004).

Nettles, et al. (2011) reported differences in passing rates between African American students and White students, both taking Praxis I for the first time. Cut scores are different in different states, but a look at passing rates is a way to discern the impact of cut scores. The discrepancy in outcomes is significant. On the Reading portion of the test, for example, 81.5% of Whites passed, compared to 40.7% of African Americans (Nettles, et al., p. 9). 79.5% of White students passed the Writing Exam, while the passing rate for African Americans was 44.2%. In Math, 78.2% of White students passed, compared to 36.8% of African American students. Nettles et. al. found that results for Praxis II showed the "range of standardized differences was from 0.74….to 1.41…all gaps on the selected Praxis II tests were considered to be large (0.80 and above)" (p. 26) with one exception, where the gap "bordered on being large" (p. 26). Flynn Ross (2005) also notes high failure rates on Praxis Exams in her work with recent immigrants and refugees in a teacher certification program in Portland, Maine, and Angrist and Guryan (2007) and Bennett, McWhorter & Kuykendall (2006) describe similar Praxis score differences between Whites and Hispanic and Latino Americans. Ross summarizes the findings of many studies demonstrating similar score gaps which result in high failure rates on the Praxis Exam for minority teacher candidates in general.

Other systemic demographic score divides exist in standardized testing as well. Diane Ravitch (2010) summarizes the findings of many researchers who demonstrate correlates between test scores and income levels, a divide Koretz (2008) says is "universally acknowledged" (p. 127). Ravitch writes, "Unfortunately, every

testing program—be it the SAT, the ACT, NAEP, or state scores—shows a tight correlation between family income and scores. Children from affluent families have the highest scores, and children from poverty have the lowest scores" (p. 286). She reports that, like the gaps between racial and ethnic groups, these scores gaps are statistically large, and they are found on both national and international tests. Koretz explains that Ravitch is referring to socioeconomic status (SES), a statistic often used in such studies. SES is an amalgam of information including not only income, but also the educational level reached by a student's parents, and parent occupational status. Each of these variables, taken independently, have been shown to "strongly predict" (p. 127) test scores. In other words, a test-takers' ethnic background and/or socio-economic status can be used to predict what that students' test score will be.

Correlation does not necessarily indicate causation, however, and Wakefield (2003) points out the difficulties inherent in distinguishing between entangled variables in human populations, as evidenced by the direct correspondence between low SES status and minority status in the United States. The research conducted for ETS by Nettles et al. (2011) was an attempt to tease out some of these many variables. Their findings? Even accounting for other factors, the team still found "very large score gaps" (p. 47) between African American and White teacher candidates on every Praxis I and the Praxis II test they studied. In fact, they found that for both Praxis I and II, "race/ethnicity explained the most variance in the scale scores" (p. 35) in comparison to other demographic factors. Their findings and the research of others, such as Bennett, McWhorter and Kuykendall (2006), show that

while it is true that income levels correspond with test results in ways that are predictive, it is also true that ethnic identities are also strong predictors of test scores.

The point is that *each* of these demographic variables negatively impact test scores, and perhaps they even work in tandem. Yet despite the many possible reasons underlying these test scores discrepancies, the fact of the matter is that test scores are often communicated as if they are valid measures of student learning. Score gaps are therefore *assumed* to be indicative of content knowledge and skills potential teacher candidates are lacking, making those candidates seem unqualified to teach.

Koretz (2008) explains how *un*scientific the leap to such an assumption is. He says that when a scientist gets a result that seems to match a particular hypothesis, (such as low test scores indicate low levels of knowledge and skills), those data are not simply accepted as confirmation of the hypothesis. Instead, *always*, the next step must be to ask what other explanations for that result are possible. The cause might also be found in a confounding variable. Clearly, among human populations, there are always *many* confounding variables. Koretz explains that a truly scientific model would require further tests, each designed for the purpose of singling out and controlling for each of the many variables, one at a time, before an explanation for the result is ever accepted as credible. In other words, to positively claim that a poor test result is caused by low levels of knowledge and skills, all other possible factors would need to be ruled out.

Therefore, while scores on Praxis I may be an indicator of a student's content knowledge, it may also be an indicator of many other things. Koretz adds, "Everyone

who studies educational achievement knows that differences in scores arise in substantial part from noneducational factors. A huge body of research collected in the United States over half a century documents this…" (p. 115).

The assumption that standardized tests are objective and can therefore be used to accurately measure each student's achievement, makes it seem as though each student then has an equal opportunity to succeed. This is based on the notion that America operates as a meritocracy in which every person has an equal chance at success, however success is defined (Au, 2013). Citing Berliner (2012), Wayne Au (2013) finds, however, that about 20% of test score variance is due to the effects of schooling, while about 60% involves many other variables, none of which have to do with schools. Au observes that if a truly objective test is given in what is actually a meritocratic society, these large discrepancies in test scores should not exist. Au further concludes that the stated goal of policy makers, to close the so-called achievement gap, means that everyone would do well on the tests—which would supposedly indicate they all had done well in school. This is impossible, however, because of the way standardized tests themselves are created and scaled. The creation of a bell curve necessitates establishing a range of scores. Therefore, it will always appear as if some test scores indicate success and others, failure (Au 2013).

A further problem is that it is difficult, if not impossible, to control for the many entangled variables that exist. Koretz (2008) writes, it is "…unarguable that social factors have a very substantial impact on test scores…" (p. 116), a common sense notion which he says is quite clear to most casual observers. Even still, "[a]s

obvious as it is, the impact of noneducational factors on test scores is widely ignored" (p. 117) by many groups: politicians who use score results as a simple way to assign blame; policy makers who react, at least in part, to the political climate of the times (Hibpshman, 2004); the press, who report the assertions of both these groups; industries who stand to benefit; and often by educators who are pressured to accept this message, even while they see a different reality play out in the day to day life of their classrooms (Koretz, 2008).

Hibpshman (2004, 2013) and Koretz (2008) agree that test results are based on a series of inferences which have to be gleaned through a fog of many, many complicating factors, any or all of which might also be interpreted as a reason for a given test result. Arguments therefore arise over the *degree* to which social factors impact scores on standardized tests, while it is unlikely they will ever be untangled enough that their effects will ever be "proven." Based on the results of regression models that find "race/ethnicity" (Nettles, et. al., p. 23) to be "significant predictors of Praxis I score performance" (p. 23), Nettles, et. al (2011) conclude that their findings "confirm what has been traditionally observed, that the accumulation of human capital as represented by various background characteristics is related to higher test performance" (p. 23). This impact, whatever the cause, carries serious implications.

Outcomes and implications. Whatever the cause of this large discrepancy in test scores, the *impact* of these systematically low passing rates for minority students and students with low socio-economic status has the effect of, in the words of Rona Flippo (2003), "canceling diversity" (p. 42) in the teaching profession. This, Flippo

says, is "the real crisis" (p. 42). Flippo writes, "…rather than being successful [in their goal of raising teacher quality], the only true accomplishment of these testing programs has been to gradually cancel out the minority teaching force in the United States, while at the same time the number of minority and language diverse children has been increasing dramatically in our schools" (p. 42). Joshua Angrist and Jonathan Guryan (2007) conducted a study to try to determine the impact of testing on teacher quality, and because they found "no evidence of a corresponding increase in quality" with increases in test scores (Abstract, p. 1), they described their findings as "reasonably consistent with the view that testing has acted more as a barrier to entry than a quality screen" (p. 18).

Assumptions about score gaps and other possibilities. Standardized testing is often viewed as a colorblind system—objective and equitable. Yet Wells (2014) points out that many policies considered to be "colorblind" are often far from colorblind in their outcomes. This is certainly the case with standardized testing. Especially as these tests are being used to determine who will be "allowed" to become a teacher, the questioning of this assumption becomes an issue of vital importance.

Fully aware of score disparities between groups of test-takers, decisions to use Praxis Exams to establish baselines for program entry and certification must assume that gaps in test scores indicate a so-called "achievement gap" between white students and students of color. Nettles et al. (2011), for example, make this assumption in their report for ETS as they explain that the large differences in test scores and passing

rates "are but two indicators of African American Praxis test-takers' under-achievement that need to be addressed" (p. 47). There is little doubt, as Ross (2005) points out, that students who themselves have not had access to excellent teachers in their schooling careers are at a distinct disadvantage when their competency is measured by generalized standardized tests, setting up a perpetual cycle of restricted access. Therefore, when these tests are used as solitary gateways for entry into TEP, and then again before a teacher can be certified, it follows that a larger percentage of minority candidates (as well as students with lower SES) will be barred entry to the teaching profession.

Since 1965, a primary goal of educational policy has been explained as attempts to close this so-called achievement gap (Jennings,2000 in Au, 2013), a gap which has, in fact, been *defined* by standardized test scores. A question seldom raised, however, is what if even a *part* of this perceived gap is, instead, caused by other societal factors, or caused by something in the creation, interpretation, and use of the tests themselves? If test score interpretations are actually derived from a series of inferences, what if, at any point, something else might be inferred?

Stephen J. Gould's 1981 book *The Mismeasure of Man* cautioned readers to interpret standardized test results carefully. He explains that no social construct, as the development of psychometric procedures for large-scale standardized testing certainly is, occurs in a vacuum. Context, and especially social and political context, often plays an important role. Yet when simple answers to complex questions are sought, and presumably found, people sometimes forget to examine the source of

these so-called answers. Given that demographic differences in testing are based on the social construct of race, it seems all the more important to examine the roots of the use of standardized testing in U.S. education.

Rooted in bitter soil. It doesn't take much digging to find the roots of educational standardized testing, so it's interesting that they have not been exposed more often, especially among educators. In an article in *Educational Administrator* (2000) titled "Predictable Losers in Testing Schemes," Peter Sacks describes his participation on a panel gathered to discuss standardized testing. Sacks wrote of his surprise when a journalist asked if these new tests would negatively impact minority children. In his view, the answer to that question has been very clear for a very long time. He writes, "The losers in high-stakes testing schemes always have been children of the poor, the working class and undereducated. And the winners always have been children of the privileged, well educated and the affluent" (p. 1). Sacks continued to say that standardized tests in the U.S. *have always been used to divide the nation along racial and socio-economic lines*.

Further digging shows the roots of standardized testing in U.S. education to be entangled with the roots of the eugenics movement in the United States. The U.S. study of eugenics is another history that has been fairly well concealed. In a piece published on George Mason University's History News Network website, Edwin Black (2003), claims that the so-called science of eugenics picked up steam in the United States before spreading to Germany, where the work was taken up by Hitler and the Nazi party. In fact, much of the first eugenics research was conducted in the

United States, funded by the Rockefeller and Carnegie foundations (Black, E., 2003). Furthermore, links between the educational uses of standardized testing and the eugenics movement are well-established (Au, 2013, Black, 2003, Stoskopf, 2002).

In his 2002 article, "Echoes of a Forgotten Past: Eugenics, Testing, and Education Reform," Alan Stoskopf introduces Henry Goddard, an educational psychologist who ran New Jersey's Vinland Training Center for Feebleminded Boys and Girls. Goddard's interest in eugenics led him to concur that the children he worked with were genetically prone to crime, poverty and any number of societal ills. Goddard decided the standardized I.Q. test developed by Alfred Binet, a test which, although modified, is still used today, would be a perfect way for him to determine which people were, in fact, "feeble-minded." One of his studies involved recent immigrants to the U.S. The results of these tests were used to "prove" that around 80% of Jews, Hungarians, Russians, and Italians had IQ's lower than a "normal" 12 year old, leading Goddard to conclude these immigrants were mentally deficient (Stoskopf, 2002). That conclusion, however, didn't stay with Goddard alone. In 1941, the New York City school system hired Goddard and used this test to help them track their students into specialized classes and schools. Further, at his center for the "feebleminded," Goddard trained 1,000 teachers to give these tests and interpret the results (Stoskopf, 2002), the outcome of which must have impacted the lives of tens of thousands of students and their families.

Stoskopf (2002) introduces another educational psychologist, Lewis Terman, a faculty member at Stanford's College of Education from 1910-1922 who was also a

eugenicist. Terman developed standardized intelligence and achievement tests and conducted long-term studies on their use and results. Stoskopf cites Gould (1996) and Brigham (1923) when he writes of the "blatant class, cultural and ethnic bias" (p. 128) inherent in the test questions Terman first developed for use with U.S. Army recruits in 1917. During the time Terman was conducting these tests with soldiers, U.S. schools were increasingly criticized as inefficient in comparison to industry, which was booming at the time. Terman saw the perfect opportunity to promote the use of his standardized tests in education, which were clearly more efficient than other methods of assessment. Backed by his dean at Stanford, Terman began establishing scoring scales for his tests, based on a study of close to a thousand middle class, native-born Protestant children of European ancestry, which he used as a norming group (Stoskopf, 2002).

As a result of the scales developed based on this norming group, Terman identified huge numbers of mental deficiencies among peoples of color, leading him to recommend they be separated from their peers and put into classes where instruction would be "concrete and practical" (as cited in Stoskopf, 2002, p. 129). Stoskopf further explains that in Terman's view, these groups should also be kept from "their prolific breeding" (p. 129, Terman, in Stoskopf, 2002).

Just as Terman had hoped, standardized testing was praised for its efficiency, allowing Terman to continue his work in the field of education. By 1925, many city school districts were using Terman's test results as a way to track their students. In fact, Terman became so well respected in the field that in 1922, he was elected

president of the American Psychological Association and went on to become editor of *six* journals of educational research (Stoskopf, 2002). Stoskopf (2002) writes, "However flawed the methodology might have been in constructing some tests and interpreting scores, they increasingly played a major role in determining the academic fate of school children--and the fates were unequal" (pp. 129-130.) Citing his own book about testing inequality and Stephen J. Gould's *The Mismeasure of Man*, Wayne Au (2013) writes:

> Through the work of these psychologists, and with the explicit support of educational philanthropists like Carnegie (Karier, 1972), IQ in the United States became conceived of as hereditary and fixed, laying the groundwork to use standardized (sic) testing to justify the sorting and ranking of different people by race, ethnicity, gender, and class according to supposedly inborn, biologically innate intelligence (p. 8).

In fact, Au explains, by 1920, 400,000 copies of Terman's National Intelligence Tests had been sold to schools across the nation. In 1922, Terman helped create the Stanford Achievement Test which, by 1925, had sold 1.5 million copies. Also in 1925, 64% of 215 U.S. cities surveyed were using intelligence test results as a way to track their students (Au, 2013).

Stoskopf also writes about Edward Thorndike, an educational researcher who had worked with Terman in administering the Army tests and then on the norm-referencing of Terman's new tests, designed for educational use. Thorndike taught and conducted his research at Columbia Teacher's College from 1899 until his death

in 1949, writing fifty books and hundreds of articles (Stoskopf, 2002). According to Stoskopf, Thorndike "institutionalized the myth" (p. 130) that standardized tests confirm truth. Like Terman, Thorndike agreed that society would be better off with policies for "selective breeding" (p. 130, Thorndike in Stoskopf, 2002). Thorndike also promoted the idea that the educational policy should be taken out of the hands of teachers and administrators and left to the so-called educational experts, a call that was soon echoed by many district superintendents (Stoskopf, 2002). The "science" of eugenics was clearly catching on. Citing Craven (1988), Stoskopf (2002) points out that in 1928, less than 100 years ago, over 300 eugenics courses were taught in colleges and universities around the U.S. What legacy is carried forward as a result of those courses, which must have impacted the thinking of thousands of students? How many of these college graduates became the so-called experts called upon to set educational policy?

On a webpage titled "Americans Instrumental in Establishing Standardized Tests" a companion page to an episode from the PBS Frontline series called *Secrets of the SAT*, Carl Brigham is identified as the developer of the SAT. The source goes on to say that Brigham became interested in intelligence testing as a student at Princeton, where he later taught. In1923, Brigham published a study based on those first standardized tests used by the U.S. Army, which had been developed by Terman and Thorndike. Also a strong advocate of eugenics, Brigham analyzed the Army's test results in terms of race. Accepting those results as "truth" and citing his findings as evidence, he wrote a book titled *A Study of American Intelligence*, in which he

concluded that as long as races continued to mix in America, the educational system would decline. In his continued work on standardized testing, Brigham adapted the Army Alpha test for use by colleges and universities, and *renamed it the Scholastic Aptitude Test*, or SAT (Americans Instrumental, n.d.). The SAT was first adopted for educational use as an admissions test by Harvard's president, James Bryant Conant. Its use quickly spread. Later in his career, Conant organized the Educational Testing Service (ETS), a non-profit agency that merged all the companies in the standardized testing industry into one organization. Interestingly, according to the PBS website, Brigham ended up renouncing many of his own findings, concluding that there is no such thing as a universal measure of intelligence. As a result, Brigham opposed the establishment of ETS and the use of the SAT he developed, a test still in use by U.S. colleges and universities.

Jay Soares (2012) explains further:

> When the SAT was introduced in 1926 it was supposed to be an IQ test that would measure intrinsic intellectual aptitude, not academic subject mastery; it was supposed to help sort between the gems in the Nordic race from the subject-test grinds in the "Jewish race". It did not work to exclude Jews, but other tactics introduced in the 1930s of requiring mother's maiden name and place of birth, were more effective toward that goal. It also did not work to predict grades. Yale and Princeton knew that as early as 1930 (Soares, 2007). But the private sector clung to the test, first for the invidious distinction over public universities of requiring a nationally normed measuring stick, later

> *because of the convenient way it disguised SES selection as academic selection,* paying the bills along the way. The lasting legacy was a pseudo-IQ test that sorted students by family income, opening or closing doors to colleges and careers in the process (p. 9, emphasis added)

From the beginning, college presidents, researchers, educational policy"experts," and some educators accepted the results of these tests as objective measures of intelligence and achievement. Many people of color, however, have long been aware of standardized testing's legacy and the myth that America operates as a meritocracy (Au, 2013). Despite the racist and classist roots of standardized testing, the assumption of objectivity is alive and thriving, driving the accountability reform movements in U.S. education.

The legacy of racist roots. Given the origins of standardized testing in U.S. education, rooted as it is in the U.S. eugenics movement of the early 20th century, and given the broad use of standardized testing in the educational accountability movement of today, it would be reasonable to assume that much must have changed—that the psychometrics developed in the context of eugenics must surely have evolved in ways that would rule out discrepancies of race and class. Yet clearly, class and race are both still accurate predictors of test scores. In this context, Wayne Au (2013) writes:

> The historical roots of high-stakes, standardized testing in racism, nativism, and eugenics raises a critical question: why is it that, now over 100 years after the first standardized tests were administered in the United States, we have

> virtually the same test-based achievement gaps along the lines of race and economic class? (p. 12).

The racist roots of standardized testing might be of little interest if not for that question and a few critical connections to the processes inherent in the development, scaling and interpretation of standardized tests. Even though the use of tests today may not come from racist or classist *intent*, it is less than one-hundred years since such separation *was* the intent, and, hauntingly, the *impact* has not changed over time. So while it is well documented and widely accepted that a large gap exists between whites and people of color, reasons attributed to these differences in outcomes remain sources of study and debate. One possibility will require an even closer look at particular aspects of test development process.

Differential Item Functioning and testing's self-perpetuating system. Jay Rosner is the executive director of the Princeton Review Foundation. His work involves research focused on standardized testing and supporting minority students in passing standardized tests that work against them because of what Kidder and Rosner (2002) term "built-in headwinds" (p. 131).

Rosner's (2012) research has largely focused on the SAT test. In a December, 2013 email correspondence, he wrote that the correlation between "bubble tests" is so high that one standardized test score is generally a reliable predictor of another. For example, an SAT, ACT or GRE score might be used to predict a Praxis score, and vice versa. In fact, these comparisons are sometimes used in research as a way to establish the validity of a test score. For example, in a Policy Information Report for

ETS, Drew Gitomer (2007) demonstrated the strong relationship between SAT scores and Praxis Exams. Gitomer compared Praxis results with SAT scores to make the case that Praxis was, indeed, filtering out candidates with lower academic ability by showing that as Praxis cut scores were raised, the SAT scores of passing candidates also increased. The intent was to demonstrate that the academic abilities of test takers were increasing as test scores increased. The underlying assumption in this case, of course, was that SAT scores are accurate measures of academic ability, an assumption important to keep in mind while reading about Rosner's findings.

In a chapter titled "The SAT: Quantifying the Unfairness Behind the Bubbles," in *SAT Wars: The Case for Test-Optional College Admissions,* Jay Rosner (2012) asks readers to consider the possibility that at least part of the consistent score gaps between Whites and students of color lies in the tests themselves. This seems a critical consideration, given that "the differences in average scores among racial/ethnic groups on the teachers licensure tests…are generally similar to the differences found among these groups on other tests" (Mitchell, K.J., Robinson, D.Z., Plake, B.S. & Knowles, K.T.,2001, p. 99). Clearly, with such consistent score disparities even among tests, questions about the underlying reasons beg an explanation. Test item selection is one of Rosner's areas of expertise, and in this chapter, he illustrates the demographic *impact* of the test item selection process used by ETS.

The ETS website titled "How Tests and Test Questions are Developed" (2014) states that "dozens of professionals — including test specialists, test

reviewers, editors, teachers and specialists in the subject or skill being tested — are involved in developing every test question, or 'test item,'" (par. 2) in keeping with high standards of "quality and fairness in the testing industry" (par. 2). Multiple steps in the process are listed, and they include: defining criteria, putting together committees to define the content domain, writing and reviewing test questions, pre-testing, finding and deleting unfair questions, putting the test together, and then, once the test is in actual use, checking to be sure the test is functioning the way it was intended.

In an attempt to ensure fairness in each test item, ETS uses a process known as Differential Item Functioning (DIF). DIF is defined this way in the ETS Glossary of Standardized Testing Terms:

> Differential item functioning (DIF) is the tendency of a test question to be more difficult (or easy) for certain specified groups of test takers, after controlling for the overall ability of the groups. It is possible to perform a DIF analysis for any two groups of test takers, but the groups of test takers ETS is particularly concerned about are female test takers and test takers from specified ethnic groups. ETS refers to those groups as "focal groups." For each focal group, there is a corresponding "reference group" of test takers who are not members of the focal group. A DIF analysis asks, "If we compare focal-group and reference-group test takers *of the same overall ability* (as indicated by their performance on the full test), are any test questions

> significantly harder for one group than for the other?" (Differential item functioning, ETS, 2014, emphasis in the original).

ETS collects demographic information from each test-taker which allows them to analyze the performance of different demographic groups on each test item. Daniel Koretz (20008) explains that when a high level of DIF is found in a test item, it could be the result of many factors, only one of which is bias in the test question. In these cases, however, the question is most often thrown out, in an attempt to err on the side of fairness.

In a chapter from a book titled *Differential Item Functioning*, William Angoff (1993), of ETS, cites many other authors when defining item and test bias in terms for the purposes of statistics. Angoff writes, "[a]n item is biased if equally able (or proficient) individuals, from different groups, do not have equal probabilities of answering the item correctly" (p. 4). This has become known as differential item analysis (DIF). Clearly, then, Angoff explains, ETS needs some way to group test-takers by "ability" in order to see if groups of "equal ability" have answered a given question correctly. In a different chapter for the same book, Nancy Cole (1993), also of ETS, writes:

> In practice, of course, we must operationalize 'ability.' As technicians we all know that; we understand its necessity…We face the confounded problem of going from the public connotation of the word *ability* to the technical theoretical meaning and then to the operational meaning we give it. In so doing, in spite of the very central public concern that tests may not adequately

> reflect ability, we have for good reasons reduced the very word *ability* to mean operationally 'test score.'
>
> (p. 27, emphasis in the original).

For the purposes of DIF analysis, then, test takers are grouped by ability according to their scores on the test overall. In fact, Cole infers that, because an operational definition is necessary for statistical purposes, technicians and psychometricians accept the assumption that test scores *are* ability. This discrepancy between the operational definition and the public connotation of the word "ability" is crucial information for educators.

A statistical technique used for making DIF comparisons is called biserial correlation. The ETS glossary defines biserial correlation as:

> [a] statistic used at ETS to describe the relationship between performance on a single test item and on the full test. It is an estimate of the correlation between the test score and an unobservable variable assumed to determine performance on the item and assumed to have a normal distribution (the familiar "bell curve")
>
> (Biserial correlation, 2014, Glossary of Standardized Testing Terms)

This, then, is the statistical method used to determine DIF. Whether "ability" is accurately measured by the overall test score, this is an assumption made for statistical purposes. The result is that the quality of each test item is measured against the results of the test itself. The Pearson *r,* which represents the Pearson product-moment correlation coefficient, is the statistical symbol used to demonstrate the

strength of the relationship between two variables (Cowan, 2007). Seema Varma (n.d.) explains that for the purposes of test item selection, a "good" test item would show a positive correlation, meaning that a student who did well on the test overall got the test item in question right, and a person who did poorly on the test also did poorly on that particular test item. This is expressed as a variance of +1 (Cowan, 2007). A negative correlation would indicate the opposite, so that if a person scores poorly on the test overall but gets that question right (or vice versa), there is a negative variance, expressed as -1 (Cowan, 2007). Varma explains that point bi-serial correlation is a "special type of correlation between a dichotomous variable (the multiple-choice item score which is right or wrong, 0 or 1) and a continuous variable (the total score on the test ranging from 0 to the maximum number of multiple-choice items on the test" (p. 3).

Angoff (1993) adds:

> The process of matching the two groups for ability, although certainly a preferred procedure, nevertheless does raise, in the public's mind, some interesting questions of its own. For example, when there already exists a marked difference between the Black and White groups, what does it mean to 'match'?...Does the analysis now have the intended value? In reply to this concern we can say that we have extracted the ability differences between the groups and have laid bare for comparison the remaining differences, those having to do with the students' color, culture, and group identity. In the last analysis, this is the essence of the comparison we wish to make.

(p. 17-18)

The *intent* of DIF, then is fairness and, as Bennett et al. (2006) point out, cultural neutrality. Jay Rosner explains the resulting *impact* of this process. While acknowledging its complexity, Rosner's (2012) results lead him to question the fairness and acceptability of the outcome of DIF: the items which are ultimately selected for inclusion on the test.

Rosner's (2012) study focused on the SAT test, and he begins by explaining that every SAT test contains a set of questions that are in a pre-testing phase of development. This section of the test will not be included in the test-taker's overall score, and the test-taker has no idea which section is being tested. This is how ETS field tests items being considered for inclusion on future versions of the test. ETS also collects demographic data on each test taker, which can then be associated with each test item, allowing ETS psychometricians to study the *demographic outcome* for each individual test item as it comes out of its pre-testing phase. Rosner claims that, although these data exist, they are rarely made available to people outside of ETS. Rosner managed to purchase two years of these item level data, which became the basis for his study. Rosner explains that ETS has since refused to release more of this information and, since there is no regulatory oversight on the testing industry in the U.S., ETS retains that right of refusal.

Rosner (2012) was interested in breaking down the test items one at a time, looking for possible clues as to why these demographic gaps in scores have remained steady "for decades" (p. 106). Using the math test as his first example, and examining

two years of SAT results as his data set, Rosner was interested in examining the 35 point gap between males and females. To accomplish this, he attached one of three labels to each test question: male, female, or neutral. A question labeled "male" meant that more males than females answered the question correctly. A "female" question meant more females than males chose the correct answer, and a neutral question meant equal numbers of males and females answered the questions correctly.

His findings? Out of 117 math questions analyzed, only one was answered correctly by more women than men. It is important to keep in mind that these were questions *selected* and *used* on the SAT. In other words, despite, or perhaps because of the way in which the DIF analysis is performed, using bi-serial correlation, the questions selected for use on the Math section of the test gave a clear advantage to males taking the test (Rosner, 2012). In the context of maintaining bi-serial correlation, this might make sense. Since females have *historically* performed less well overall on the test, and individual test questions are selected in order to maintain a correlation between performance on individual items and overall test performance, it stands to reason that the questions selected would perpetuate this status quo.

Using the same two years of data and the same methodology, Rosner (2012) then analyzed individual SAT questions in order to compare outcomes for test-takers who self-identified as Mexican American and those who identified as White. This time, he studied items from both the math and verbal sections of the SAT. Of 276 questions analyzed from both test sections, *one* question was "neutral," or answered correctly by equal percentages of White and Mexican American test-takers, and *274*

were answered correctly by higher percentages of people identifying as White. In other words, only one question selected for inclusion on the SAT was answered correctly more often by Mexican American test-takers than White test-takers.

The third analysis Rosner (2012) conducted compared results of a demographic breakdown of individual questions comparing correct answers of African American and White test-takers in both the math and verbal sections of the SAT, based on information from the same two years of data. This time, Rosner found that *every single one* of the 276 questions selected for use on both the math and verbal sections of the SAT were more often answered correctly by Whites than by African-Americans. There were no neutral questions. As an example, Rosner did use a question that had been answered correctly by African Americans more often than by Whites, but that question came from a pre-test and it had been rejected for inclusion in a final version of the SAT.

Clearly, no matter what the intent, the *impact* of the methods used for test item selection *favors White test-takers* (and for the math section, test-takers who are both male and White.) Throughout the article, Rosner raises questions of fairness. In order to emphasize the issue of fairness, he asks readers to consider reactions if the opposite of these findings were true. What if data showed that SAT test questions were selected in a way that resulted in each question selected favoring females over males, or Mexican Americans or African Americans over Whites? Would test results still be so widely accepted as a fair and unbiased (Rosner, 2012)? Consider the magnified impact of outcomes for African American or Mexican American women taking the

math section of the SAT, where the selected test items favor outcomes for both Whites and males.

How can this happen? As explained earlier, because of the statistical methodologies used in large-scale standardized testing, Rosner (2012) writes, "the profile of the answering cohort for each individual question should parallel the answering cohort of the test overall" (p. 115), creating a self-perpetuating system. *Given the historic roots of U.S. educational testing in the eugenics movement, this raises critical questions.* Are the standardized test results of the eugenics era, less than 100 years ago, still being perpetuated? Rosner's interpretation of the data he analyzed lead him to conclude that, in order to maintain the profile required by point bi-serial correlation, "99% of SAT math questions chosen to appear on scored sections have to be answered correctly by a higher percentage of males than females, a higher percentage of whites than Mexican Americans, *and* a higher percentage of whites than blacks, *simultaneously"* (p. 115, emphasis in the original). Digging a little deeper into the history of point bi-serial correlation, it seems that this statistical technique and other psychometric frameworks *also* originated in the context of eugenics studies.

***The roots of statistical models for standardized testing*.** Sir Francis Galton, creator of the correlation coefficient was, interestingly, a cousin of Charles Darwin (Brutlag, 2007) and was "a man motivated by strong eugenic views" (Norton, 1978, p. 9). From 1850-52, he explored Africa, following his interest in heredity and "possible improvement of the Human Race" (Brutlag, 2007). His field observations

led him to believe that human intelligence followed a normal distribution curve, scaled on the group he termed Anglo-Saxons (Jensen, 2002). On this scale, Galton's observations suggested to him "that Africans were 'two grades' below Anglo-Saxons, a difference equivalent to 1.33 standard deviations…"(Jensen, 2002, p. 149). Galton also "concluded that men were higher than women in general ability" (Jensen, p. 149).

Based on his belief that intelligence was an inherited trait, Galton's goal was to create a statistical procedure to demonstrate that trends in intelligence correlated from one generation to the next (Brutlag, 2007). Based on the normal curve, Galton developed statistical models, first in the context of sweet peas, then of physical human characteristics, and finally applying them to human intelligence (Brutlag, 2007). Jensen (2002) calls Galton "the father of psychometrics, the measurement of quantitative social behaviors" (p. 148). Psychometrics is the field of study used in the creation and scoring of standardized testing today. Jensen, citing Burt, 1962 and Stigler, 1986, writes,

> Before the invention of inferential statistics…Galton provided psychometrics with some of its most fundamental measurement tools and descriptive statistics, and all of these are still in use. He invented the measures of bivariate correlation and regression (further developed by Karl Pearson), the use of percentile scores for measuring relative standing on various measurements, and the use of the Gaussian or normal curve as a means for scaling variables

> that theoretically are normally distributed but cannot be directly measured on an interval scale or a true ratio scale.
>
> (Jensen, 2002)

Jensen (2002) adds that because Galton decided to accept that general ability follows a normal distribution, this allowed for him to convert measures of rank order to percentiles, which could then be interpreted in terms of normal deviates, or standard deviation. Daniel Koretz (2008) posits that the bell curve (what Jensen refers to as normal distribution) frequently occurs in nature, and has useful mathematical properties that allow for descriptions of a given distribution. However, Koretz writes,"[t]he fact is that we don't really know what the 'true' distribution of reading or mathematics achievement really should look like, and we can design tests to change the shape of the distribution" (p. 79), so that scale scores generally follow a bell curve, which makes them easier to interpret. But what information is lost or distorted in the quest for ease of interpretation?

Karl Pearson, who worked with Galton, created the mathematical framework for correlation coefficients and other statistical models (Brutlag, 2007). "Widely regarded as the founder of the modern discipline of statistics," (Norton, 1978, p.1), Pearson was also a leading promoter of social Darwinism and eugenics (Norton, 1978). In 1893, Pearson's first paper introduced the term standard deviation, as Pearson too, assumed sample populations would match normal distributions (Brutlag, 2007). Pearson became director of both a "Biometric Laboratory" and the "Galton Laboratory for National Eugenics," which he and Galton established in 1906 (Norton,

1978). Norton defines biometry as the application of mathematics to the study of variation in human populations. Norton explains that statistical methods and models were developed in Pearson's Biometric lab and then applied in the Eugenics lab. These labs attracted many students who went on to hold posts and produce papers that greatly influenced many fields. To summarize Pearson's influence, and therefore the link between eugenics and the development of statistical models for use in the social and psychological sciences, Norton (1978) writes, "Certainly, in Pearson's time, statistics was always associated with eugenics, and, more generally, was strongly promoted as a mathematical methodology that was capable of elevating several disciplines—for instance, psychology, anthropology, sociology and craniometry—into truly scientific ones" (Norton, 1978, p. 5). Norton quotes Pearson to explain his purpose for developing these statistical models:

> The purpose of the mathematical theory of statistics is to deal with the relationship between 2 or more variable quantities without assuming that one is a single-valued mathematical function of the rest. The statistician does not think a certain x will produce a single-valued y; not a causative relation but a correlation…somewhere within a zone…Our treatment will fit all the vagueness of biology, sociology, etc. A very wide science.
>
> (Pearson in Norton, p. 10)

The complexities and vagueness of these social fields, then, were represented through statistical models developed *in order to demonstrate the superiority of Anglo*

Saxons, on whom they were normed. These trends continue to be represented in test results today.

Adverse impact. Rosner (2012) reports that the response of ETS to his analysis of their data is simply that the DIF process they've used since 1983 ensures fairness in test item selection (Schmitt & Dorans, 1988). Despite this process, Rosner found that on the SAT overall, women score about 35 fewer points than men on the Math section, Latino/as score about 70 points lower than whites on both Math and Reading sections and African Americans scores around 100 points less than whites on both Math and Reading tests created for the SAT. The failure rates on Praxis Exams, as shown earlier, reflect similarly large discrepancies.

Daniel Koretz (2008) is careful to explain that a difference in scores among groups does not necessarily indicate test bias. He explains that a test can be psychometrically sound and still produce discrepant scores. Rather than going ever deeper into the psychometrics of standardized testing, looking at questions of bias in design and use, it might be best to analyze what Koretz calls *adverse impact,* an issue outside the realm of psychometrics. Adverse impact means, simply, "a group has been harmed by testing" (p. 265). Clearly, this is the case when tests with large group discrepancies in scores are used in ways that impact people's lives, such as when scoring below a cut score prevents people from being permitted to pursue a chosen career. Koretz explains that adverse impact can occur without test bias in any technical sense, because it is used more as a legal term.

Regardless of intent or the presumed meaning of the inferences represented by test scores, these score discrepancies clearly have an adverse impact on students of color and students from lower SES. The outcome of the use of these scores is that they disproportionately bar entrance for some groups and advantage other groups.

Further, because of the way tests are designed and scaled to a bell curve, this adverse impact for lower scoring groups increases exponentially due to what Koretz (2008) calls the "Berkeley effect" (p. 268), which "can be expected even in the total absence of bias" (p. 268). As a standard (such as a cut score) is raised to become more selective, lower scoring groups will be increasingly underrepresented, according to Kortez, and higher scoring groups will become statistically overrepresented, exacerbating the already negative effect of systematic score differences. This occurs because the creation of a bell curve "bunches up" (p. 268) a group's scores close to the mean score for that group. When this happens, Koretz explains, the phenomenon he has named the Berkeley effect is both a "mathematical certainty" (p. 268) and "very powerful" (p. 268).

To illustrate this effect, Koretz (2008) shares an example of what happens when a cut score on a test is used to allow or deny admission to a given program, in much the way Kentucky uses Praxis cut scores to allow or deny students access to program entry and/or teacher certification. Koretz reminds readers the Berkeley effect is one reason a test score alone should never be used to allow or deny admission, because "the results are dramatic" (p. 268).

For the sake of example, if there were no set cut score for test-takers, groups would be represented in proportion to their representation in the population. Scores would provide some inferential information, but would not be used in ways that allow or deny access. Setting cut scores as if test scores can be used to establish proficiency or lack thereof is a choice; one that Koretz says results in large adverse impacts.

If a cut score is set at the mean score for all test-takers, for example, the representation of groups with scores lower than the mean is, then, disproportionately lower, because the establishment of a bell curve means scores for that group will be bunched up around their group mean below the cut score. As you move a cut score up the scale from the overall mean, the representation of groups with lower mean scores drops disproportionately, while an overrepresentation of the higher scoring group is admitted (Koretz, 2008). In terms of the large or very large discrepancies between the Praxis scores of whites and minority students found by Nettles, et al.(2011), the use of cut scores and the Berkeley effect must have an impact, compounding the underrepresentation of peoples of color.

A Cut Score Framework Procedure for Kentucky, adopted in 2012, is outlined in a document titled "Internal Procedures" for the Education Professional Standards Board (EPSB) (Cut Score Framework Procedure, 2012). It states that now that Kentucky teachers have participated in every multi-state standard setting session, the policy is that Kentucky will use the recommended scores resulting from those sessions "if they equate to the 25th percentile. If the cut scores are below the 25th

percentile, the scores could be raised" (p. E-12. Internal Procedures, Approved, 2012).

In a 2013 report for CAEP, authors Edward Crowe, with Michael Allen and Charles Coble, cite three main problems with testing systems used with teacher candidates. They write that, in general, cut scores on teacher tests are set too low "to ensure that those who pass have the content and professional knowledge to be effective classroom teachers" (p. 12). Interestingly, the very next bullet point says, "The tests themselves have little demonstrable relationship to the knowledge, skills, and teaching performance required in today's schools" (p. 12). How do these two statements go together? Does the second not negate the first? Nevertheless, later in the document, the authors advise CAEP to make a move to improve teacher quality by substantially raising cut scores within the current testing system "until better tests…are in place" (p. 15). They recommend that to gain national accreditation, a program's students must surpass

> uniform national passing cut scores set at the 75th percentile for all test takers in the nation. Setting a high bar at this level would ensure that only the strongest candidates would be allowed to enter the profession. An alternative to the 75th percentile is to set passing cut scores one standard deviation above the mean for all national test-takers. For a normal distribution of test scores values, this would be about the 68th percentile. In effect, the 68th percentile standard means that candidates would have to score in the top third of all test-takers. (p. 16).

In 2000, Gitomer and Latham's research for ETS discussed what they called a "myopic perspective" (p. 6) that raising cut scores equated to "raising the bar." They call these "deceptively simple solutions" (p. 6) which ignore "the complexity of the issues embedded in teacher reform" (p. 6). They write:

> If the highest passing scores currently used in any one state were implemented across all states, fewer than half the candidates would pass Praxis I, and fewer than two thirds would pass Praxis II. Without other interventions, the supply of minority candidates would be hit the most severely. For example, only 17% of the African American candidates would pass Praxis I, and just one third would pass Praxis II. The dramatic effects that would be brought about by raising passing standards require very careful policy analysis.
>
> (p. 6)

Do Crowe, et al. (2013) know the well documented, easily predictable impact their suggestion of a 75th percentile pass rate would have on candidates of color, who already struggle to meet current cut scores? If they don't know, why not? The impact on peoples of color would surely be more dramatic than Gitomer and Latham's (2000) illustration. Because of the increase in the recommended cut score levels, the impact of the Berkeley effect (Koretz, 2008) would be far greater. If Crowe, et al., do know the predicted impact, are the results acceptable to them, in the name of *appearing* to raise the bar for teacher standards? The authors add a caveat to presumably raise standards for teacher education programs, by suggesting that "[a]t least 80% of all program graduates would have to pass all relevant tests at this 75th

percentile passing score" (p. 16). CAEP has adopted the suggestion that programs must show an 80% pass rate (CAEP, 2014) on the tests they use, but as of 2014, CAEP has not required nationwide adoption of a particular test, nor agreed to set standard cut scores nation-wide

Why is such disparate impact *allowed* to happen? Given all the possibilities for inaccurate inferences from test results and the potential for adverse impacts amplified by what Koretz (2008) calls the Berkeley effect, why are Praxis scores used as if they actually determine which students do or do not have the prerequisite knowledge and skills to enter a teacher education program (or even take the coursework), and later, to become certified? In an attempt to explain, Koretz (2008) refers to a common saying: "the perfect is the enemy of the good" (p. 118). Koretz says this might be acceptable reasoning if "flawed data provide a relatively correct answer. But it is very bad advice if the flawed data suggest fundamentally misleading answers" (p. 118). The fact that test data are necessarily "flawed" is the reason psychometricians advise against using test scores alone in making high-stakes decisions. The fact that ETS declares it valid to use Praxis I to determine admission to teacher education programs (Praxis Technical Manual, 2010), from which policy makers might infer Praxis I results to be predictive of future teaching success, further confuses the issue.

Justification from the judicial branch. With such systematic discrepancies in test results across racial, ethnic and socio-economic lines, one might wonder about the legal ramifications of using test results to screen teacher candidates. In a 2004

report to advise Kentucky policy makers regarding the setting of cut scores, Terry Hibpshman writes, "…while the use of tests is a legitimate and popular mechanism for selecting candidates for certification, the question of how tests may best be used for this purpose ultimately has legal implications" (p. 3). Hibpshman warns that setting cut scores is an "arbitrary matter," (p. 4) and although no legal statutes regulate the use of any standardized test, there is legal precedent that acts as a guide. Hibpshman (2004) points specifically to the 14th Amendment, which has been interpreted to include "the right to practice one's chosen profession" (p, 4), but he encourages policy makers to note that those terms "do not allow an individual to bring action merely because a state policy is unfair: it must deprive them of either liberty or property" (p. 4).

Most important in its application to discrepant test results among groups, Hibpshman explains, is the Code of Federal Regulations, Title 29, which contains what is known as the "four-fifths rule" in 1607.4.D. This Code reads:

> A selection rate for any race, sex, or ethnic group which is less than four-fifths…(or eighty percent) of the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, although a less than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.
>
> (in Hibpshman, 2004, p. 5)

There it is: bias written right into law. Applied to teacher certification or admission to program, the four-fifths rule means that as long as a policy allows

admission to minorities at a rate of at least 80% of that of White students, the policy is legally considered to be unbiased. Interestingly, this 20% discrepancy, which works to advantage white candidates on Praxis and in the application of any other policy, is not described as a racial advantage. Instead, the difference has been declared to be legally acceptable. Therefore, when Praxis tests are used as gatekeepers for program entry and certification, it is perceived as fair if minority candidates pass at a rate 20% lower than that of white candidates. Hibpshman (2004) goes on to say that not only is the acceptance of such disparity written into law, but in legal cases, "the courts usually give test administrators a great deal *more* latitude than the regulation would seem to grant" (p. 5, emphasis added). This may be the reason cut scores that produce score discrepancies even greater than 0.80 have remained in place.

This is clearly an example of what john a. powell (2012) means when he describes court rulings that uphold laws even when their impact is discriminatory. In *Racing to Justice: Transforming Our Conceptions of Self and Other to Build an Inclusive Society*, powell introduces the term "racialization." He writes, "[b]y racialization, I refer to the set of practices, cultural norms, and institutional arrangements that both reflect and help to create and maintain race-based outcomes in society" (p. 4). Powell contends that "racialization is a set of historical and cultural processes" (p. 4) that help maintain the "racial status quo" (p. 6). Clearly, standardized testing is one of these processes.

To demonstrate one means used to keep such processes in place, powell (2012) cites 426 U.S. at 245-246, the1976 *Washington v Davis* ruling "which sets out

the Supreme Court's discriminatory purpose doctrine, requiring that a plaintiff prove *intent* in racial discrimination claims" (p. 5, emphasis added). This ruling set a precedent that in order for any claim of discrimination to be upheld, the person bringing the claim has to prove that discrimination was intended. Intent, of course, is an internal process and therefore exceedingly hard to prove.

In order to explain how this ruling applies in cases regarding the disparate impact in the use of standardized test scores to select teacher candidates, Hibpshman (2004) confirms:

> Equal protection cases are especially difficult to prove. 29 CFR established the four-fifths rule as a presumptive test of disparate impact, but courts have ruled that disparate impact by itself does not constitute discrimination, and require…additional proofs. In addition to disparate impact cases, some cases have been brought on the basis of *discriminatory intent,* the idea that a state policy intentionally discriminates against some protected group. Such cases are very difficult to prove. When plaintiffs have been successful, it has usually been when it was shown that the state knowingly ignored the advice of experts who had warned of the possibly discriminatory consequences of adopting some policy
>
> (p. 13, emphasis in the original)

Powell (2012) explains that this ruling is a result of what he calls the color-blind stance of the Supreme Court, which is regarded as neutral when issues of race are considered. However, powell points out that racial neutrality would only be

neutral in its effects if all groups had been allowed equal opportunity throughout our country's history (powell). Yet despite a history of white advantage, racial discrimination, and clearly unequal policy outcomes, powell (2012) says that the Court's stance is that to be "fair" is to ignore racial disparities as if racism had never played a role in establishing a social hierarchy in our country. To view this as a neutral position is to ignore the fact that in the U.S, white people have historically been advantaged in every way, and these advantages have been passed down from generation to generation (powell).

High stakes uses for standardized test scores are a case in point. White people, and particularly middle and upper class white people, are still clearly advantaged when standardized tests are used for any high-stakes purpose. Current use of these tests now directly impacts the education of every child in America, as well as the selection of their teachers. "Fairness is not advanced by treating those who are situated differently as if they were the same" (p. 9), powell continues. "Although a policy neutral in design is not necessarily neutral in effect, the courts and the public seem all but obsessed with the design and, even more narrowly, with the *intent* of the design, rather than the results" (p. 9, emphasis in the original).

As Hibpshman points out, proving discriminatory intent in the use of Praxis Exam results would be almost impossible, especially since the new CAEP standards explicitly *require* programs of teacher education to document their efforts to recruit candidates of color (CAEP, 2013), as if the lack of such efforts were the source of the problem. Discriminatory *impact* as a result of testing policies, however, is clearly

accepted. In fact, the law *expects and accepts* discriminatory impact, as is clear by the existence of the "four-fifths" rule.

Hibpshman (2004) further explains that 29 CFR *seems* to require two things: "that a test be anchored in some reasonable expectation about job performance, and that there not be evidence of disparate impact" (p. 5). 1607.5.H. adds, "Where cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force" (in Hibpshman, p. 5).

Defining those "normal expectations of acceptable proficiency" is the role played by the job analysis described earlier. Yet Hibpshman (2004) acknowledges that in "countless analyses by psychometricians" (p. 8), "criterion-related validity is virtually impossible to establish…because adequate measures of performance are difficult to obtain and …it is difficult to arrive at an operational definition of good performance" (p. 8). Clearly, definitions of good performance vary among individuals according to their philosophical stances.

In addition, it is difficult to find ways to truly measure performance quality, which can only be based on observable behaviors (Hibpshman, 2004). Because of these difficulties, while 29 CFR would *seem* to require a test to predict job performance, Hibpshman explains that in practice even that has not been required by the courts. Perhaps this is why in the Praxis Technical Manual (2010), ETS carefully explains that Praxis is not a predictor of job performance, yet also claims the tests can validly be used at program entrance and as a requirement for certification. These

claims must be based on what Hibpshman explains as the modern, rather than the traditional, definition of test validity. In modern terms, because of the complexity of the processes created to try to establish test validity, validity is now "taken as a chain of inference about the usability of a test for specific purposes" (p. 9).

Because little evidence exists to establish a relationship between test scores and teacher performance, Hibpshman (2004) explains how difficult it is to establish cut scores that make it appear that such a relationship exists (Hibpshman). The Angoff Procedure, as described earlier, was created for the purpose of defining performance standards, and its use is accepted by the courts as evidence that a minimally accepted criteria has been established (Hibpshman). This is one of the assumptions upon which a link in the chain of inference relies. Verifying powell's (2012) point about the legal system's focus on intent alone, Hibpshman writes, "[c]ourts will often accept test programs with weak validity studies if they are convinced that the developer and administrator made a good-faith effort to establish validity and have a coherent plan for remediating the deficiencies of the validity studies" (Hibpshman, p. 12).

In fact, Mitchell, et.al (2001) write, "court decisions have been inconsistent about whether the Civil Rights Act applies to teacher licensing tests. In two of three cases in which teacher testing programs were challenged on Title VII grounds, the court upheld use of the tests…ruling that the evidence of the relevance of test content was meaningful and valid…and that valid alternatives with less disparate impacts were not available" (p. 112).

The court's acceptance of the series of processes created to try to establish the relationship on which criterion-related validity is based, a process Hibpshman describes as "virtually impossible" (p. 8), makes it *appear* the tests can ensure that state policy is, indeed, serving its intended purpose of raising standards of teacher quality. Koretz (2008) carefully explains that no test can really do that. Yet policy makers have decided to use standardized tests for this purpose anyway, and clearly, the courts will allow it. This decision, Hibpshman explains, "is more political than technical" (p. 9). He writes, "[p]ublic perception is not a validity issue in a technical sense, but it must be considered in any test development effort. However well founded formal validity studies and cut score procedures happen to be, they will usually be unconvincing if the public perceives teacher quality to be a serious problem and the cut scores too lenient" (p. 9).

This statement seems to be borne out even in the 2013 Title II Report, in which states are criticized for setting cut scores below the overall average test score (U.S. Dept. of Education). This criticism comes apparently despite the known outcome of raising cut scores for candidates of color. This criticism reinforces the recommendations of Crowe, et al. (2013) to CAEP that, in order to appear as if it is raising the bar of teacher quality, CAEP should set cut scores at the 75th percentile. The question must be raised: Do current policy makers and advisers find the known outcomes for teachers of color acceptable? These outcomes affect every U.S. citizen, and in many ways, the future of our nation.

Dehumanization, Collateral Damage and the Real Crisis.

There are clear implications in the decision to use Praxis I cut scores as a barrier to program entry and of Praxis II as a gatekeeper to certification in Kentucky: higher percentages of people of color and people from low SEC households will be excluded from even the possibility of becoming teachers. Some would argue that candidates scoring below the selected cut scores should be excluded from candidacy test scores because those scores indicate "unacceptably" low levels of the knowledge and skills needed for successful teaching. From the analysis above, however, it is clear that this assumption is based on standardized test score data and uses of those data that truly *are* based on a long series of inferences. And while these inferences and their outcomes have been accepted by the courts, are they ethical? Are they equitable? Should they be accepted by the American people?

Every policy must be based on a cost-benefit analysis. The use of test scores for political purposes, to give the *appearance* that this will raise teacher quality, comes at great cost. The resulting exclusion of peoples of color from becoming teachers is what Flippo (2003) calls "the real crisis" (p. 42). Flippo names impacts on individuals, teacher education programs, and the future students these educators of color might have taught, as stakeholder who will be negatively impacted in this crisis. She writes that this use of testing in teacher education has been in place long enough to clearly see results and to predict future impact. She writes,

> …the only true accomplishment of these testing programs has been to gradually cancel out the minority teaching force in the United States, while at

> the same time the number of minority and language diverse children has been increasing dramatically in our schools"
>
> (p. 42).

Rather than regarding this impact as a reason to back away from using Praxis Exams as sole gatekeeper at multiple points of entry, however, Flippo (2003) points out that instead, these policies have spread nation-wide, as increasing numbers of states have set regulations mandating the use of these tests. This has occurred, Flippo says, despite the many, many teacher educators and researchers who continue to question the costs inherent in the result: the maintenance of what Christine Sleeter (2001) calls the overwhelming presence of whiteness in the teaching population.

Weighing the costs. Although policy decisions have been made, the public and even many educators may not be fully aware of the costs and the potential impact on the future of our nation. In his book *Racing to Justice: Transforming Our Conceptions of Self and Other to Build an Inclusive Society*, john a. powell (2012) provides a historical perspective of U.S. law and current policy that establishes and maintains the social and economic hierarchies in the United States. While powell does not mention the use of high-stakes standardized testing specifically, its use in U.S. education seems to exemplify many of the points he makes. Powell's book gives a context for understanding how such racially disparate outcomes are still accepted in our nation's institutions. Further, powell elucidates the benefits for the future if racial justice truly were to be prioritized as a national goal in the United States.

The outcomes of standardized testing are not a surprise. They are well documented and have apparently been accepted by education policy makers. In 2003, Wakefield wrote, "[m]inority and low-income teacher candidates are among the casualties of high-stakes testing's rise from comparative obscurity to federal policy. High-stakes tests, with the same ethnic and socioeconomic shortcomings as the SAT, guard the door to the teaching profession…" (p. 380). In 2004, Hibpshman informed Kentucky policy makers that "[it is an unfortunate but undeniable fact that minorities as a group tend to perform less well on standardized tests than do whites…the lower relative performance of minorities presents unavoidable complications in the process of setting cut scores" (p. 10). He backed this statement with evidence in data tables of teacher certification exam results on which the mean scores for African-Americans on certification exams are significantly lower than for whites on *every* test, most by wide margins. Hibpshman highlighted the fact that of the 30 tests listed only *four* would even meet the criteria of the four-fifths law. This means that with the cut scores that were used, on all but four of the tests in this sample, African Americans passed at a rate that was *less* than 80% of the passing rate of Whites.

Hibpshman (2004) uses these test results to demonstrate what the outcome would be if cut scores were raised to be set at the overall median score, as was suggested in the 2013 Title II report by the U.S. Department of Education. Using the Praxis test referenced earlier as an example, Principles of Learning and Teaching: Grades K-6, the White pass rate would be 52% and the African American pass rate would be 2%. Other tests have similar results. By this point, Hibpshman's conclusion

is obvious: “in most cases there would be severe consequences for African-American candidates” (p. 14).

Even contemplating using test scores with these outcomes exemplifies the position john powell (2012) calls false neutrality:

> One way of expressing color blindness is to be neutral on the issue of race. Proponents of this position apparently are most interested in neutrality in the design of policies and programs. They pay less attention to the administration or implementation of what they design and, more importantly, often ignore the effects of the policies and procedures they create.
>
> (p. 9)

Costs of a stance of false neutrality. What can really be inferred from such large race and class-based gaps in test results? What evidence exists that the so-called achievement gap, a crisis claimed as the basis for U.S. education policy reform, truly indicate gaps in achievement? Jay Soares (2012) points out “[t]he SAT has retained the same bell curve distribution since 1926, which some take as a measure of its validity, rather than an indicator of its role in transmitting social disparities” (p. 8). Scheunemann and Slaughter (1991) name five common explanations for historically large test score differences among groups, rooted in psychometrics, education, biology, culture, and/or history. Since some of these explanations are clearly social, Wakefield (2003) raises an important question: “What is biased—life or tests?” (p. 385). She continues, “If economic conditions, racial inequality, insufficient funding for schools, gender, lack of family support, or poor testing conditions are the

problem, states must address these life issues before the tests can work effectively" (p. 385). Put another way, for test results to be treated as if they are equitable measures of academic achievement, all other variables would need to be isolated and accounted for, requiring the elimination of any social hierarchy. The United States is, of course, far from this goal. That is one reason that Paul Gorski (2013) insists on naming these test score gaps as opportunity gaps rather than achievement gaps.

The common assumption of objectivity behind the use of standardized testing is based on a colorblind stance similar to the one john powell (2012) describes as taken by the Supreme Court since the 1976 *Washington v Davis* ruling. This standard, powell explains,

> feeds and is fed by an ideal of 'neutrality,' in which individuals live outside of any social, historical, or political context. This decontextualization is one of the rhetorical devices that the courts commonly used to justify the exclusion and subordination of different groups of people. Implicit in this approach is yet another rhetorical device: arguments and language that draw on dominant norms to convey the impression of objective decision-making.
>
> (p. 104)

By any measure, however, the so-called playing field (as if this were a game) of life in the U.S. has been and is far from level. Regardless of the sources of bias, and Wakefield believes they are many, when standardized tests are used as gatekeepers to a profession, the impact is the same… "fail the test, find another vocation" (p. 385).

Powell (2012) points out the combined impact of multiple programs and policies in the U.S. based on this false universalism:

> [w]hat false universalism fails to address is that groups of people are differently situated in relation to institutional and policy dynamics. If one looks at only one or two constraints, one is likely to inaccurately assume that groups in very different circumstances are in fact quite similar.
>
> (pp. 15-16)

For a group that has, say, ten social constraints, powell explains that even if one or two of were removed, as is often the case with anti-discrimination laws, the remaining eight would continue to restrict that group's opportunity. Yet when a group continues to struggle even after a perceived barrier is removed, the dominant group's perception is sometimes that the failure of group members to achieve a given goal lies *with the group,* rather than in the policies impacting the group. Powell continues, "[b]ut in order for progress to occur, that group's situation must be seen as a whole, including prior discrimination in education, housing, and health care…" (p. 16). He concludes:

> …[I]t is critical in a democracy that we be attentive to how opportunity is distributed, and to and for whom…It would also be useful for policymakers to deliberately consider how groups of people are situated, and the relevance of these situations, when designing and adopting policies"
>
> (p. 16).

Further, powell states that "[u]universal programs often operate on the unstated assumption that the particular conditions of the more favored group exist for all groups—that they are universal" (p.17). This statement certainly applies to the assumption of objectivity in high stakes use of large-scale standardized testing, and the assumption that these test scores measure *only* academic achievement. Apparently, they measure placement in the social and economic hierarchy as well. As explained earlier, historic evidence exists that this is just what they were statistically designed to do.

Whatever the causes, the outcomes of this stance of false neutrality are clear, and their costs are great. In a book titled *Contradictions in School Reform: Educational Costs of Standardized Testing*, Linda McNeil (2000) writes:

> The educational losses to minority students created by a centralized, standardized system of testing are many. What is taught, how their learning is assessed and represented in school records, what is omitted from their education—all these are factors that are invisible in the system of testing and in the accounting system reporting its results. Standardization of educational testing and content is creating a new kind of discrimination—one based not on a blatant stratification of knowledge access through tracking, but one which uses the appearance of sameness to mask persistent inequities.
>
> (pp. 251-252).

Human costs. In an article titled, "Oppression, Privilege and High Stakes Testing", Carl Grant (2004) writes about the mental and physical oppression of

people of color with regard to high-stakes standardized testing. Grant refers to the *American Heritage Dictionary* (1985, p. 872) definition of oppression as "a feeling of being heavily weighted down, either mentally or physically" (p. 3). He says high stakes testing causes this feeling of mental anxiety for students, teachers, administrators and parents. He cites the same dictionary to define privilege as "a special advantage, immunity, permission, right or benefit granted to or enjoyed by an individual, class or cast" (p.986 in Grant, p. 3). Grant ascribes privilege to the testing industry and politicians, but it seems clear that white upper and middle class test takers, too, have always been privileged when large-scale standardized tests have been used as the basis for educational decisions. Further, because so much educational research in the U.S. is based on large scale standardized test score results, reported as if they are true measures of academic achievement, this research, too, might be called to question. According to the Education Commission of the States and the Mid-continent Research for Education and Learning (2004), most participants in educational research are white, "which calls into question whether the results apply to participants from ethnic minority backgrounds" ("How Do I Know if the Research Warrants Policy Changes?" 2004, A Policymaker's Primer on Education Research).

Citing an earlier article he authored, powell (2012) writes "The color-blind argument is based on the seriously flawed assumption that whatever is not grounded in objective scientific data is not real" (powell, p. 31). In fact, this seemingly neutral stance allows this system of privilege and oppression to be masked—even in terms of language use. Powell contends that racialized messages are often sent in terms that

seem to be neutral. An example lies in terms that have recently come into the language of classroom teachers, used to describe the group of students affected by the so-called achievement gap. In 2014, a pre-service teacher asked me to define "gap students," a term she had heard used by one of her field-work teachers. Further research finds that Kentucky has developed a "Gap Delivery Plan," for which they have defined a

> Student Gap Group -- an aggregate count of student groups that historically have had achievement gaps. Student groups combined into the Student Gap Group include ethnicity/race (African American, Hispanic, Native American), special education, poverty (free/reduced-price meals), gender and limited English proficiency that score at proficient or higher.
>
> (Gap Delivery Plan Draft, 2011, Kentucky Department of Education)

Based on situating students in terms of test scores alone, this dehumanizing way to describe students epitomizes the deficit view about which Paul Gorski (2013) warns. Powell (2013) refers to this as part of the depersonalization of many aspects of our society, which Gorski (2013) and Sleeter (2013) contend are part and parcel of neoliberal initiatives. This depersonalization is now impacting teacher education programs as well, as faculty are told students must be accepted or denied admission to program based on whether or not they can meet the "acceptable" cut score on a state-selected test.

Costs of standardization and dehumanization in teacher education.

Although she doesn't use the term, Wakefield (2003) writes about the

dehumanization of education that is occurring as a result of education policy, including, but not exclusive to, the reliance on standardized testing as measures of student learning. In policy initiatives that demonstrate a general mistrust of the professional judgment of K-12 teachers, who had long been relied upon to assess the learning of students in their classes, education policy in the "accountability era" has instead come to rely on large-scale, multiple choice standardized test results, reporting them as if these judgments are more "scientific" and therefore more accurate than teacher-created assessment results. This practice, and the set of assumptions upon which it is based, is now finding its way into teacher education. ETS is now making decisions that used to be made by teacher-educators (Wakefield, 381). Koretz (2008) warns against ready acceptance of "misleading answers" (p. 118), but how is "misleading" to be defined? Is a test result "misleading" if a teacher educator determines, through a variety of assessments, that a candidate has the necessary grasp of content knowledge and pedagogical skills to teach? Often a faculty member has followed and supported a student in practice during many hours of clinical work, assessing and evaluating all the while. If the student has demonstrated excellence in classroom practice, yet doesn't make the cut score on the Praxis Exam, is the test score resulting from the series of inferences (Koretz; Hibpshman, 2004) "misleading"? Wakefield (2003) defines the dehumanization of teaching and learning inherent when test results are relied upon over human judgment, as if the test scores reveal truth. She writes, "

> Testing replaces a personal review of strengths and weaknesses and two years of face-to-face encounters. All candidates become statistics. Disadvantaged candidates often become victims of redundant testing. Moreover, with data shared among state departments of education and professional standards commissions, those from disadvantaged districts are not given the opportunity to rectify the inequities forced upon them by history and economics"
>
> (p. 386).

Costs of data systems re-defining the purpose of schooling. In the last thirty years, the idea that test scores can be used to measure student learning, the quality of schools, and the quality of teachers has become widely accepted (Wells,2014). In a brief for the National Education Policy Center called "Seeing Past the 'Colorblind' Myth of Education Policy," author Amy Stuart Wells finds that since 1994, through policy and in order to keep federal funding, each state has been forced to create a so-called accountability system, similar to the system Kentucky has named "Unbridled Learning" (Unbridled Learning, 2014). Intended to link K-16 student data with teacher education program data and teacher accountability data, these systems have become more common along with widespread adoption of the Common Core Standards and accompanying accountability systems. The problem, writes Wells, is that these test scores are fast becoming the sole factor in defining and determining what it means to be educated. Wells asks who in our society, other than the testing industry itself, could possibly benefit from such a narrow definition?

Among the many difficulties inherent in that assumption is the high correlation between schools deemed to be "failing" on the basis of standardized testing alone and the number of students of color and/or students from lower socio-economic classes in those schools (Wells, 2014). It has increasingly become standard operating procedure that punitive measures are then used to "motivate" failing schools to raise test scores. Hibpshman (2013) recommends sanctions for "failing" teacher education programs as well. The schools that get punished, then, are most often the schools with students from low SES backgrounds and/or students of color, which further exacerbates existing inequities. John powell (2013) warns of just such systems that, no matter what their stated intent, establish a cycle that results in maintaining the racial and socio-economic status quo. This is one reason it is so important that we study the *outcomes* of policy and systems, not just intent. Clearly, this system of testing and sanctions maintains just such a cycle.

Teacher education programs are required to report the passing rates of their candidates on state-required standardized tests. These passing rates are taken as an indicator of the quality of the programs. This reporting is mandatory in that it is tied to federal funding (Wakefield, 2003). "A teacher program's failure to report may result in a fine of up to $25,000 and loss of federal funding. This information includes the pass rate of candidates on assessments required by the state for teacher licensure or certification, the statewide pass rate on those assessments, and other basic information on teacher-preparation programs" (Wakefield, 2003, p. 383). Based on

the assumption that test scores indicate program quality, Wakefield (2003), citing Salzer and staff (2000) writes:

> Praxis I blocks the entry into teacher education for many minority and low-income candidates, while Praxis II blocks the exit. If education programs earn licensure rights according to their Praxis II pass rates, we can expect schools serving disadvantaged populations to discontinue teacher education as well as a decrease in diversity among teachers.
>
> (p. 384-5)

Federal proposals released in December of 2014 further increase this reliance on standardized test scores, requiring programs to gather data that will be used to link the test scores of a teacher candidate's future *students* back to the teacher preparation program. These data will then be presumed to be an indicator of the quality of the teacher education program the candidate attended (Federal Register, 2014). Given Koretz's (2008) words of caution about the careful use of test data as only one indicator of the single thing the test was designed to measure, it seems impossible to account for the many, many variables impacting the data the Federal Government proposes to collect, and the even longer string of inferences required to make any resulting assumptions about program quality. Yet Hibpshman (2013) recommends state sanctions against any program not meeting set quality standards, up to and including the closing of programs.

Costs of limited focus of instruction and assessment in teacher preparation programs. Wakefield (2003) observes the loss of many passionate, gifted and

committed teachers to the teaching profession as a result of the use of standardized testing as a gatekeeper. Besides the loss to these individuals, society also suffers a loss to the students these teacher candidates will never serve. How can a society predict and prioritize which factors are the *most* important predictors in high-quality teaching? What if teacher candidates who have struggled ultimately make the best teachers? Wakefield (2003) cites the critical importance of a teacher's ability to make human connections with students. Surely this is a factor in high-quality teaching. Is it possible that the ability to build relationships with students in order to reach them where they are is *more* important than content knowledge? If so, are these policies blindly screening out some candidates who might have superior interpersonal abilities? Of the multiple intelligences named in Howard Gardner's (1983) research, Wakefield acknowledges that, by attempting to measure only linguistic and mathematical intelligence, at least five others are excluded. What if these are important pieces of the complex puzzle that makes for high quality teaching? What dispositions are critical? Wakefield asks,"[e]ven granting Praxis the ability to measure knowledge of content and pedagogy, can it screen for compassion, character, understanding, and commitment?" (p. 387). Wakefield concludes that because of the use of Praxis as a sole gatekeeper, "…some passionate and gifted teachers will fail to find places of service in public schools" (p. 386).

Grant (2004) and Tucker (2011) contend that high-stakes tests have a large impact on what educators teach. Both authors regard the effect on U.S. instruction as negative because of the kinds of tests now in place. Popham (2014), however, feels

that criterion referenced tests with carefully established criterion rather than cut scores, could play an important and positive role in driving instruction. Bejar (2008) contends that the setting of cut scores plays an important role in establishing education policy with regard to what is taught.

In "Oppression, Privilege and High-stakes Testing," one source of oppression Grant (2004) names is the high-stakes use of multiple choice tests, which undermines quality teaching and learning by narrowing the curriculum to include as precisely as possible the knowledge and skills on the test that will be given. Koretz (2008) explains these test items are selected to represent a much larger domain of knowledge and skills. As educators try to match what is taught to what is tested on multiple choice tests, the quality of teaching and learning are negatively impacted.

In the book *Surpassing Shanghai: An Agenda for American Education Built on the World's Leading Systems*, Marc Tucker (2011) agrees. Tucker contends that the U.S. reliance on multiple-choice tests that can be computer graded is "heavily biasing the curriculum toward the teaching of [English and mathematics] and away from the teaching of other subjects that top-performing countries view as critical" (p. 176). Tucker goes on to say that while the emphasis in other countries is on the mastery of complex skills involved in problem solving tasks, the U.S. has emphasized so-called basic skills because of the kinds of tests it has chosen to use. He points out that the creation of the Common Core standards still emphasizes only two primary subject areas. Further, Tucker claims that no other country that outperforms the U.S. would even consider the use of computer scored multiple choice tests, because that

test format can't measure the educational outcomes they expect their students to achieve. In other words, Tucker contends the tests themselves set the bar too low for the kinds of teaching and learning that should take place in schools. Grant (2004) adds that the instructional hours spent preparing students to take high-stakes multiple choice tests are hours lost from high quality, culturally relevant teaching and learning.

Now that high-stakes multiple-choice standardized testing has also moved into teacher education, there will likely be a corresponding impact on instruction. Some may see this as a positive thing; others as negative, depending at least partially on their philosophical stance. Without a doubt, and especially *because* of the known outcomes for candidates of color and from low SES backgrounds, teacher education programs are being forced into finding instructional time for test preparation. After all, if this is the system in place, it only seems fair to do everything possible to help students pass the tests, especially because of the discrepant results for low SES candidates and candidates of color. This, of course, draws time and resources away from the kinds of education that may have a more direct impact on candidates' classroom teaching. Again, this can be viewed from many perspectives. If, as is proposed, a teacher's success will be measured in part by students' standardized test scores (Federal Register, 2014), it might indeed be best for a candidate to be well-schooled in how to teach to the test, by experiencing such teaching themselves. Marc Tucker (2011), however, might disagree on the grounds that the computer scored, multiple-choice Praxis tests promote exactly the kind of low level education that many other countries take steps to avoid. If a policy encourages teachers to teach in

ways that might directly help their students pass the test, it might be a logical stance that the impact of their teaching should be measured by the same kinds of tests. This proposed next step in federal regulation may have the effect of locking the entire system of what Tucker considers low-level teaching firmly into place.

Costs of test preparation. Ironically, Koretz (2008) demonstrates how test preparation is a form of test bias so strong that it actually invalidates test results. Koretz explains that because test items comprise a very small part of a subject area meant to represent a much larger domain, if a teacher teaches only what is represented on the test, some critical skills and knowledge will likely be ignored. Koretz observes that reform rhetoric is focused on alignment, emphasizing the importance of aligning standards with teaching, and teaching with assessment. Koretz argues, however, that the notion that a large-scale norm-referenced test "worth teaching to" (p. 254) might be developed is "nonsense" (p. 254), precisely *because* such tests always represent a very small subsection of what is important to learn.

Koretz (2008) also criticizes coaching for any test, beyond simple exercises to acclimate students to a test's format. Using the common example of a teacher showing students how to eliminate some answer choices on a multiple choice test, Koretz shows how this can easily invalidate the test results, because it focuses learning on the multiple choice test format itself. In other words, a valid test should give information about the learning the test was designed to measure. No test is designed to measure student success at *taking* the test, yet with widespread coaching, that is a likely outcome. In a high-stakes testing climate, the very purpose of the

testing seems all too easily forgotten. To truly test to see what students have *learned*, test items should be presented in many different formats, to see if the knowledge is transferrable. This is another of many reasons a test score alone should never be considered an indicator of student achievement (Koretz, 2008).

Coaching for a test artificially inflates test scores, because it can make "performance on the test unrepresentative of the larger domain" (Koretz, 2008, p. 256). Koretz continues, "[t]he acid test is whether the gains in test scores produced by test preparation truly represent meaningful gains in student achievement" (p. 258). Koretz' argument makes good sense, yet many educators now seem to take what they see as the necessity of test preparation for granted. Teacher education programs and colleges are finding new ways to coach their students, and especially students of color and from low SES backgrounds, in the hopes that they will be able to pursue their career of choice. These intentions, of course, are good. Koretz, however, reminds us, "What we should be concerned about is proficiency, the knowledge and skills, that the test scores are meant to represent. Gains…that do not generalize…to performance in the real world are worthless" (258). Yet Hibpshman (2004) points out that the transfer of learning represented by test scores into the real world of teaching is exactly where research is lacking. With such high-stakes costs for individuals and for society, this research base seems the least that should be expected. For these reasons, too, it seems valid to question whether the use of these tests is actually doing more harm than good, as teacher preparation programs adjust their curricula and spend time and money on test preparation which invalidates the tests themselves.

Costs of standardizing curriculum. It might seem obvious that, with the expectation of alignment and the use of large-scale, nationally-normed standardized tests for assessment purposes, standardized testing leads to standardization of curriculum. Some think that would be a good thing. Yet the implications of a standardized curriculum in a democratic nation as diverse as that of the United States are daunting. From whose perspectives, whose ways of knowing, will this standardized education be based? Who is included and who is left out? Traditionally, European perspective has been front and center in American education, a perspective so taken for granted as to often be considered simply "normal," and rarely questioned (Loewen, 2007). Implications for students who don't come from European backgrounds are well documented by multi-cultural educator James Banks (1993) and many others, including Grant (2004), who concludes:

> I have learned that during this high-stakes testing reform, the chances are diminishing of public schools richly contributing to their students becoming reflective, enlightened and critical learners who have an appreciation and acceptance of social justice and global and national ethnic and racial diversity (p. 10).

Standardization can all too easily lead to the expectation of assimilation, in this case to the white norm john powell (2012) describes. The system of oppression and privilege in standardized testing identified by Grant (2004) is masked by a pretense of neutrality, based on a misconception of "sameness" among Americans (powell, 2013). The view presumes there is one cultural norm, established by the

European Americans who colonized the United States (powell, 2013). The American Evaluation Association (AEA) cautions about making similar assumptions regarding evaluation as well. Their Statement on Cultural Competence in Evaluation (2011) says:

> Evaluation cannot be culture free. Those who engage in evaluation do so from perspectives that reflect their values, their ways of viewing the world, and their culture. Culture shapes the ways in which evaluation questions are conceptualized, which in turn influence what data are collected, how the data will be collected and analyzed, and how data are interpreted.
>
> (n.p.)

The statement continues that cultural competence is an "ethical imperative" and that an evaluation's validity depends on cultural competence.

As outlined earlier, the educational use of standardized tests and the statistics developed for their large-scale use arose during a time where the researchers' intent was to validate eugenics as a science. In most cases, the researchers and statisticians themselves were eugenicists. The people who advocated the educational use of these were also eugenicists. If that is the cultural norm that shaped the way these evaluations were "conceptualized, which in turn influence[s] what data are collected, how the data will be collected and analyzed, and how data are interpreted" (AEA, 2011, n.p.), should the ongoing score discrepancies between white people and peoples of color be raising alarm bells, especially now that test data are being used in such high-stakes ways? Should the assumptions upon which high-stakes uses for these

tests go unquestioned? Should they simply be accepted as neutral, objective and scientific? Should the large racial and socioeconomic discrepancies in their outcomes simply be overlooked? What is the human and societal cost?

The AEA statement on Cultural Competence in Evaluation emphasizes that theories, too, are "inherently cultural" (n.p.) As established in the first part of this Capstone, educational philosophies and the theories on which they are based have now been chosen *for* teacher education programs by regulating bodies. This raises questions about the academic freedoms which have long been carefully protected within U.S. higher education (Accreditation and Academic Freedom, 2012). Christine Sleeter (2013) and Paul Gorski (2013) point out that the theories upon which these policies are based arise from what they agree is a neoliberal agenda, which has dominated the past twenty years of school reform. These policies are now reaching into higher education at lightning speed, by way of state and federal regulations for teacher education programs. As explained previously in this document, regulations now dictate who will be considered unqualified to teach, how that determination will be made, how selected pre-service candidates will be "trained," the philosophy on which this training is based, and how student learning will be evaluated.

These policies will impact generations of students to come. The increased high-stakes use of large-scale standardized tests in U.S. K-12 education, and the addition of subsequent regulations requiring high-stakes uses of these tests in teacher education, has a predictable outcome: the teaching force will remain mostly white.

This raises a chilling question: Might this be the last cog in the gears that lock the U.S. social and economic hierarchy firmly into place?

Maintaining a White, Middle Class Norm: Affirmative Action for White Folks.

The historic and contemporary use of standardized test scores as if they were a measure of educational success is, ultimately, a way to justify white, middle class elitism (Au, 2013). As Au points out, when we look at the continued impact of using standardized test scores, such as those from Praxis Exams, as valid measures of their claims, only a few conclusions can be drawn. Au notes that the efforts over these past thirty years to close test-based "achievement gaps" have proven futile. The score gaps remain firmly in place. Given such large discrepancies in the demographic breakdown of test scores in over one-hundred years of the use of these tests, (always with white, middle class people from European ancestry with top scores), Au says only two conclusions can be drawn: either the eugenicists were right (!), or the tests themselves are neither objective nor accurate.

This is a consideration of utmost importance, since these tests are now being used as the basis of almost all educational research conducted in the U.S., and as the basis for almost all educational decision-making (Au, 2013). The title of Au's article clearly states his claim: "Hiding Behind High-Stakes Testing: Meritocracy, Objectivity and Inequality in U.S. Education." Jay Soares (2012) writes that even Charles Murray (1994), author of *The Bell Curve,* has joined the ranks of educators and researchers asking for the removal of the SAT as a criteria in college admission, due to the overwhelming evidence of harm to those at the bottom of the curve.

In an ETS Research Memorandum focused on the use of testing in teacher education, Gitomer and Latham (2000) concur that the impact of Praxis testing is clear. They confirm that the high-stakes use of these tests will assure the continuation of the social status quo among teacher education candidates. They write, "[m]ore than 4 in 5 Praxis Candidates are White, and more than 3 in 4 are female. Since, across all racial/ethnic groups, minority candidates tend to pass Praxis at lower rates than majority candidates, testing causes a predominantly White pool of prospective teachers to grow even whiter" (p. 5).

Closing the circle: Impacts for the future. If our nation's leaders continue to choose to bolster the cycle of privilege and oppression by adding teacher education students to the list of those whose fates are unfairly decided through the use of standardized test scores, the consequences for the future of social and economic justice in the U.S. are dire. Clearly, when minority groups are disadvantaged by the use of test data or any other policy, the dominant group, in this case white middle and upper class students, is advantaged. Legacies of these advantages and disadvantages have had long-term impacts on individuals and on the social hierarchy of society as a whole. What has been lost as members of various groups have been barred from participation in professions requiring standardized test scores as the sole measure of preparedness? How has the absence of representation in impacted fields like law, psychology, and education, fields that create the framework of the U.S. social order, affected us as a nation? What ways of knowing, of living, of being, have been

excluded from consideration as a result of the loss of those whose voices have been barred?

Research has clearly shown how important it is for the successful education of all children, including children of color, to have a teacher who looks like them (Villagas A.M. & Irvine, J.J., 2010). It's important for all students to be able to see themselves in the schools they attend: in the curriculum, on the schoolhouse walls, in the materials used for learning, and in their classroom leaders. Yet, Wakefield (2003) writes, "[u]nder the high-stakes screening tests, minorities can expect to see fewer teachers of their own race teaching their children" (p. 386).

"There are many reasons to be concerned about the small numbers of minority teachers" (p. 112), writes Mitchell, et al. (2006). Drawing from ideas of Choy et al. (1993) and the National Education Association (2002), the authors continue,"[t]he importance of minority teachers as role models for minority and majority students is one source of concern. Second, minority teachers can bring a special level of understanding to the experience of their minority students and a perspective on school policies and practices that is important to include. Finally, minority teachers are more likely to teach in central cities and schools with large minority populations" (p. 112).

In their study titled "Race, Gender, and Teacher Testing," Goldhaber and Hansen (2009) found

> evidence that Black teachers have more consistent success than White teachers in teaching minority students, and this matching effect is greatest in magnitude for Black teachers at the lower end of the licensure performance

> distribution. Moreover, the point estimates suggest that these matching effects are as important as any information conveyed through either the signaling or the screening functions of the tests when it comes to the achievement of minority students
>
> (p. 27).

It follows that if test scores continue to be used as a screen, and further, if cut scores are raised, the negative impact on the number of teachers of color will also have a significant negative impact on learning outcomes for students of color in this country, completing a cycle that reflects a racist history.

The impact of the use of standardized testing as a gatekeeper in teacher education means that anyone who does not fit the demographic of those who score well on these tests, white, middle and upper class folks, will be kept out of classrooms in disproportionately large numbers. This brings us back to the question of whether this is, in fact, a method of ensuring the preservation of the social and economic status quo in the U.S., the institutionalization of the white norm that john powell (2012) identifies?

These policies will have the impact of allowing only limited numbers of people of color or people raised in lower socio-economic circumstances from teaching our children. That these policies have been so quickly introduced and so eagerly embraced by people in positions of power, politicians and policy makers, rather than educators, may illuminate the powerful role teachers play in shaping our

society. Why do policy makers suddenly seek a larger degree of control over every aspect of education, from kindergarten to teacher education?

It seems that what is being created in education reform policy is actually a house of mirrors, reflecting so many fragmented images that practitioners and onlookers alike are confused by distorted images. Perhaps even some policy makers themselves are, at this point, unsure of which images are real and which illusion. Yet these distortions, in the form of test scores, are heralded as scientific and objective. Decisions that affect people's lives are taken out of the human context of teacher student relationships, and are instead based on data collected in ways we are told can be trusted, when the truth is that the data generated by large-scale, nationally normed standardized tests was never designed to stand alone or to be used to make high stakes decisions (Koretz, 2008). Politicians and policy makers then tell educators and the public what they should see in these test scores, how they should be interpreted, and, in fact, many pretend to see it, even as a different reality plays out right in front of them. This happens, for example, when a teacher listens to a child read, talks to him about what he has read, yet is told by his test scores that he is a non-reader; or when a student with a high test score in reading cannot relay any information about the meaning of the words she has just read from the page.

Most significantly, the human context of the eugenics movement, which guided both the psychometric development and rationale for the educational use of large-scale standardized tests, shows they were created in order to accomplish exactly the kinds of social and socio-economic sorting they continue to do. These are the

mechanisms that have created the U.S. social hierarchy since the time of colonization. It used to take place with guns. Now, Bennett, et al. (2006) claim this overreliance on test results is a new kind of discrimination, which uses the illusion of sameness to mask ongoing social inequalities. The outcome is the same, and the cycles that keep the social and economic hierarchy firmly in place continue. It is the professional obligation of educators to shatter the illusions created by this house of mirrors. Our very democracy is at stake.

Alternatives: A Call for Continued Research

In a report for ETS, Gitomer and Latham (2000) conclude that candidates passing teacher licensure tests appear to be "more academically able than those who do not" (p. 4). Their results rely on comparisons between scores on Praxis Exams and SAT scores, one of many examples of research resting on the assumption that test scores are true measures of academic achievement. Despite that conclusion, however, the authors call to question the implications of going for the "best and the brightest" in teacher education as determined by test scores alone. They suggest an alternative focus: working to ensure that teacher candidates develop into excellent teachers. Inherent in this statement is the acknowledgement that the relationship between test scores and excellent teaching has not been established. In fact, it doesn't even make much sense. Gitomer and Latham contend that "blindly raising testing standards may well do more harm than good" (p. 9) as "the supply of minority candidates was reduced much more drastically than the supply of majority candidates…" (p. 5). They therefore conclude that "Gross generalizations about supply, demand, and impact of

licensure miss the point that there are disparate effects across licensing areas and population groups that require much more complex and strategic analysis in order to support sound policy decisions" (p. 5). Yet the use of Praxis Exams as a way to narrow the pool of "qualified" candidates has increasingly become part of state and federal policy requirements.

Gitomer and Latham (2000) conclude their ETS Research Memorandum by writing, "The stakes are high, the problems exceedingly complex, and measures that redress one problem often exacerbate another. There will be no quick fixes or easy solutions" (p. 11).

Changing the question. Flynn Ross (2005) suggests that a crucial piece is missing from the teacher education policy reform equation: a focus on equity. Adding a single word to the question driving policy efforts might change everything that follows. What might happen if the question became: How can we *equitably* ensure that teacher education candidates are prepared to teach? The will to ask this question, of course, depends on whether social and economic equity is a priority for the United States. Working toward that goal seems an essential part of democracy. Author john powell (2012) writes that striving for unity, rather than the human separation created by a social and economic hierarchy that continues to grow wider, is the only way our nation will remain strong. Powell continues: "Altering structural barriers that impede progress for non-whites would result in a wide array of benefits: freeing minds, increasing social and geographical diversity and space, reducing poverty and alienation, and healing communities" (p. xxiii).

Questioning assumptions in educational research. Before exploring answers to this new question, which will involve alternatives to the use of large-scale standardized tests in teacher education, there is another important assumption to question. Just as standardized test scores are now often referred to in education policy as if they directly equate to student achievement, the same is true in educational research (Sleeter, 2011). The use of standardized test results as a proxy for student achievement has become so widely accepted, it is rarely questioned (Sleeter, 2011). Yet David Berliner calls educational research "the hardest science of all" (Berliner, 2002, title), in part, because a true definition for achievement is so difficult to achieve. Berliner and Sleeter (2011) explain the importance of context in educational research, and the inability to control the myriad of variables around teacher behavior, student behavior, and the interactions between the two as well as the influence of curricula and school policy. With so many variables involved, one begins to wonder if meaningful quantitative research is even possible in a school setting. Yet this is *the* type of research the U.S. Government has deemed worthy of attention and funding (Berliner, 2002). Yet Berliner, himself an educational researcher, contends "a single method is not what the government should be promoting for educational researchers" (Berliner, p. 20). Based on the importance of context, "...ethnographic research is crucial, as are case studies, survey research, time series, design experiments, action research and other means to collect reliable evidence for engaging in unfettered argument about education issues" (p. 20).

At the very least, the current emphasis on quantitative research alone restricts even the kinds of questions that can be raised (Sleeter, 2011). Far from demonstrating systems thinking, this view of research narrows the view of a complex, interrelated system to such a small speck, it is like studying a piece of lichen on a particular type of tree and then using that information to extrapolate its meaning for every kind of tree in the forest. To take the analogy one step further, those data would then be used to set policy for the management of the forest as a whole.

Much educational research in the United States today is focused on improving student achievement. Although achievement can be defined in many different ways, the current, often unquestioned assumption by many researchers is that large-scale, multiple choice standardized tests do, indeed, accurately measure the academic achievement of students (Sleeter, 2011). One wonders how questioning this assumption might impact the outcomes. Indeed, this type of questioning is an inherent part of any kind of scientific research (Berliner, 2002). Especially given the hidden roots of the educational testing industry itself, it seems crucial to call to question educational researchers' use of test scores as lone indicators of student achievement. Much might be learned from studying ways researchers in other countries define academic achievement, because widespread use of multiple-choice standardized tests is not the global norm (Sleeter, 2011).

In this era of accountability, policy-makers and politicians increasingly require evidence that methods of instruction and assessment are research-based. Like assessment, research occurs in a cultural context (Berliner, 2002). When

contemplating the use of high-stakes standardized testing in teacher education, then, as well as possible alternatives, it is important to continue to critique research methods and measurements before accepting results as evidence. Because of the importance of context, this scientific practice of questioning at every step is especially important in social research.

Questioning fairness in standardization. In this era where demands for accountability have become the highest educational priority, even things that seem obvious are sometimes hidden from view. John powell (2012), for example, aids readers in seeing the white norm that is institutionalized in the United States. A similar truism exists in the cycle of teaching and learning. Students and teachers are not standard. Their cultural situations are not standard. This truth underlies the difficulties in attempting to standardize *anything* in education, let alone attempts to standardize almost everything. As john powell (2012) points out, equity is not achieved by treating everyone as if they were the same, living under the same conditions. Educators seem to agree: differentiation of instruction has become an important point of conversation in educational circles. It stands to reason, then, that teacher education policy will need to allow teacher educators the same flexibility (Ross, 2005) in differentiating instruction and assessment to meet the individual needs of their students as the teacher candidates are expected to demonstrate with students in their field work classrooms. In teacher education, as in general education, one size will never fit all. Treating students as if they are standard is simply not fair. Bennett, et al. (2006) concur:

> PRAXIS I, as it is currently used in most settings, is an inequitable TEP admissions tool because it establishes a single standard to assess the capabilities of talented students who have had unequal educational opportunities and unequal access to the knowledge needed to attain passing scores on the test. We are not advocating 'special consideration' for students of color who take the test; we ask for fairness. The test does not ensure high standards, as advocates have hoped, and instead excludes many talented students of color both because PRAXIS I itself is unfair and because P-12 schools do not provide a high-quality education for *all* students from all ethnic, linguistic, socioeconomic, and geographic backgrounds
>
> (p. 567).

Exploring Alternatives. It is clear that more research is needed in the search for equitable evaluation measures that might predict future teaching success. In addition, the research methods themselves must be based on equitable measures of academic success. Alternatives to the use of standardized test scores do exist. Wakefield (2003) gives this reminder of a not-too-distant past:

> Before high-stakes tests dominated the educational landscape, teacher candidates were screened, on a case by case basis, by experienced teacher education professionals in nationally and/or state accredited programs. Though grades and test scores of these candidates were open for review, decisions were based on a variety of considerations. Future teachers typically filed an entrance application or portfolio and were personally interviewed.

> Applications included recommendations, references, academic highlights, philosophy of education, and career goals. Using a holistic approach to screen incoming candidates, teacher education programs addressed some of the following issues: Experience…Relevance…Strengths and weaknesses…Intellectuals and toilers…Brains and heart…[and] Learning styles. (Wakefield, 2003, p. 386).

As Wakefield acknowledges, the assumption that there should be some standardized way to compare evaluations across programs and institutions is just that, an assumption, based on a set of philosophical beliefs. In fact, other than for purposes of certification, these evaluations used to occur within the context of student/teacher relationships at the institutional level.

Some alternative forms of evaluation include the use of standardized test scores as part of a holistic approach. Despite their critiques of high-stakes uses for standardized tests, Koretz (2008) and Robert Sternberg (in Jaschick, 2010), have no problem with their low-stakes use, as one source of data in a much larger picture.

Sternberg has developed a new system of evaluation for college admissions called Kaleidoscope (Jaschick, 2010). It is intriguing to consider possible applications of this system in teacher education. In an interview published on the *Inside Higher Ed* website in 2010, Sternberg explained that what sets the Kaleidoscope Project apart is its ability to evaluate some of the qualities Wakefield (2003) describes. Sternberg said:

The Kaleidoscope Project has three features that are perhaps distinctive. These features emanate from the view that the purpose of college/university education is to produce the leaders of tomorrow who will make a positive, meaningful, and enduring difference to the world.

> First, the questions are based on a theory of leadership, WICS -- wisdom, intelligence, creativity, synthesized -- according to which positive leaders need a synthesis of (a) creative skills and attitudes in order to generate new ideas; (b) analytical skills and attitudes in order to ensure that the ideas are good ones; (c) practical skills and attitudes to implement their ideas and to persuade others of the value of these ideas; and (d) wisdom-based skills and attitudes to ensure that the ideas help to achieve a common good, over the long and short terms, through the infusion of positive ethical values. So the questions in Kaleidoscope are designed to measure these creative, analytical, practical, and wisdom-based skills and attitudes.
>
> (Jaschik, S. 2010, n.p.)

Sternberg explains that student responses are evaluated holistically and based on the application as a whole through the use of rubrics. He claims scores on Kaleidoscope do not show the "substantial ethnic group differences," that standardized tests do. The reason, he asserts, based on his research, is that Kaleidoscope assesses much more than a narrow subset of skills (Jaschik, 2010).

Arguably, classroom teachers need to be able to demonstrate a similar set of skills as those Sternberg describes in the interview with Jaschik (2010). One

suggestion for future research, then, would be a study using college entrance data gathered from the Kaleidoscope evaluations of teacher education candidates who went on to teach after graduation. These results might then be compared to multiple measures of teaching success (beyond just standardized test scores) to see if results from the Kaleidoscope evaluation system might correlate with future teaching success in ways that might make it a valid predictor. If correlations are indicated, further research might then lead to the creation of a similar type of evaluation system that could be administered for purposes of teacher certification.

Another alternative to the use of standardized test scores as gatekeepers in teacher education is the use of portfolio evaluations. Currently, some states require portfolio evaluation for teacher licensure. (Darling-Hammond, Pacheco, Michelli, LePage, Hammerness, Youngs, 2005). EdTPA is one example of a performance-based portfolio assessment designed to supplement other evaluations of basic skills and knowledge (EdTPA, n.d.). To avoid the perpetuation of widely disparate racial and social economic outcomes, however, any standardized testing information included as part of these portfolio data would need to exclude high-stakes implications, eliminating the need for setting cut-scores. Test scores could instead be used as Koretz (2008) says they are intended: as useful indicators that are part of a larger evaluation package.

In the conclusions based on their research, Bennett, et al. (2006) make five recommendations that might be used as "stopgap measures intended to be used until a fair test can be created" (p.568). These include: waiving Praxis I requirements for

candidates with SAT scores of 1,000 or above in order to avoid the added expense of a redundant test; pre-testing potential teacher education candidates on basic skills so that supplemental coursework can be offered in needed areas; allowing candidates to pass Praxis I with a composite score for all sections equal to the combination of required cut-scores on each section; admitting students to teacher education programs on a provisional bases if they have a high GPA and "exhibit strong evidence of teaching ability" (p. 568) even if they do not meet the Praxis cut-scores; and allowing candidates whose first language is not *standard* English unlimited time to take the Praxis. In proposing these as stopgap measures, Bennett, et al. apparently have either concluded that the use of large-scale standardized evaluation is now inevitable in teacher education programs, and/or they hold onto the hope that development of a large-scale standardized evaluation tool that produces equitable results might be possible.

Exhibiting a similar hope, Mitchell et al. (2001), in a report for the Committee on Assessment and Teacher Quality write:

> The committee contends that the effects of groups differences on licensure tests are so substantial that it will be difficult to offset their impact without confronting them directly…it is critically important that, where there is evidence of substantial disparate impact, work must be done to evaluate the validity of tests and to strengthen the relationships between tests and the knowledge, skills, abilities, and dispositions needed for teaching. In these instances the *quality of the validity of evidence* is very important.

(p. 113, emphasis added).

Mitchell et al. (2001) continue:

> The initial licensure tests currently in use rely almost exclusively on content-related evidence of validity. Few, if any, developers are collecting evidence about how test results relate to other relevant measures of candidates' knowledge, skills, and abilities. It is important to collect validity data that go beyond content-related validity evidence for initial licensing tests. However, conducting high-quality research of this kind is complex and costly. Examples of relevant research include investigations of the relationships between test results and other measures of candidate knowledge and skills or on the extent to which tests distinguish candidates who are at least minimally competent from those who are not.
>
> (p. 18).

Until the United States has invested in the kinds of research Mitchell et al. (2001) refer to, research which will allow for the development of systems capable of truly evaluating qualities that might predict a candidate's potential for high-quality teaching, reason would call for a moratorium on the use of large-scale, multiple-choice, standardized test results for *any* high-stakes purpose. The costs and consequences are simply too great. In teacher education, that would eliminate the use of Praxis as a gatekeeper at program entry as well as the use of Praxis II cut scores to bar the certification of teacher education candidates who meet all other criteria.

Even though legal precedent allows for the injustice, the tests in current use in teacher education simply do not meet the validity criteria which correspond to their use, nor do they meet ethical standards of equity. The resulting losses to individuals, to society as a whole, to future students, and therefore to the future potential of the United States as a nation are far too great. Beyond any reasonable doubt, the purported benefits of high-stakes uses of Praxis tests in teacher education simply cannot be worth the costs to equity and justice in a nation striving to live up to its democratic principles.

References

A policymaker's primer on education research: How to understand, evaluate and use it. (2004). Mid-continent Research for Education and Learning (McREL) and the Education Commission of the States (ECS). Retrieved from http://www.ecs.org/html/educationissues/research/primer/researchwarrants.asp

About Berea College. (2013). *Berea College*. Retrieved December 7, 2013, from http://www.berea.edu/about/

Accreditation and academic freedom: An American Association of University Professors Council for Higher Education Accreditation advisory statement. (2012). Retrieved from http://www.chea.org/pdf/AAUP-CHEA%20-%20FINAL.pdf

Achieve. (2014). *College and Career Readiness.* Retrieved August 31, 2014, from http://www.achieve.org/college-and-career-readiness

Adichie, C. (2009). Chimamanda Ngozi Adichie: The danger of a single story. Retrieved from http://www.ted.com/talks/chimamanda_adichie

Ahmad, F.Z. & Boser, U. (2014). *American's leaky pipeline for teachers of color: Getting more teachers of color into the classroom.* Center for American Progress. Retrieved from http://www.americanprogress.org/issues/race/report/2014/05/04/88960/americas-leaky-pipeline-for-teachers-of-color/

Allred, S., Draut, A., Ellis, A. & Liguori, B. (2014). *Student growth and the growth and effectiveness system* [Powerpoint]. Retrieved July 25, 2014 from http://education.ky.gov/teachers/PGES/TPGES/Pages/TPGES-Student-Growth-Page.aspx

Americans instrumental in establishing standardized tests. (n.d.). Frontline: *Secrets of the SAT*. Retrieved March 5, 2014, from http://www.pbs.org/wgbh/pages/frontline/

American Evaluation Association statement on cultural competence in evaluation. (2011). Retrieved from http://www.eval.org/p/cm/ld/fid=92

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In Holland, P.W. & Wainer, H., Eds. *Differential item functioning*, pp. 2-23. Retrieved from http://books.google.com/books?id=6YAXJfswvfYC&printsec=frontcover#v=onepage&q&f=false

Angrist, J.D., & Guryan, J. (2007). Does teacher testing raise teacher quality? Evience from state certification requirements. *Economics of Education Review*. Retrieved from doi:10.1016/j.econedurev.2007.03.002.

Au, W. (2013). Hiding behind high-stakes testing: Meritocracy, objectivity and inequality in U.S. education. *International Education Journal: Comparative Perspectives 12*(2). Retrieved from http://ojs-prod.library.usyd.edu.au/index.php/IEJ/article/view/7453

Bachelor of Education (Primary). (2012). Deakin -. Retrieved November 27, 2013, from http://www.deakin.edu.au/course/bachelor-of-education-primary-education-and-teaching

Bachelor of Teaching (Secondary)/Bachelor of Arts. (2012). Deakin -. Retrieved November 27, 2013, from http://www.deakin.edu.au/course/bachelor-of-teaching-secondary-bachelor- of-arts-education-and-teaching

Badia, E. (2014, March 27). New York defends standardized tests for students as movement against them builds. *NY Daily News*. Retrieved May 18, 2014, from http://www.nydailynews.com/new-york/education/movement-standardized-testing-growing-article-1.1736703

Baker, B., & Oluwole, J. (2013). Deconstructing disinformation on student growth percentiles & teacher evaluation in New Jersey. *New Jersey Education Policy Forum*. Retrieved July 15, 2014, from http://njedpolicy.wordpress.com/2013/05/02/deconstructing-disinformation-on-student-growth-percentiles-teacher-evaluation-in-new-jersey/

Banks, J. (1993). Approaches to multicultural curriculum reform. In J. Banks and C. Banks (Eds.), *Multicultural education: Issues and perspectives.* Boston: Allyn & Bacon.

Barnett, D., Christian, C., Hughes, R., & Wallace, R. (2010). *Privileged thinking in today's schools: The implications for social justice*. Lanham, Md.: Rowman & Littlefield Education.

Bejar, I. (2008). Standard setting: What is it? Why is it important? *R & D Connections*. Educational Testing Service.

Bennett, C.I., McWhorter, L.M., Kuykendall, J.A. (2006). Will I ever teach? Latino and African American students' perspectives on Praxis I. *American Educational Research Journal.* Retrieved from DOI: 10.3102/00028312043003531

Berea College. (2013). *Excerpts from the Berea College catalog and handbook: The academic program 2012-13*. (2012). Berea College.

Berea College Office of Institutional Research and Assessment. (2012-2013). *Berea College Fact Book*. Retrieved from http://www.berea.edu/ira/files/2013/11/factbook1314.pdf

Berea College Strategic Planning Committee. (2011). *Being and becoming in the 21st century: The strategic plan for Berea College*. Retrieved from http://www.berea.edu/cisrk/files/2012/08/being-and-becoming-june-2011.pdf

Berliner, D.C. (2002) Educational research: The hardest science of all. *Educational Researcher (31)* 8 p18-20. Retrieved from . http://mkoehler.educ.msu.edu/hybridphd/hybridphd_summer_2010/wp-content/uploads/2010/06/Berliner-2002.pdf

Berry, S. (2014). Common Core: How 'nonprofits' reaped millions. *Breitbart News Network.* Retrieved July 13, 2014, from http://www.breitbart.com/Big-Government/2014/01/25/Common-Core-Where-Nonprofits-Reaped-Millions

Bigelow, B. (2012). From Johannesburg to Tucson. *Rethinking Schools 26*(4). Retrieved from http://www.rethinkingschools.org/archive/26_04/26_04_bigelow.shtml

Biserial correlation, (2014). *Glossary of standardized testing terms*. Educational Testing Service. Retrieved from https://www.ets.org/understanding_testing/glossary/

Black, E. (2003). The horrifying American roots of Nazi eugenics. *George Mason University History News Network.* Retrieved March 5, 2014, from http://hnn.us/article/1796.

Brutlag, J.D. (2007). The development of correlation and association in statistics. Retrieved from http://buttelake.com/corr.htm

Bushaw, W.J. & Lopez, S.J. (2013) Which way do we go? *Phi Delta Kappan 95*(1), 9-25. Retrieved from EBSCOhost Academic Search Premier Database.

CAEP Accreditation Standards. (2013). Council for the Accreditation of Teacher Preparation. Retrieved from http://caepnet.org/standards/standards/

Calkins, L. (2013, June 11). In the wake of the ELA exam: Lessons from NYS. *The Reading & Writing Project*. Retrieved May 19, 2014, from http://readingandwritingproject.com/news/2013/06/11/In-the-Wake-of-the-ELA-exam-Lessons-from-NYS.html

California Commission on Teacher Credentialing. (July, 2012). *CTC Statistic of the Month.* Retrieved from www.ctc.ca.gov/educator-prep/statistics/2012-07-stat.pdf

Carjuzza, J., Jetty, M., Munson, M. & Veltkamp, T. (2010). Montana's Indian Education for All: Applying multicultural education theory *Multicultural Perspectives 12*(4), pp. 192-198. Retrieved from EBSCO host.

CCSSO Task Force on Educator Preparation and Entry into the Profession. (2012). *Our responsibility, our promise: Transforming educator preparation and entry into the Profession.* Retrieved from http://ccsso.org/Resources/Publications/Our_Responsibility_Our_Promise_Transforming_Educator_Preparation_and_Entry_into_the_Profession.html#stha sh.ZSgIEti3.dpuf

Cole, N.S. (1993). History and development of DIF. In Holland P.W. & Wainer, H. Eds. *Differential item functioning,* pp. 25-29. Retrieved from http://books.google.com/books?id=6YAXJfswvfYC&printsec=frontcover#v=onepage&q&f=false

Collins, C. (2004). Envisaging a new education studies major: what are the core educational knowledges to be addressed in pre-service teacher education? *Asia Pacific Journal of Teacher Education 32*(3), 227-240. Retrieved from EBSCOhost Academic Search Premier Database.

Common Core Standards Initiative: Preparing America's Students for College and Career. (2014). Retrieved from http://www.corestandards.org/

Cowan, G. (2007) *Understanding and conducting research: A user-friendly approach.* Dubuque, IA: Kendall-Hunt Publishing Co.

Crowe, E., with Allen, M. and Coble, C. (2013). *Outcomes, measures, and data systems: A paper prepared for the CAEP Commission on Standards and Performance Reporting*. Teacher Preparation Analytics, LLC. Retrieved from https://caepnet.files.wordpress.com/2012/12/caep-paper-final-1-19-2013.pdf

Criterion referencing. (2014). *Glossary of standardized testing terms*. Educational Testing Service. Retrieved from https://www.ets.org/understanding_testing/glossary/

Cut Score Framework Procedure. (approved, 2012). Internal procedures. Kentucky Education Professional Standards Board Policies and Procedures. Retrieved from http://www.epsb.ky.gov/documents/BoardInfo/EPSB_Internal_External_Procedures_Policies.pdf

Danielson, Charlotte. (adapted for Kentucky 2011). *Framework for Teaching.* Retrieved from http://education.ky.gov/teachers/hieffteach/pages/pges--overview-series.aspx

Darling-Hammond, L., Chung Wei., R., with Johnson, C. (2009). Teacher preparation and teacher learning. In Sykes, G., Schneider, B. L., Plank, D. N., & Ford, T. G. (Eds.) *Handbook of Education Policy Research* (pp. 613-636). New York: Routledge. Retrieved from https://scale.stanford.edu/.../Teacher_Preparation_and_Teacher_Learning

Darling-Hammond, L. & Bransford, J. (2005). *Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able to Do.* San Francisco: Jossey-Bass.

Darling-Hammond, L, Pacheco, A., Michelli, N., LePage, P., Hammerness, K., with Youngs, P. (2005) Implementing curriculum renewal in teacher education: Managing organizational and policy change. In Darling-Hammond, L & Bransford, J., (Eds.), *Preparing teachers for a changing world: What teachers should earn and be able to do* (pp. 442-479). San Francisco: Jossey-Bass.

Differential item functioning. *Glossary of Standard Testing Terms*. ETS. Retrieved from https://www.ets.org/understanding_testing/glossary/

Dutro, E. (2009). Children writing "hard times": Lived experiences of poverty and the class-privileged assumptions of a mandated curriculum. *Language Arts, 87*(2), 89-98. Retrieved from EBSCOhost Academic Search Premier database.

Kentucky Educational Professional Standards Board cut score framework procedure. (Approved Jan. 9, 2012). *Internal procedures*. Retrieved from http://www.epsb.ky.gov/documents/BoardInfo/EPSB_Internal_External_Proceedures_Policies.pdf

FAQ General information. (n.d.) *EdTPA.* American Association of Colleges for Teacher Education. Retrieved from http://edtpa.aacte.org/faq#51

EPSB data dashboard. (2014). Educational Professional Standards Board (EPSB). Retrieved from https://wd.kyepsb.net/EPSB.WebApps/Dashboard/DashbrdWeb/

Federal Register. (2014). *Teacher preparation issues: A proposed rule from the Education Department on 12/3/2014*. Retrieved on Dec. 30, 2014 from https://www.federalregister.gov/articles/2014/12/03/2014-28218/teacher-preparation-issues

Feiman-Nemser, S. (1990). Teacher preparation: Structural and conceptual alternatives. In Houston, W. R., Haberman, M., & Sikula, J. P. (Eds.) *Handbook of Research on Teacher Education* (pp. 1-61). New York: Macmillan. Retrieved from ncrtl.msu.edu/http/ipapers/html/pdf/ip895.pdf

Feistritzer, C.E. *Profile of teachers in the U.S. 2011.* (2011). National Center for Education Information. Retrieved from http://www.edweek.org/media/pot2011final-blog.pdf

Field test. (2014). Smarter Balanced Assessment Consortium. Retrieved June 19, 2014, from http://www.smarterbalanced.org/field-test/

Flippo, R.F. (2003). Canceling diversity: High-stakes teacher testing and the real crisis. *Multicultural Perspectives 5*(4), 42-25.

Frederickson, N. (1994). *The influence of minimum competency tests on teaching and learning.* Educational Testing Service, Princeton, New Jersey. Retrieved from http://files.eric.ed.gov/fulltext/ED369820.pdf

Gap delivery plan draft. (2011). Kentucky Department of Education. Retrieved from http://education.ky.gov/districts/tech/kmp/documents/gapdeliveryplan010312.pdf

Gardner, H.(1983). *Frames of .mind*: *The theory of multiple intelligences.* New York: Basic Books.

Gitomer, D. (2007). *Teacher quality in a changing policy landscape: Improvements in the teacher pool*. Educational Testing Service. Retrieved from http://www.ets.org/Media/Education_Topics/pdf/TQ_full_report.pdf

Gitomer, D. H. & Latham A.S. (2000). Generalizations in teacher education: Seductive and misleading. *ETS Research Memorandum*. Retrieved from http://www.ets.org/research/policy_research_reports/publications/report/2000/imcz

Glossary of Standardized Testing Terms. (2014). *How ETS Approaches Testing.* Educational Testing Service. Retrieved June 13, 2014, from https://www.ets.org/understanding_testing/glossary

Goldhaber, D. & Hansen, M.(2009). Race, gender, and teacher testing: How informative a tool is teacher licensure testing? *American Educational Research Journal OnlineFirst.* Retrieved from doi: 10.3102/0002831209348970

Gorski, P. C. (2013). *Reaching and teaching students in poverty: strategies for erasing the opportunity gap*. New York: Teacher College Press.

Gould, S.J. (1996). *The mismeasure of man*. New York, N.Y.: WW. Norton & Company.

Grossman, P., Schoenfeld, A., with Lee, C. (2005). Teaching subject matter. In Darling-Hammond, L & Bransford, J., (Eds.), *Preparing teachers for a*

changing world: What teachers should learn and be able to do (pp. 201-231). San Francisco: Jossey-Bass.

Grant, C. A.(2004). Oppression, privilege and high-stakes testing. *Multicultural Perspectives 6* (1), 3-11.

Hammond, L., Bransford, J. (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do.* San Francisco, CA: Jossey-Bass.

Harvill, L.M. (1991). *An NCME Instructional Module on standard error of measurement*. National Council on Measurement in Education. Retrieved from http://ncme.org/linkservid/6606715E-1320-5CAE-6E9DDC581EE47F88/showMeta/0/

Herz, A. (2014, February 26). Seattle teachers vow to boycott a new standardized test. *Seattle News*. Retrieved May 19, 2014, from http://www.thestranger.com/seattle/seattle-teachers-vow-to-boycott-a-new-standardized-test/Content?oid=18966894

Hibpshman, T. (2013) *Design of an Education Professional Standards Board (EPSB) preparation and accountability system for teacher training programs.* Kentucky Education Professional Standards Board. Retrieved from http://www.epsb.ky.gov/dataresearch/index.asp

Hibpshman, T. (2004). *Considerations related to setting cut scores for teacher tests.* Kentucky Education Professional Standards Board. Retrieved from www.kyepsb.net/documents/Stats/Journals/cut_score_analysis.pdf

Hill, H., Unland, K., Litke, E., Kapitula, L. (2012). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education* 118. University of Chicago.

Hoffman, J. V. (2000). The de-democratization of schools and literacy in America. *The Reading Teacher*, 53(8), 616-620. Retrieved from EBSCOhost Academic Search Premier database.

How tests and test questions are developed. (2014). *How ETS approaches testing.* Retrieved June 19, 2014, from https://www.ets.org/understanding_testing

Interpreting your Praxis Examinee score report. (2009). Educational Testing Service. Retrieved Dec. 17, 2014 from http://www.ets.org/s/praxis/pdf/sample_score_report.pdf

Introduction to psychometric tests. (n.d.). Institute of Psychometric Coaching. Retrieved June 13, 2014, from http://www.psychometricinstitute.com.au/Psychometric-Guide/Introduction_to_Psychometric_Tests.html

Jacobs, N. (2013). Traditional teacher education still matters: Healthy doses of theory plus clinical practice still lays a solid foundation for new teachers. *Phi Delta Kappan 94*(7), 18-22. Retrieved from EBSCOhost Academic Search Premier Database.

Jaschick, S. (2010). College admissions for the 21st century. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/news/2010/09/28/sternberg

Jennings, J. (2012). *Reflections on a half-century of school reform: Why have we fallen short? Where do we go from here?* Center of Education Policy. Washington, D.C.

Jensen, A.R. Galton's legacy to research on intelligence. (2002). *Journal of Biosocial Science 34* (2), pp. 145-172. Retrieved from http://arthurjensen.net/wp-content/uploads/2014/06/Galtons-Legacy-to-Research-on-Intelligence-2002-by-Arthur-Robert-Jensen.pdf

Karp, S. & Sokolower, J. (2014). Colonialism, not reform: New Orleans schools since Katrina. *Rethinking Schools 28* (4), pp. 32-38.

Keeping test content current. (2014). *Praxis National Advisory Committees.* The Praxis Series. Educational Testing Service. Retrieved from http://www.ets.org/praxis/states_agencies/about/content_current/nac

Kentucky teaching certificates. (2014). 16 KAR 2:010. Kentucky Legislature. Retrieved from http://www.lrc.state.ky.us/kar/016/002/010.htm

Kentucky test requirements. (2014). *Praxis: Kentucky: Test requirements*. Retrieved October 11, 2014, from http://www.ets.org/praxis/ky/requirements

Kidder, W.C.& Rosner, J. (2002). How the SAT creates built-in-headwinds: An educational and legal analysis of disparate impact, *Santa Clara L. Review* 43. Retrieved from: http://digitalcommons.law.scu.edu/lawreview/vol43/iss1/3

Kiley, K. (2013) Education in the liberal arts. *Inside Higher Ed.* Retrieved from http://www.insidehighered.com/news/2013/05/24/colorado-colleges-education-major-challenges-whether-disciplines-still-define

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press: Cambridge, MA.

Koretz, D. (2008). A measured approach: Value-added models are a promising improvement, but no one measure can evaluate teacher performance. *American Educator*. Retrieved from http://www.aft.org/pdfs/americaneducator/fall2008/koretz.pdf

Kozol, J. (2005). *The shame of the nation: the restoration of apartheid schooling in America*. New York: Crown Publishers.

Knapp, J.E. & Knapp, L.G. (1995) Practice analysis: Building the foundation for validity. In *Licensure testing: Purposes, procedures, and practices.* Buros Institute of Mental Measurements: Lincoln, Nebraska. Retrieved from http://digitalcommons.unl.edu/buroslicensure/9/

Ladson-Billings, G.(1995). But that's just good teaching! The case for culturally relevant pedagogy. *Theory into Practice 34*(3). Retrieved from http://equity.spps.org/uploads/but_that_s_just_ladson-billings_pdf.pdf

Layton, L. (2014, June 10). Gates Foundation urges delay in using tests for teacher evaluation. *Washington Post.* Retrieved September 27, 2014, from http://www.washingtonpost.com/local/education/gates-foundation-urges-delay-in-using-tests-for-teacher-evaluation/2014/06/10/d037c7fa-f0e1-11e3-914c-1fbd0614e2d4_story.html

Lee, T. (2014, May 7). Education racial gap wide as ever according to NAEP. *MSNBC*. Retrieved July 13, 2014, from http://www.msnbc.com/msnbc/student-proficiency-stagnant-race-gap-wide

Livingston, S. & Zieky, M. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Educational Testing Service.

Loewen, J. (2007). *Lies my teacher told me: Everything your American history textbook got wrong*. New York, NY: Touchstone.

Lyman, T. (1973). *A history of Indian policy.* Document Resume ED 092 279 BC 07 929. Bureau of Indian Affairs (Dept. of Interior), Washington, D.C. Retrieved from http://files.eric.ed.gov/fulltext/ED092279.pdf

McNeil, L.M. (2000). *Contradictions of reform: The educational costs of standardized testing*. New York: Routledge.

Mitchell, K.J., Robinson, D.Z., Plake, B.S. & Knowles, K.T., ed. (2001). Testing teacher candidates: The role of licensure tests in improving teacher quality. *Committee on Assessment and Teacher Quality: Center for Education Board on Testing and Assessment: Division of Behavioral an Social Sciences and Education National Research Council*. Washington, D.C. National Academy of Sciences. Retrieved from http://www.nap.edu/openbook.php?record_id=10090&page=R2

Mission. (2013). *About Berea College*. Retrieved January 1, 2014 from http://www.berea.edu/about/mission/

Murphy, C. M. (Oct. 25, 2013). KACTE fall 2013 retreat Kentucky Praxis program update [Powerpoint]. Retrieved from http://www.kacte.org/assets/ky-kacte-fall-retreat-10-25-2013.pdf

National Alliance of Black School Educators- legislative national standards. (n.d.). National Alliance of Black School Educators. Retrieved July 13, 2014, from http://www.nabse.org/leg_standards.html

National Center for Education Statistics. (2010). *Status and trends in the education of racial and ethnic groups.* U.S. Department of Education. Retrieved from http://nces.ed.gov/pubs2010/2010015.pdf

National Center for Research on Teacher Learning (1993). *An annotated bibliography: Findings on learning to teach.* College of Education, Michigan State University. Retrieved from ncrtl.msu.edu/http/annot.pdf.

NCATE. (2010-2014). *Unit standards in effect 2008*. Retrieved August 31, 2014, from http://www.ncate.org/Standards/UnitStandandardsineffect2008/tabid/476/Default.aspx

Nettles, M.T., Scatton, L.H., Steinberg, J.H., Tyler, L.L. (2011). *Performance and passing rate differences of African American and White prospective teachers on Praxis Examination*. Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR- 11-08.pdf

Norton, B. (1978). Karl Pearson and statistics: The social origins of scientific innovation. *Social Studies of Science 8*(1), pp. 3-34. Retrieved from http://www.jstor.org/stable/284855

Ochshorn, S. (2014, April 4). Test talk: Not a pretty picture, Lucy Calkins. *ECE PolicyMatters* RSS. Retrieved May 19, 2014, from http://www.ecepolicymatters.com/archives

O'Malley, K. (2012). *Standardized testing. What is it and how does it work?* - Pearson Research & Innovation Network. Retrieved June 13, 2014, from http://researchnetwork.pearson.com/college-career-success/standardized-testing-what-is-it-and-how-does-it-work

Our History. *Achieve.* (2014). Retrieved May 16, 2014, from http://www.achieve.org/history-achieve

Pink, D. (2009). Daniel Pink: *The puzzle of motivation.*. Retrieved July 14, 2014, from http://www.ted.com/talks/dan_pink_on_motivation

Popham, W.J. (2014). Criterion-referenced measurement: Half a century wasted? *Educational Leadership 71*(6), pp. 62-66. Retrieved from EBSCOhost Academic Search Premier Database.

powell, j. a. (2012). *Racing to justice: Transforming our conceptions of self and other to build an inclusive society.* Bloomington: Indiana University Press.

Praxis National Advisory Committees (NAC). (2014). Educational Testing Service. Retrieved from http://www.ets.org/praxis/states_agencies/about/content_current/nac

Praxis Study Companion. (2013). *English to Speakers of Other Languages.* Educational Testing Service. Retrieved from https://www.ets.org/praxis/prepare/materials

Praxis technical manual. (2010). Educational Testing Service. Retrieved from https://www.ets.org/praxis/institutions/resources

Preparing and credentialing the nation's teachers: The secretary's ninth report on teacher quality. (2013). U.S. Department of Education, Office of Postsecondary Education. Washington, D.C. Retrieved from http://www2.ed.gov/about/reports/annual/teachprep/index.html

Princeton Review. (2014). *What makes a good test?* Retrieved October 4, 2014, from http://www.princetonreview.com/corporate/what-makes-a-good-test.aspx

Proper use of Praxis Exams. (2000). Educational Testing Service. Retrieved from http://www.ets.org/praxis/about/bulletin/

psychometrics. (n.d.). *Merriam-Webster*. Retrieved June 13, 2014, from http://www.merriam-webster.com/dictionary/psychometrics

Rhoades, K. & Maddaus, G. (2003). Errors in standardized tests: A systemic problem. *National Board on Standardized Testing and Public Policy.*

Rosner, J. (2012). The SAT: Quantifying the unfairness behind the bubbles. In Soares, J.A. *SAT wars: The case for test-optional college admission.* (pp. 104-117). New York, NY: Teachers College Press.

Ross, F. (2005). Creating flexibility in teacher-certification policy to ensure quality and equity. *Maine Policy Review 14*(1). Retrieved from

http://digitalcommons.library.umaine.edu/cgi/viewcontent.cgi?article=1188&context=mpr

Scheunemann, J.D. and Slaughter, C. (1991). *Issues of test bias, item bias, and group differences and what to do while waiting for the answers.* Educational Testing Service. Retrieved from http://eric.ed.gov/?id=ED400294

Schuls, J. & Ritter, G. (2013) Teacher preparation not an either-or: Traditional and alternative routes to teaching are both good ideas—for certain subjects and grade levels. *Phi Delta Kappan 94*(7), 28-32. Retrieved from EBSCOhost Academic Search Premier Database.

Schmitt, A. & Dorans, N. (1988). *Differential item functioning for minority examinees on the SAT.* Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-88-32.pdf

Senate Bill 1 Highlights. (2009). *2009 Session of the Kentucky General Assembly.* Retrieved from http://education.ky.gov/comm/UL/Documents/SENATE%20BILL%201%20HIGHLIGHTS.pdf

Sleeter, C. (2013). *Power, teaching, and teacher education: Confronting injustice with critical research and action*. New York: Peter Lang.

Sleeter, C. (2011). *Professional development for culturally responsive and relationship-based pedagogy.* New York: Peter Lang.

Sleeter, C. (2001). Preparing teachers for culturally diverse schools: Research and the overwhelming presence of whiteness. *Journal of Teacher Education 52* (94).

Retrieved from http://www.unco.edu/cebs/diversity/pdfs/Sleeter_Preparing%20Teachers%20for%20Culturally%20Diverse%20Schools.pdf

Slover, L. (n.d.). *What's next?*. Partnership for Assessment of Readiness for College and Careers. Retrieved June 19, 2014, from http://www.parcconline.org/whats-next

Soares, J.A. (2012). For tests that are predictively powerful and without social prejudice. *Research & Practice in Assessment 7*, 5-11. Retrieved from http://www.rpajournal.com/dev/wp-content/uploads/2012/07/SF.pdf

Standards in your state. (2014). Common Core State Standards Initiative. Retrieved May 19, 2014, from http://www.corestandards.org/standards-in-your-state/

Standards for admission to educator preparation. (2014). 16 KAR 5:020. Kentucky Legislature. Retrieved from http://www.lrc.state.ky.us/kar/016/005/020.htm

Starnes, B. (2006). Montana's Indian Education for All: Toward an education worthy of American ideals. *Phi Delta Kappan, 88* (3), 184-192.

Starnes, B. (2006) What we don't know can hurt them: White teachers, Indian children. *Phi Delta Kappan (87)* 5, 384-392.

Stoskopf, A.(2002). Echoes of a forgotten past: Eugenics, testing, and education reform. *Educational Forum 66*(2), pp. 126-133. Retrieved from EBSCOhost Academic Search Premier Database.

Strategic Planning Committee, Berea College. (2011). *Being and becoming: Berea in the 21st century: The strategic plan for Berea College*. Retrieved from http://www.berea.edu/cisrk/files/2012/08/being-and-becoming-june-2011.pdf

State and country quick facts. United States Census Bureau. (2014). Retrieved from http://quickfacts.census.gov/qfd/states/21/2105842.html on May 10, 2014.

Strauss, V. (2013, January 11). Teachers refuse to give standardized test at Seattle high schools: Update. *Washington Post.* Retrieved May 19, 2014, from http://www.washingtonpost.com/blogs/answer-sheet/wp/2013/01/11/teachers-refuse-to-give-standardized-test-at-seattle-high-school/

Strauss, V. (2013, January 29). A tough critique of Common Core on early childhood education. *Washington Post.* Retrieved September 6, 2014, from http://www.washingtonpost.com/blogs/answer-sheet/wp/2013/01/29/a-tough-critique-of-common-core-on-early-childhood-education/

Tannenbaum, R. (2011). Setting standards on the Praxis Series Tests: A multi-state approach. *R & D Connections*. Educational Testing Service. Retrieved from https://www.ets.org/praxis/states_agencies/adoption_process/standard_setting_studies/multistate/

Tucker, M.S. (2011). *Surpassing Shanghai.* Cambridge, MA: Harvard Education Press.

Unbridled learning. (2014). Kentucky Department of Education. Retrieved from http://education.ky.gov/comm/UL/Pages/default.aspx

Understanding your Praxis scores. (2014-15). Educational Testing Service. Retrieved from https://www.ets.org/praxis/scores/understand

Varma, S. (n.d.). *Preliminary item statistics using point-biserial correlation and p-values*. Educational Data Systems, Inc. Retrieved from http://www.eddata.com/resources/publications/EDS_Point_Biserial.pdf

Villagas, A.M. & Irvine, J.J. (2010). *Diversifying the teaching force: An examination of major arguments.* Published online, Springer Science & Business Media LLC. Retrieved from http://www.montclair.edu/profilepages/media/439/user/Villegas_%26_Irvine--2010.pdf

Wakefield, D. (2003). Screening teacher candidates: Problems with high-stakes testing. *The Educational Forum 67*(4), pp 380-388. Retrieved from EBSCO host Academic Search Premier Database.

Walters-Parker, K. (March, 2014). *We're in this together: A model for shared P-12 & EPP accountability.* [PowerPoint slides]. Retrieved from http://caepnet.files.wordpress.com/2014/04/were_in_this_together.pdf

Wells, A.S. (2014). *Seeing past the "colorblind" myth of education policy: Addressing racial and ethnic inequality and supporting culturally diverse schools*. University of Colorado, Boulder: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/seeing-past-the-colorblind-myth

What Makes a Good Test. (2014). The Princeton Review. Retrieved June 13, 2014, from http://www.princetonreview.com/corporate/what-makes-a-good-test.aspx

Quality Performance Index (QPI). (2000-2013). Education Professional Standards Board (EPSB). Retrieved from http://www.kyepsb.net/documents/EduPrep/KEPP/qpi.asp

VITA

WENDY ZAGRAY WARREN

EDUCATION

May, 1983	Bachelor of Philosophy Miami University Oxford, Ohio
July, 2012	Master of Arts University of Montana Missoula, Montana
Pending	Doctor of Education Morehead State University Morehead, Kentucky

PROFESSIONAL EXPERIENCES

2012-2015	Professor Berea College Berea, Kentucky
1996-2012	Teacher Columbia Falls Schools Columbia Falls, Montana
1991-1996	Teacher Lake Mills Schools Lake Mills, Wisconsin
1987-1991	Teacher LaMotte School Bozeman, Montana
1983-1984	Teacher Oxford School for Young Children Oxford, Ohio

PUBLICATIONS

Starnes, B.A., Warren, W.Z., Warren, R.W. When a Door Closes… (2008). *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (4): 5-6.

Warren, W.Z. Becoming Upstanders: Humanizing Faces of Poverty Using Literature in a Middle School Classroom (2014) in *The Poverty and Education Reader*. Gorski, P. and Landsman, J., Editors.

Warren, W.Z. Remembering Our Humanity. (2010). *Journal of the Montana Writing Project* 4(2): 3-6.

Warren, W.Z. What if Simon Ortiz is Right? (2010). *Journal of the Montana Writing Project* 4(3): 11-12.

Warren, W.Z. Life in Four Directions. (2010). *Journal of the Montana Writing Project*, 4 (3): 3.

Warren, W.Z. Bearing Witness. (2010). *Journal of the Montana Writing Project 4* (1): 7-8.

Warren, W.Z. Since I've Seen You Last. (2009). *Journal of the Montana Writing Project 4* (1): 21.

Warren, W.Z. Poetic Justice. *Journal of the Montana Writing Project* (2009). 3 (4): 10.

Warren, W.Z. Programs in Action: New Voices on the E-Anthology: Creating a Space for Access Relevance and Diversity. (2008). *National Writing Project website.*

Warren, W.Z. Everybody Has a Piece of the Story: An Interview with Brenda Snider, Little Shell Chippewa/Metis Historian. (2008). *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (4): 7-9.

Warren, W.Z. I'm Learning As Fast As I Can: Federal Policies and Land Issues. (2008). *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (3): 32-34.

Warren, W.Z. Meet Jocelyn Acheson. (2008). *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (4): 22-26.

Warren, W.Z. My Guide. (2008). *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (4): 36-37.

Warren, W.Z. The Power of Story. (2008) *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (3): 5-6.

Warren, W.Z. Meteor. (2008). *Journal of the Montana Writing Project 2* (2): 16.

Warren, W.Z. Rising to the Power of Transformation. (2008). *Journal of the Montana Writing Project 2* (2): 20.

Warren, W.Z. Letter from New York City. (2008). *Journal of the Montana Writing Project 2* (2): 16.

Warren, W.Z. Fear Factor. (2007). *Journal of the Montana Writing Project 2* (1): 17.

Warren, W.Z. Photo-Essay: Montana Writing Project Browning Summer Institute. (2007). *Journal of the Montana Writing Project* 2 (1): 18-19.

Warren, W.Z. An Interview with Shannon Augere, Montana Legislator. (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (2): 17-20.

Warren, W.Z. I'm Learning As Fast As I Can: The 'First' Thanksgiving? (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for Al 3* (2): 31-34.

Warren, W.Z. Leo Bird: Native Science. (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (2): 25-27.

Warren, W.Z. Meet Brittany and Tyler Allen. (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (2): 7-10.

Warren, W.Z. Dressing Up for Halloween: Cultural Sensitivity in the Classroom. (2007).\3 (1): 25-27.

Warren, W.Z. I'm Learning As Fast As I Can: Perspective-Taking: The Columbus Story. (2007) *Full Circle: A Journal for Teachers Implementing Indian Education for All* 3 (1): 31-36.

Warren, W.Z. Place-Based History…From the Classroom of Lorrie Tatsey. (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (1): 13-16.

Warren, W.Z. The Bureau of Indian Affairs. (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for All 2* (2): 35-37.

Warren, W.Z. Federal Recognition. (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for All 2* (2): 31-34.

Warren, W.Z. I'm Learning As Fast As I Can: Selecting Accurate Materials. (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for All 2* (2): 15-17.

Warren, W.Z. I'm Learning As Fast As I Can: Names: Native American/American Indian/ Indian. (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for All 1* (2): 32-36.

Warren, W.Z. Honoring Sacred Places: a Conversation with Curly Bear Wagner. (2007) *Full Circle: A Journal for Teachers Implementing Indian Education for All 1*(2): 25-30.

Warren, W.Z. IEFA, Justice and Democracy. (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (2): 6.

Warren, W.Z. Supporting One Another. (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for All 3* (1): 6.

Warren, W.Z. Change. (2007). *Full Circle: A Journal for Teachers Implementing Indian Education for All 2* (2): 2.

Warren, W.Z. On the Importance of Partnerships: Coming to Know Better. (2007) *Full Circle: A Journal for Teachers Implementing Indian Education for All 1* (2): 6.

Warren, W.Z. I'm Learning As Fast As I Can: More Than Vegetables and Place Names: The Iroquois Confederacy. (2006). *Full Circle: A Journal for Teachers Implementing Indian Education for All 1* (1): 10-13.

Warren, W.Z. Editor's Column: Full Circle. *Full Circle: A Journal for Teachers Implementing Indian Education for All* (2006) 1 (1): 7.

Warren, W.Z. One Teacher's Story: Creating a New Future or Living Up to Our Own History? (2006). *Phi Delta Kappan 88* (3), 198-203.