ALTERNATIVE MEASURES TO ASSESS TEACHERS' PRACTICES, ATTITUDES,

AND KNOWLEDGE OF THE CREDE STANDARDS FOR EFFECTIVE PEDAGOGY

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAI'I AT MĀNOA IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

EDUCATIONAL PSYCHOLOGY

DECEMBER 2011

By
Alice K. Holt Taum

Dissertation Committee:

Lois A. Yamauchi, Chairperson
Richard J. Shavelson
Seongah Im
Paul R. Brandon
Ronald H. Heck

TABLE OF CONTENTS

Abstract

Fifteen K-12 teachers participated in a one-year university professional development course focused on the Center for Research on Education, Diversity and Excellence (CREDE) Five Standards for Effective Pedagogy. The standards are a set of instructional strategies that were developed through classroom research conducted by the Kamehameha Early Education Program. In general, the strategies serve as good practice; however evidence suggests that they are particularly effective when working in classrooms with culturally diverse learners. The standards are Joint Productive Activity, Language and Literacy Development, Contextualization, Complex Thinking, and Instructional Conversation. This study examined how teachers' changed their practices, attitudes, and understandings of the standards using four different measures. Three measures were surveys developed for this study. The fourth measure was an instrument used in previous CREDE research to code videotaped instruction. Reliability was established for the surveys and trend analyses using repeated measures ANOVA was used to examine teacher change during their participation in the professional development course. Results from the practices and attitudes surveys were compared with the classroom video analyses using a multitrait-multimethod matrix to determine the extent to which teachers' self-reported behaviors and attitudes aligned with video ratings. Convergent validity between the practices survey and video scores was established for Joint Productive Activity, but not for the other standards.

Four CREDE expert instructors who had between 7 and 35 years of teaching experience completed the three surveys on one occasion to establish the validity of the

instruments. Expert results using intraclass correlations were high for the practices and

attitudes surveys and acceptable for the vignette survey.

List of Tables

Attitudes Toward Teaching Survey

Video Observations using the Classroom Observation Rubric

List of Figures

Chapter 1

Introduction

From 1989 through 2000, approximately $286,600,000 was appropriated by

Congress to fund programs to increase access to education for Native Hawaiians (U. S.

Department of Interior, 2000), of which approximately $30 million was funded annually

through the Native Hawaiian Education Act (NHEA) (U. S. Department of Education,

2008). This study was part of a larger project that was funded by NHEA, which

prioritizes supporting organizations and projects that strive to "improve native Hawaiian

education statistics throughout the State of Hawai'i" (p. 2). The purpose of the larger

project was to provide professional development through a university course to teachers

who were working in schools with high concentrations of Native Hawaiian students. The

course focused on teachers learning about the Five CREDE (Center for Research on

Education, Diversity and Excellence) standards, a set of research-based instructional

strategies, which evidence suggests promote academic success for culturally diverse

learners (Doherty, Hilberg, Epaloose, & Tharp, 2002). Research on the CREDE

standards originated in the 1980s in early childhood classrooms for Native Hawaiian

students (Hao, 1988; Hao & Hartley-Forsyth, 1993; Tharp 1982; Tharp, Jordan, Speidel,

Au, Klein, Calkins, Sloat, & Gallimore, 1984). For this study, I developed three

measures to assess how K-12 teachers' instructional practices, attitudes about teaching

and learning, and knowledge of the CREDE standards changed over time while

participating in the professional development course.

There are different labels used for different types of workshops, programs, and

study groups designed to support teachers. There are induction, professional

development, coaching, and mentoring programs—however, they are often similar in their general intent to expose teachers to new ideas while providing support for them throughout the process of implementing these ideas. Despite the many variations of teacher professional development, there are fewer teacher support programs that specifically address effective instructional strategies for diverse student populations. Yet, because classrooms are becoming increasingly diverse, there is a need for more of these specialized programs (Santoro, 2009).

In this chapter, I describe the context relevant to the Hawai'i public school system during the 2008-2009 year, when this study was conducted. This includes information at the state, school, teacher and student levels. Following this brief introduction are the research questions for this study.

Need for CREDE Model

Approximately 47% of the teacher participants included in this study were originally from various states throughout the continental United States, and five of them lived in Hawai'i for five years or less. As described previously, all of the CREDE standards are useful instructional strategies; however, one in particular is particularly relevant to teachers who are new to the island: Contextualization. This standard emphasizes the importance of teachers' understanding their students' home and community lives so that they are better able to make connections between the content area subjects they are teaching and what students already know (Dalton, 1998, 2008; Tharp, Estrada, Dalton & Yamauchi, 2000). The Kahua Induction Program on the island of Hawai'i is an example of a pilot program developed for teachers who were new to the island that not only provided academic support, but contextualized community support

about the place and culture (Kahumoku & Kekahio, 2010). The course included three instructional components; regular meetings, coaching sessions, and video recordings of the teachers' classroom instruction.

Hawai'i Report Card

According to the 12th annual edition of *Education Week's Quality Counts 2008* (Swanson, 2008) report, Hawai'i's public schools were given an overall C (74.8%) grade for their performance. The grade was based on six performance and policy criteria: "chance for success; K-12 achievement; standards, assessments, and accountability; transitions and alignment; the teaching profession, and school finance" (p. 36). Only the teaching profession criteria will be discussed here, for which Hawai'i earned a grade of C+ (77.8%). There were fourteen categories calculated into the teaching profession overall grade, one of which included professional development. States were given credit if they established the following five policies: the state (a) had formal professional development standards, (b) financed professional development for all districts, (c) required districts/schools to set aside time for professional development, (d) required districts to align professional development with local priorities and goals, and (e) had standards for licensure of school administers. Hawai'i enacted all except for the requirement for districts and schools to set aside time for professional development.

Another category discussed in *Education Week's Quality Counts 2008* teaching profession criteria was *support for beginning teachers*, in which states were given credit if the following policies were established: (a) all new teachers were required to participate in a state-funded induction program; (b) all new teachers were required to participate in a state-funded mentoring program; (c) the state had standards for selecting,

training, and/or matching mentors, and (d) the state had a reduced-workload policy for first-year teachers. Compared with other states, which adopted all four policies at the time of this study, Hawai'i did not have any of these support systems in place for beginning teachers, suggesting that this was an area in which the state could benefit from investing greater resources.

These policy decisions—not requiring states to set aside time for professional development and not establishing a support system for beginning teachers—may need to be revisited. Changes are imminent on the educational horizon, from a federal level as seen with the Common Core Standards, to the classroom level, where classroom diversity is growing, which will have a direct impact on the country's educational system.

Hawai'i Schools

Hawai'i has a unitary statewide school system. In 2008, there were 412 K-12 schools, with 287 public schools (News Release, 2008) spread across fifteen complex areas throughout the state and 125 private schools (Hawai'i Private Schools, 2008). Complex areas are made up of a group of high schools and their middle and elementary feeder schools within a geographically designated area.

Hawai'i's educational system has expanded over the past 116 years, from employing 221 teachers (Benham & Heck, 1998) in 1892 to approximately 11,500 public schools teacher in 2007 (Hawai'i Educational Policy Center, 2008). An estimated 1,700 teachers in Hawai'i's were not fully licensed at the time of this study, or in terms of meeting one of the Hawai'i Department of Education's (HIDOE) objectives, were not "highly qualified."

Over the eight years prior to the study, public school enrollment steadily declined, dropping from 163,382 in 2000-2001, to 159,719 in 2007-2008. Enrollment in special education programs followed a similar trend, while the number of schools increased during the same time frame from 279 to 287 (Teacher Quality, 2008). Of Hawai'i's total population of 1.3 million, roughly 10% were identified as Native Hawaiian or Other Pacific Islanders (U. S. Census Bureau, 2010). Hawaiian and Part-Hawaiian students comprised nearly 28% of the population enrolled in the public schools—the greatest ethnic group represented across the eleven groups identified in the report—with 9% of teachers identifying themselves as Hawaiian or Part-Hawaiian (Hawai'i Department of Education, 2009).

In rural and remote regions where high enrollment of Hawaiian students was common, recruitment and retention rates for licensed teachers was a challenge (Hawai'i Department of Education, 2009). For example, parts of the Windward and Leeward complex areas, as well as some rural areas on the island of Hawai'i, were challenged because of the difficultly in maintaining fully licensed teachers and teachers who remained at the same school for five or more years.

The complex area with the lowest percentage of fully licensed teachers across the state, Nanakuli, a community where five participating project teachers taught, also had one of the greatest percentages of economically disadvantaged students and students enrolled in special education classes. The ratio of fully licensed teachers in Nanakuli was 78%, with 72% of students categorized as economically disadvantaged. The community with the second highest percentage of economically disadvantaged students in the State was Pahoa on the island of Hawai'i at 77% (Hawai'i Department of Education, 2009).

Additionally, this complex area had the lowest on-time graduation rates, 61%, in the entire state. According to the educational accountability results measuring Hawaiʻi's schools' Adequate Yearly Progress (AYP) from 2007 through 2009, required by NCLB, the percentage for all schools "meeting" the measures steadily declined (65% percent of schools in 2007, 42% in 2008, and 36% in 2009, respectively). At the same time, the percentage of schools "not meeting" the measures steadily increased (35% percent of schools in 2007, 58% in 2008 and 64% in 2009).

Teachers Working in Schools with High Concentrations of Hawaiian Students

As a result of the NCLB Act, the HIDOE identified two major objectives relating to teacher quality, the first of which addressed the issue of having all core academic subjects taught by a "highly qualified teacher," defined as someone who had at least a bachelor's degree, was fully certified by the State, and demonstrated knowledge in the subject area. The second objective proposed by the HIDOE was to reduce the rates of poor and minority students who were taught by inexperienced, unqualified, or out-of-field teachers (Teacher Quality, 2008). According to Darling-Hammond and Ball (1997), schools nationwide with larger concentrations of minorities were more likely to have teachers who (a) had fewer years of teaching experience, (b) had fewer years of teaching at a particular school, and (c) were not fully credentialed. Like other minorities who tend to experience more negative school outcomes, Hawaiian students enrolled in public schools often had teachers who were less prepared to teach students who came from diverse cultural backgrounds (Hawaiʻi Department of Education, 2009; Kanaiaupuni & Ishibashi, 2003).

Hawaiian Students

There are a number of indicators suggesting that Hawaiian students who are enrolled in public schools are at greater risk of negative educational outcomes compared to students from other ethnic groups. For many Hawaiian students, the complexities of their home lives do not easily intersect with school. Standardized achievement scores for Hawaiian students are among the lowest in the state (Kanaiaupuni & Ishibashi, 2003; Office of Hawaiian Affairs, 1994; Singh, 2011; Takenaka, 1995). In a recent study examining Hawai'i State Assessment (HSA) scores of Native Hawaiian and white students' in grades 3, 5, 8 and 10 using a hierarchical linear model, Singh (2011) found that students' individual performance in third grade was a predictor for future academic success in fifth, eighth, and 10th grade. Results were based on both Native Hawaiian and white students' comparable performances on the HSA test in third grade, but increasing gaps between the two ethnic groups in fifth and eighth grade. By 10th grade, scores between the two groups more than doubled, revealing Hawaiian students were the most disadvantaged.

Hawaiians also have higher dropout and grade retention rates and are over-represented in special education and under-represented in higher education (Kanaiaupuni & Ishibashi, 2003; Office of Hawaiian Affairs, 1994). Considering this combination of increased Native Hawaiian student enrollment in special education programs within the context of schools that have high concentrations of Native Hawaiian students, evidence suggests a need for greater efforts to provide experienced, fully qualified, and subject area content specialist teachers for these already challenged schools.

The three surveys developed for this study were intended to collect information from the teacher participants and examine how teachers' self reports compared with their videotaped lessons that were coded by two, trained observers. The surveys were developed using the same implementation criteria for the standards as they are described in the CREDE rubric that is used for rating videotaped classroom observations. Following are the six research questions address in this study.

Research Questions for This Study

(1)   To what extent did CREDE expert instructors' agree on the measurement of the CREDE standards for the practices, attitudes, and vignette surveys?

(2)   How did teachers' responses on the practices and attitudes surveys compare with results from teachers' video recorded instruction?

(3)   To what extent did teachers' instructional practices change over time based on a survey that was designed to measure teachers' performance using the CREDE standards in the classroom?

(4)   To what extent did teachers' attitudes change over time based on a survey that was designed to measure teachers' beliefs about teaching and learning using the CREDE standards in the classroom?

(5)   To what extent did teachers' instructional practices change over time based on results from their videotaped lessons using the Classroom Observation Rubric?

(6)   To what extent did teachers' reliably respond to the survey designed to measure teachers' understandings of the CREDE standards?

Chapter 2

Literature Review

Policy Issues Surrounding Education

In this section, some of the policy issues surrounding public education at both the federal and state levels will be highlighted, followed by a discussion on eight areas relating to professional development.

There is an increased emphasis on National and State Standards, which according to one California Superintendant, has changed how teachers were once prepared for their teaching careers that used to emphasize "how to teach," but not "what to teach" (Cash, 2007). This paradigm shift is evidenced by the adoption of the Common Core State Standards (CCSS) (2011) in 46 states, which focus on *what* teachers' teach. The CCSS define the knowledge and skills that students should have by the time they graduate from high school so that they are prepared for college courses, or workforce training programs.

There have been two major changes in education at both the federal and state levels since 2007-2008, when this study was conducted. First, our country is now under a different presidential administration. Second, Hawai'i was selected as one of nine states to be awarded a portion of $4.35 billion "Race to the Top" funding for school reform.

The largest request for increased funding for the Elementary and Secondary Education Act since the law was enacted in 1965 was approved to fund programs such as "Race to the Top and Investing in Innovation (also known as the 3-Is program), school turnarounds, charters, school safety and programs around preparing, retaining, and rewarding effective teachers and leaders" (Education Government, 2010). Some of the target areas for the budgeted $49.7 billion for the Department of Education's

discretionary programs for 2011-2012 include innovative teacher and leader reforms such as performance pay, teacher and leader recruitment and preparation, special education students, programs designed to promote a well-rounded education that support comprehensive literacy, Science Technology Engineering Mathematics, and other core subjects including history and arts and funding for charter and other autonomous schools.

The Race to the Top funding lent support to the No Child Left Behind (NCLB) Act, which required local educational agencies to collaborate with other agencies by providing them with technical assistance, including professional development[1] (NCLB Act, 2001). Because this study involved pubic schoolteachers in 2007-2008 and focused on how they changed while they were involved in a professional development course, it is important to re-visit what the NCLB Act stated about teacher professional development.

Among the 670 pages of guidelines describing the Act, there are 193 references to "professional development."[2] Passed through legislation in 2002, the NCLB Act was approved to address what some have described as a failing public school system (Berliner & Biddle, 1995; Tyack & Cuban, 1995). The intent of the NCLB Act was to:

> ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging State academic achievement standards and state academic assessments (NCLB Act, 2001).

---

[1] An example of these efforts, the Hawaiʻi Department of Education collaborated with another agency, Kamehameha Schools, in which an entire complex area participated in an induction program offering teachers academic and community support, which will be discussed further in a later section (Kahumoku & Kekahio, 2010).
[2] The phrase "teacher training" occurs 37 times in the NCLB Act document.

The first part of this statement involves equity, the importance for all students to have the opportunity to obtain an education—rich or poor, ethnic minority or majority, disabled or gifted. The Act purports an educational system supporting the acquisition of skills and knowledge for all students to become self sufficient, contributing members of society. However, the second part of this statement evokes concerns from some educators and researchers who argue that demanding proficiency on assessment measures, especially for minority learners, will not inherently close the achievement gap (Hanushek & Raymond, 2005; Mabry, Poole, Redmond, & Schultz, 2003; Valenzuela, Prieto, & Hamilton, 2005). According to Hanushek and Raymond (2005):

> given the generally lower achievement by minority groups, an implicit
> assumption is that accountability–as revealed through mandatory disaggregation
> of performance for racial and ethnic groups–will simultaneously close the large
> achievement racial/ethnic gaps along with improving all performance. (p. 4)

Although there is some evidence that the achievement gap was reduced for Hispanic students following legislation of the NCLB Act, this was not the case for African Americans (Hanushek & Raymond, 2005), or arguably, for Native Hawaiians (Kanaiaupuni & Ishibashi, 2003; Kekahio, 2007) whose educational history is both unique and complex (Benham & Heck, 1998).

As mentioned earlier, Hawai'i enacted all professional development policy criteria set forth in the *Education Week Quality* report (Swanson, 2008), except for the requirement for districts/schools to set aside time for professional development. The $75 million Race to the Top funding awarded to Hawai'i provided the opportunity for various agencies and institutions to assemble and develop a strategic plan to improve educational

outcomes. The results of these efforts led to a set of shared goals by educational leaders in the state.

There are five performance outcomes that have been adopted by the Governor, Board of Education, HIDOE and the University of Hawai'i, which include: (a) raising the overall achievement of K-12 students; (b) ensuring college- and career-readiness; (c) increasing higher education enrollment and completion rates; (d) increasing science, technology, engineering and mathematics (STEM) proficiency statewide and highly effective STEM instruction in Title I schools, and (e) ensuring equity and effectiveness by closing achievement gaps. The last goal aims to reduce the gap between groups and all students in state assessment scores, graduation rates, and college enrollment rates by 50 percent by the year 2018, at which time the gaps are projected to be eliminated (Hawai'i Race to the Top Application, 2010).

Teacher Professional Development

Over the past two decades, a considerable amount of research has been conducted on teacher professional development, teacher learning, and teacher change (Garet, Porter, Desimone, Birman, & Yoon, 2001). The literature on professional development is extensive and includes discussions covering a broad range of research topics. Some of the common themes focus on teacher quality (Darling-Hammond & Ball, 1997; Darling-Hammond, 2006a; Guskey, 2002); teacher change (Cohen & Barnes, 1993); pedagogy (Carpenter, Fennema, Peterson, Chiang, & Loef, 1989); pedagogical content knowledge (Borko & Putnam, 2007; Shulman, 1987; Shulman & Grossman, 1988; Shulman, 2004), and the decision-making processes, thoughts, behaviors, and judgments of teachers (Borko, 2004; Engstrom & Danielson, 2006; Sawchuk, 2010; Scher & O'Reilly, 2009;

Shavelson & Stern, 1981). Some of the literature focuses on methodological approaches

(Guskey, 1995; Spillane, 1999) and includes practical examples of how to promote

teacher change, whereas, others are more theoretical, grounded in the ideas and principles

supporting best practices (Elmore & McLaughlin, 1988).

To manage the vast amount of literature, I categorized articles into sub-categories:

(a) national reform; (b) core standards; (c) State Race to The Top funding; (d) high stakes

testing; (e) instrument development; (f) teacher coaching; (g) assessing teacher change,

(h) Indigenous education, and (i) culturally relevant pedagogy.

In the following section, I review the literature on eight areas that provide a

context for this study: (a) theoretical frameworks for professional development; (b)

professional development studies; (c) professional development focused on Hawaiian

learners; (d) culturally relevant tools and resources; (e) the CREDE standards for

Effective Pedagogy professional development model; (f) the validation of instruments;

(g) examples of studies that validated different measures of instruction; and (h) examples

of studies that validated the measure used in the current study.

Theoretical Frameworks for Professional Development

Because no two school districts, schools, classrooms, teachers, or student

populations are exactly alike, it is difficult to advocate a single professional development

model that will work well for all educators and students (Guskey, 1995). Teachers who

are at suburban schools likely have very different student populations than those in rural

schools and, although the students share the same role of being students, the contexts of

their environments are very different. Such differences include their parents' careers or

sources of income, the role that each parent plays within the family, the roles that older

siblings versus hired caretakers (Tharp, Estrada, Dalton, & Yamauchi, 2000) play within

the family, and the degree in which there exists a sense of community. This is

particularly true in Hawai'i, where schools are not only separated by land distance by

island, but by water, across islands.

Educational researchers generally agree that in order for professional

development to be successful, it must be viewed as a process, not an event (Garet, Porter,

Desimone, Birman, & Yoon, 2001; Guskey, 1995, 2002; Loucks-Horsley, Harding,

Arbuckle, Murray, Dubea & Williams, 1987; Tillema & Imants, 2007). Lemlech,

Hertzog-Foliart, and Hackl (2005) suggest that the "learning process itself is more

valuable than the end result; that the planned and unplanned learning experiences

empower the student as problem-solver and inventor/creator in his own right" (p. 165).[3]

However, measuring a process and the outcomes of the process is much more

complicated than measuring an isolated event, as it involves a series of actions that are

working toward a particular aim. Considering the complexities of any given classroom

(Milner, 2011), it is difficult to isolate specific variables that can be attributed to teacher

change (Guskey, 2002). For example, teachers' content knowledge, instructional style,

and years of teaching experience (Darling-Hammond, 2000; Shulman, 1987), combined

with students' prior knowledge, learning styles, developmental levels and personalities,

are all variables that influence instruction. As Shulman (2004) asserts, "an instructional

method simply cannot be conceived in isolation from the teacher who employs it, the

learning climate of the classroom, and school in which it is deployed" (p. 150). Given

---

[3] Professional development is recognized as being a process, just as Berliner (1987) describes the process of student learning at the turn of the 20th century, emphasizing that it is the process, not the outcome that determined good teaching.

this dynamic environment, it is important to explore studies that have examined professional development activities, the processes that were involved, and the outcomes that were measured.

Borko and Putnam (1995) argue that teachers' knowledge and beliefs should be the central focus for reforming teaching practices. Building upon categories proposed earlier by Shulman (1987; Shulman & Grossman, 1988; Shulman, 2004), they suggest that there are three knowledge domains that are important elements of good instruction: (a) general pedagogical knowledge, (b) subject-matter knowledge, and (c) pedagogical content knowledge.

The first domain, *general pedagogical knowledge* encompasses the knowledge and beliefs that teachers have about teaching and learning and strategies that best promote effective learning environments. General pedagogical knowledge includes classroom management strategies and the means for organizing a productive learning environment. The second domain involves *subject-matter knowledge*, which requires a balance between understanding content and instructional practices. The latter involves teaching for understanding using materials and programs that facilitate the learning process. The last domain, *pedagogical content knowledge*, involves understanding a subject area and the topics and issues that surround it and how this information can be organized and represented for teaching (Borko & Putnam, 1995; Putnam & Borko, 2000). This type of knowledge takes time to acquire through experiences because it requires that teachers not only understand the subject matter they are teaching, but also that they are able to organize how information will be presented to students with varying interests and abilities and respond to students based on their needs. Each of the three knowledge

domains is embedded in many different professional development models. Garet et al. (2001) support this notion that changes in teachers knowledge and beliefs improve their instructional practices.

Another interest area among educational researchers involves studies focused on examining teacher quality, considered to be one of the most important factors in explaining differences in student performance (Darling-Hammond & Ball, 1997). The research measuring the effects of teacher quality and training on student learning commonly focuses on teachers' education and experience levels, licensure exams, and teacher certification; however, the results of these studies are not consistent. Goldhaber (2002) suggests that although we know that good teachers make a difference in student learning outcomes, it is unclear "what makes for a good teacher" (p. 55). Whereas some researchers have determined that the results from these types of studies are inconclusive in terms of teachers' direct impact on student learning (Goldhaber & Brewer, 1997; Hanushek, 1986), others argue that the evidence is well documented (Darling-Hammond & Ball, 1997).

In 1996, the U.S. Department of Education (USDOE) created a Professional Development Team to assist with preparing and supporting educators to succeed in teaching their students. Secretary Richard W. Riley established the team to:

> Examine the best available research and exemplary practices related to
> professional development, and to summarize the lessons learned from this
> knowledge base in the form of principles that might inform practitioners and
> policymakers across the country and guide the Department's efforts in the area of
> professional development (U. S. Department of Education, 1996).

The mission of the team was to place professional development as part of a system-wide effort to improve and integrate teacher recruitment and preparation, including licensing, induction and ongoing support. The team agreed on 10 principles that promote effective professional development, built around collaboration between the teacher, school and community (see Table 1). The principles are broad in nature and are not intended to be prescriptive; rather they serve as a framework to guide those who develop and offer professional development opportunities so that they are successful in preparing and supporting teachers' professional advancement. The principles described by the team do not reference the importance of teachers' learning about students' different cultural beliefs and practices, their home environments, learning styles, or address disparities among students from disadvantaged or impoverished families and communities, which was addressed in the professional development of the current study.

Professional Development Studies

Large Scale Studies. There are few rigorous large-scale studies that have examined the effectiveness of teacher professional development. The Institute of Sciences (IES) in the U.S. Department of Education commissioned a study by the American Institutes for Research (AIR) and Manpower Demonstration Research Corporation (Quint, 2011). Out of hundreds of studies that were reviewed in the literature, five used random assignment designs that yielded causal effects of professional development on student outcomes. All five studies involved elementary grades and collectively studied 14 measures of student achievement in reading and mathematics. For all 14 measures, students whose teachers participated in professional development had

Table 1

Professional Development Principles and Participant Targets[4]

Principle	Teacher	School	Community
Professional development should focus on teachers as central to student learning, while including other members of the school community.	√	√	√
Professional development should focus on individual, collegial, and organizational improvement.	√	√	
Professional development should respect and nurture the intellectual and leadership capacity of teachers, principals, and others in the school community.	√	√	√
Professional development should reflect the best available research and practice in teaching, learning, and leadership.	√	√	
Professional development should enable teachers to develop further expertise in subject content, teaching strategies, uses of technologies, and other essential elements in teaching to high standards.	√		
Professional development should promote continuous inquiry and improvement embedded in the daily life of schools.	√	√	
Professional development should include collaborative planning by those who will participate in and facilitate that development.	√	√	
Professional development should require substantial time and other resources.		√	
Professional development should be driven by a coherent long-term plan.		√	
Professional development should be evaluated on the basis of its impact on teacher effectiveness and student learning, and this assessment should guide subsequent professional development efforts.		√	

[4] (U. S. Department of Education, 1996)

higher test scores than students whose teachers had not been involved in any training. However, in nine of the 14 cases, the differences were not statistically significant.

Building upon these research studies, AIR and MDRC conducted two studies including almost 170 elementary and middle schools, which examined the causal role between professional development and improving student outcomes. In the first study involving reading, 270 randomly assigned second grade teachers participated in professional development where they learned about essential components of early reading instruction such as phonemic awareness, phonics, fluency, vocabulary and comprehension. In the second study, professional development for 287 randomly assigned seventh grade mathematics teachers focused on rational numbers—fractions, decimals, ratios, rates, proportions, and percentages—with an emphasis on students understanding of rational numbers and computational skills. Unlike many studies where professional development is an important aspect of another intervention, such as a new curriculum, in both the reading and mathematics studies, the professional development was the intervention tested.

For the reading study, greater teacher knowledge was associated with higher student test scores. First year results of the math study showed no significant relationships between teacher knowledge or practice and student achievement. However, results from the second year resembled that of the reading study, with higher levels of teacher knowledge associated with higher student test scores.

In a national study involving more than 1,000 mathematics and science teachers, Garet et al. (2001) identified three "structural features" and three "core features" which teachers reported positively influenced their knowledge and instructional practices. The

three structural features were: (a) the form of the activity, whether a workshop or a study group; (b) the degree of commonalities across members of the group (same school, grade, or subject area), and (c) the duration of the activity, including the number of contact hours and time span. The core features were: (a) the degree of focus on content to improve teachers knowledge; (b) the extent of opportunities for active learning, and (c) the degree to which activities promoted a coherent teacher learning program that was consistent with teachers' goals and aligned with state standards and assessments. Teachers were surveyed on each of these features and results of all standard path coefficients were statistically significant at the .05 level. For example, when examining the relationship between reform types of professional development such as study groups involving coaching and mentoring, with traditional types of professional development such as structured workshops occurring outside a teacher's own classroom, the form of the activity had an important influence on the duration of the activity; longer periods tended to involve greater number of contact hours than traditional activities. There were also significant path analysis correlations between time span and contact hours, such that both positively influenced opportunities for active learning and coherence. Longer activities tended to afford more opportunities for engagement, or active learning and were more likely to support teachers' efforts in reaching identified goals.

Case Studies. In addition to large scale studies that explored teacher professional development, following are two examples of professional development schools (PDS) that offered sustained learning opportunities by partnering new teachers with veterans, university adjuncts and teacher leaders. Professional development schools include large schools in restructuring and small, new schools where student teachers are placed. There

are three goals associated with PDSs: to (a) provide for continued development and professional growth of experienced teachers; (b) provide a context for thinking and reinventing schools so they can build and sustain the best educational practices, and (c) contribute to preservice teacher education. The number of PDSs is increasing, especially in schools serving diverse student populations (Darling-Hammond, 2005; Lemlech et al., 2005).

One study investigating a PDS was Wells Junior High School, a school that partnered with the University of Southern Maine to develop the Extended Teacher Education Program for preservice teachers. The program consisted of 30-credit hours, embedded within a 1-year graduate intern program (Miller & Silvernail, 2005). This case study involved 15 interns and was framed around a set of principles and practices including videotapes for professional development. The process began with videotaping teacher interns while they were teaching a lesson, which was followed by discussions with experienced teachers. Initially, the intern, supervisor—a college professor—and cooperating teacher—coach—all viewed the videotaped lesson together, but only the intern commented and was encouraged by both teacher and supervisor to critique the lesson. Next, the intern met with the cooperating teacher, and the supervisor videotaped their discussion. After the second videotaping, the intern was excused and the cooperating teacher and supervisor viewed the coaching session together. The final step involved the cooperating teacher, who commented on and analyzed the coaching experience, discussing the video with her supervisor. The videotaping process that was developed primarily to assist new teachers proved to be equally powerful as a professional development tool for the Wells faculty. The veteran teachers' believed that

the opportunity to meet with the supervisor to critique the intern-coaching session provided a valuable opportunity for reflection on the teaching and learning process.

The second study involved a partnership between an elementary school and university in Los Angeles, California, that collaborated to create a PDS. There were nine teachers, a media resource teacher and the principal who participated in the project (Lemlech et al., 2005). These teachers' discussed the issues encountered from the different perspectives of the university and school-based participants that involved restructuring. Among all of the issues that were discussed, two included teacher professionalism and decision-making. The goals of the project were to improve instructional programs, provide a supportive educational environment for student teachers, and promote inquiry on teaching.

The case study was framed as an inquiry, a set of questions that were driving a restructuring project that reported participant perceptions about the PDS. Questions about how teachers and teacher educators would share the responsibility for structuring a professional development school and expected outcomes for both the project teachers and university teacher educators were discussed. Participants were also interested in changes in teaching practices, beliefs and whether an increased sense of professionalism would occur. Outcomes from the study were positive and teachers and teacher educators supported working together, as long as clear parameters were defined. Participants also appreciated the sense of shared collegiality. Their partnership enabled them to focus on curriculum and instruction to improve student performance, linking what preservice teachers studied to what was reinforced in practice, while highlighting the importance of teachers' understandings that learning to teach was a process. Finally, through shared

readings, discussions, presentations, and reflections, teachers were able to reassess their beliefs about their instructional practices.

These two case studies are examples of PDSs, which offer systems that promote teacher change for both novice and experienced teachers through the collaborative efforts of pre- and in-service school and university partnerships.

Professional Development Focused on Hawaiian Learners

The literature surrounding multicultural education, examining both teacher professional development programs and culturally diverse learners is helpful in identifying the different terms and phrases that are commonly used (Reyhner, Gilbert, & Lockard, 2011; Santoro, 2009; Thomas & Kearney, 2008). Culturally relevant pedagogy (Ladson-Billings, 1995a, 1995b; Milner, 2011), culturally responsive teaching (Farmer, Hauk, & Neumann, 2005), culture based education (Demmert & Towner, 2003), culturally relevant strategies (Kahumoku & Kekahio, 2010), cultural compatibility (Vogt, Jordan, & Tharp, 1987), culturally responsive pedagogies (Santoro, 2009), and culturally responsive professional development (Farmer, Hauk & Neumann, 2005) each convey the relevance of culture in the context of teaching and learning.

As noted earlier, the literature shows that high-need students are often taught by educators with little or no teaching experience or who have limited backgrounds and training in special education or theories on multicultural instructional practices (Jorgensen, Grootenboer, Niesche, & Lerman, 2010).

In general, teachers tend to instruct based on how they learned themselves as students, making it unlikely for one-size-fits-all, or isolated teacher professional development trainings to significantly impact teachers' ideas and behaviors (Darling-

Hammond, 2006a; Shavelson & Stern, 1981). This is particularly true for novice

teachers (Yamauchi, 2003; Yamauchi, Wyatt, & Taum, 2005) whose teaching repertoires

have not yet been established. In a research study on the Leeward coast of Oʻahu,

teachers participated in a study group in which they learned about the CREDE standards

(Yamauchi, 2003). Data collected through surveys and teacher interviews revealed that

some teachers believed that their instructional approach was similar to their own learning

style. By the end of the professional development study, one teacher who was adamant at

the beginning about expecting students to learn the way that he did, transformed his

beliefs and began focusing on using strategies that supported his students' varied learning

styles. For teachers new to Hawaiʻi who are unfamiliar with its history and culture and

who know little about the community they are entering into, professional development,

mentoring, and induction programs are particularly important.

Ladson-Billings (1995a, 1995b) emphasizes the value of incorporating students'

cultures into instructional practices to reduce the effects of what she calls the dominant

culture. This involves using culturally relevant instructional strategies that will

encourage students to recognize differences between what they are learning in the

classroom and what they already know from other experiences outside the classroom.

According to Milner (2011) there is a difference between culturally relevant *pedagogy*

and culturally relevant *teaching*. The former is often used to discuss or describe the

theory of culturally relevant teaching, while the latter is used to describe the practice of

the theory. Teachers who support instruction through "culturally relevant learning

contexts" also view "culture as an asset, not a detriment" to students' success. In the

Hawaiian culture, there are many ʻolelo noʻeau (Pukuʻi, 1983), wise sayings depicting

how children learn largely through observation, listening, and doing as reflected in the Hawaiian sayings, "Nānā ka maka; hoʻolohe ka pepeiao; paʻa ka waha" (p. 248), which translates into "Observe with the eyes; listen with the ears; shut the mouth," "I ka nānā no a ʻike," "by observing, one learns" (p. 129), or "Nānā ka maka; hana ka lima," "observe with the eyes; work with the hands" (p. 247). Similarly, some non-English speaking students may not be as skilled as native English speakers in asking questions in English or using language to the same extent in the learning process. Therefore, it is important for teachers to be cognizant of students learning styles and how relevant pedagogical practices may better align with students' diverse home and school learning environments.

Culturally Relevant Tools and Resources

Demmert, Grissmer, and Towner (2006) described six criteria that support culture-based education (CBE) learning environments for Native learners: (a) recognition and use of Native American, American Indian, Alaska Native and Native Hawaiian languages as the language of instruction; (b) pedagogy that stresses traditional cultural adult/child interactions; (c) congruent teaching strategies between traditional cultural and contemporary ways of knowing and learning curriculum; (d) curriculum based on traditional cultures, recognizing Native spirituality and places, (e) strong Native community participation and (f) knowledge and use of the social and political mores of the community. As part of the Hawaiʻi Indigenous Education Teaching framework, Kanaʻiaupuni and Kawaiʻaeʻa (2008) purport a similar framework for operationalizing culture-based instruction, which include five components: recognizing and using native or heritage language; family and community involvement; making learning meaningful

and culturally relevant; structuring the learning environment in culturally appropriate ways, and data collection to measure students' progress. To measure teachers' use of these five instructional components, the Hawaiian Indigenous Education Rubric was developed.

Another familiar resource for teachers who are working with culturally diverse students in Hawai'i is *Na Honua Mauli Ola (NHMO): Hawai'i Guidelines for a Culturally Healthy and Responsive Learning Environment* (Native Hawaiian Education Council, 2002). Although not explicitly used as a teacher professional development tool, NHMO offers guidelines for learners, educators, schools, families and communities and is a resource that can be particularly useful for teachers who are new to the islands. There have not been any formal studies conducted using NHMO, however, there are educational programs that utilized this cultural resource, such as the Maui Fishpond which used the guidelines to teach students about their environment, and the Hawai'i Dryland Forest, a non-profit organization that offered NHMO as a supplement to other curricula.

Similarly, Moenaha is a culture-based instructional method that integrates teaching and learning strategies from McCarthy's (1981, 1985) 4MAT instructional cycle, with Hawaiian culturally relevant practices. McCarthy's method is based on research from various sources including education, neuroscience, neuropsychology, and behavioral science. The parts of the 4MAT include adaptations, meanings, skills, and concepts. Moenaha offers teachers opportunities for collegial sharing, access to coaching and mentoring, activities promoting reflection and a culminating portfolio activity (Kahumoku & Kekahio, 2010). The training includes assistance in developing culturally

relevant lesson plans. Teachers are asked to keep journals describing "moments that touch" them, when a realization hits that something is not working and needs to change. Teachers consider the facts surrounding why a situation happened, what the implications were, and most importantly, what behaviors the teacher needs to change so that the situation does not happen again. This provides them with the opportunity to first reflect upon their behaviors, understand how the behaviors impact their students and subsequently change them.

The Kahua Induction Program involved a partnership between the HIDOE, Kamehameha Schools, and various other community organizations to provide first-year teachers with mentoring, academic, social-emotional, and place-focused support (Kahumoku & Kekahio, 2010). The Kahua Program identified three goals and evaluation questions to measure the programs success. The goals were to (a) increase retention rates of new HIDOE teachers; (b) increase the number of new HIDOE teachers' understandings and implementation of cultural, place-based, indigenous practices, and (c) assist new HIDOE teachers in acquiring strategies to strengthen relationships with their students' parents, guardians, and other family and community members. To measure the outcomes, the Kahua Program examined (a) the effects of the program on Complex area teacher retention rates; (b) changes in knowledge, attitudes and skills in the area of culture-based education, and (c) the extent to which participants acquired successful strategies to strengthen relationships with families and communities. Upon completion of the induction program, teachers were more likely to continue teaching in year two, had clearer understandings of culture-based education, and established stronger relationships with families within their communities (Kahumoku & Kekahio, 2010).

In a pilot study beginning in spring 2007, 35 teachers elected to participate in the yearlong Kahua Program on the island of Hawai'i. Approximately 53 percent of participants were from the continental United States; roughly 48 percent of teachers lived in Hawai'i for five years or less, and nearly 89 percent were raised outside of the area where they were teaching. Researchers used multiple data sources including retrospective teacher surveys, individual event surveys—through a Likert-scaled and short answer questionnaire—and DVD recordings. The retrospective survey was designed to measure changes in participants' knowledge and self-perceptions surrounding their teaching beliefs. The individual survey was a modified version of the measure that was used in the Hawaiian Cultural Influences in Education study, to collect information from participants on the impact of the Kahua Program. Results indicated improved retention rates and increased knowledge, attitudes, and skills toward culture based-education. Paired *t*-test results reflected an increase in teachers' knowledge, attitudes and skills for all constructs (context, content, knowledge, family, language, and assessment).

A recent study examined the relationship between culture-based education (CBE) and student outcomes (Kana'iaupuni, Ledward, & Jensen, 2010; Thomas & Heck, 2009). In the study, researchers defined CBE as part of a holistic and comprehensive application that includes educational approaches which stem from a particular worldview (Demmert & Towner, 2003), in this case, the Hawaiian culture. More specifically, CBE grounds instruction and student learning in the values, beliefs, practices, and language of a particular community. There were five components to the CBE model, including language, family and community, content, context and assessment. Through the collaborative efforts of Kamehameha Schools, the HIDOE, several charter schools, and

Hawaiian organizations in the state, this study examined cultural influences in education for Hawaiians (Hawaiian Cultural Influences in Education, HCIE), with the goal of acquiring a better understanding of how schools can provide engaging and relevant educational experiences for Hawaiʻi's children.

To measure the five CBE components, the Hawaiian Indigenous Education Rubric was administered to teachers and their students using a four-point scale (none, emerging, developing, enacting) to determine the degree to which their classrooms measured the following: (a) heritage language in teaching; (b) ʻohana (family) and community involvement; (c) culture and place-based content; (d) context, and (e) data and accountability.

Using a three-level hierarchical linear model, Thomas and Heck (2009) analyzed survey data from approximately 11,000 student/teacher records. Results indicated that CBE was an important predictor within schools where the five components—language, family and community, content, context and assessment—were present.

The CREDE Standards for Effective Pedagogy Professional Development Model

The professional development model used in this study came from CREDE, previously at the University of California, Santa Cruz and now at the University of California, Berkeley, its predecessors, the National Center for Research on Cultural Diversity & Second Language Learning, and the Kamehameha Early Education Program (KEEP). There are seven CREDE standards. Standards 1 through 5 are considered to be generic principles identified as useful for all groups of students. Standards 6 (Modeling) and 7 (Student Directed Activity), developed specifically for indigenous students (Tharp, 2006), were not included in this study. The five teaching standards are believed to

"maximize excellence of achievement among students generally considered to be at risk for educational failure: the poor, the culturally and linguistically diverse, the isolated of the deep ghetto and the deep hills, the excluded, and the unfairly treated" (Tharp, Estrada, Dalton, & Yamauchi, 2000, p. 17). The five principles, or Five Standards for Effective Pedagogy, are believed to be effective across all subject areas, and will be referred to throughout this paper as the CREDE standards.

The CREDE standards are grounded in Vygotsky's (1978) sociocultural theory, in which learners create their own knowledge through their interactions with materials and activities within a social setting. According to Vygotsky, there are two developmental levels of learning, a child's actual developmental level and what a child can do with assistance. The space between the two performance levels is considered a child's zone of proximal development. Vygotsky believed that learning takes place when children are working in their zone of proximal development on tasks that they cannot do alone but can do with the assistance from more capable peers or adults. Vygotsky also believed that higher mental functioning results from conversations between individuals before it exists within an individual. Although all of the CREDE standards fit within Vygotsky's sociocultural framework, the Joint Productive Activity and Instructional Conversation standards may promote the greatest opportunities for student collaboration and also student-teacher discussions, helping students move from their actual performance levels toward what they are capable of doing with assistance. This combination of small group conversations between students and their teacher are consistent with traditional teachings for Native Hawaiians, which were often determined by a parent or elder based on a child's ability (Benham & Heck, 1998).

Kamehameha Early Education Program. The CREDE standards are

instructional practices, which largely derive from 15 years of research and development

conducted at KEEP. The purpose of KEEP was to improve literacy learning using an

alternative pedagogy that encouraged teachers to incorporate meaningful social and

linguistic instructional activities from the Hawaiian culture (Tharp, Jordan, Speidel, Au,

Klein, Calkins, Sloat, & Gallimore, 2007; Tharp & Dalton, 2007). The program was

organized as a research-and-demonstration school to examine concerns about Native

Hawaiian students who were struggling in school, based on school achievement data

placing them as the "lowest achievers of all state ethnic groups" (Tharp et al., 2000). In

KEEP, students were taught that the teacher played an adult role that many students were

unfamiliar with, because in their home environment, older siblings often played the role

of adult. Since KEEP, CREDE researchers further developed the standards. The CREDE

standards are:

- Standard 1—Teachers and Students Producing Together. (Joint Productive

    Activity).

- Standard 2—Developing Language and Literacy across the Curriculum.

    (Language and Literacy Development).

- Standard 3—Connecting School to Students' Lives. (Contextualization).

- Standard 4—Cognitively Complex Activities. (Complex Thinking).

- Standard 5—Teaching Through Conversation. (Instructional Conversation).

    Some of the research involving the CREDE standards has focused on the positive

effects that the teaching strategies have made on student learning (Doherty et al., 2002).

For example, in multiple studies by Saunders and Goldenberg (e.g., Saunders, O'Brien,

Lennon, & McLean, 1998; Saunders & Goldenberg, 1999) there was a positive

relationship between Instructional Conversation and students' reading comprehension.

The same positive effects were observed when Instructional Conversation was used in

combination with Contextualization, particularly for English Language Learners.

Doherty et al., (2003) examined the use of Joint Productive Activity for Latino English

Language Learners during language arts instruction and found that students' reading

strategies improved. Hilberg, Tharp, and DeGeest (2000) randomly assigned eighth

grade American Indian students into two groups where they spent one week learning

about fractions, decimals and percentages. One group had an instructor teaching

mathematics using the CREDE standards, and the other had a non-CREDE instructor.

Results showed that students who were in the CREDE classroom retained information

that they learned two weeks after the study and performed better on tests of mathematics

conceptual learning. In repeated studies with Grade 1 and 4 students, Estrada (2004)

found that students who were in classrooms with teachers who were strong in

implementing the CREDE standards tended to be reading at grade level, compared with

students whose teachers used the standards to a lesser degree.

The Validation of Instruments

According to The Standards for Educational and Psychological Testing, "validity

refers to the degree to which evidence and theory support the interpretations of test scores

entailed by proposed uses of tests" (AERA, APA, &NCME, 1999, p. 9). To assure a

sound validity argument, there are 24 standards to consider which offer evidence

supporting the quality of a particular test. For example, how a test is constructed,

reliability results, and appropriate administration and scoring are all important factors

when interpreting scores. Similarly, Messick (1989; 1995) defines validity as a judgment

based on "empirical evidence and theoretical rationales" (p. 13) supporting inferences

and actions based on test scores. One of the most basic questions in construct validation

asks to what extent a test measures what it is intended to measure? Constructs provide an

organized method for interpreting observed behaviors (Messick, 1989), in this case, the

CREDE standards. Furthermore, interpreting a construct involves "an interplay of

construct definition, instrument development, and data collection" (Shavelson, Hubner, &

Stanton, 1976, p. 415).

Surveys are a common measure used to ask teachers to report what they are doing

in the classroom (Little, Goe, & Bell, 2009; Stumbo & McWalters, 2011). Self-report

data can tap into a teacher's knowledge, thought processes and beliefs (Darling-

Hammond, 2006b; Little et al., 2009). Self-report data is also useful for formative

assessment purposes and promoting teacher self-reflection. Surveys can be especially

helpful when used with other measures such as classroom observation, student ratings, or

portfolios. However, with so much weight placed upon teacher performance, it is

imperative that measures of teacher performance are valid and reliable, credible, and fair

(Stumbo & McWalters, 2011).

Most books or articles that discuss methodology emphasize that research

questions should drive methods (Desimone, 2009). For more "nuanced constructs"

(Desimone, 2009, p. 190) such as critical reflection or quality of instruction, observations

or interviews are often most appropriate. Observations allow for detailed distinctions

such as whether a teacher is implementing an instructional strategy through a perfunctory

or deliberate manner. Observations are often considered the "gold standard" for

measuring classroom practice, however, they are also time consuming and expensive

(Stecher, Le, Hamilton, Ryan, Robyn, & Lockwood, 2006).

In contrast, surveys are used to ask about specific behaviors, such as those

experienced throughout the duration of a professional development program. It is

important that survey questions align as closely as possible with the implementation

models of curricula in order to yield more program-specific information, as opposed to

general information about teacher beliefs, knowledge, or instructional strategies (Penuel,

Fishman, Yamaguchi, & Gallagher, 2007). Surveys are often cost-effective and can

capture information about classroom practice that covers a longer timeframe than is

typically feasible using observations (Stecher et al., 2006). The accuracy of self-reported

responses depends on the reliability and validity of the measure (Mullens & Kasprzyk,

1996). Mayer (1999) describes three cautions for using self-reported data. First, learning

environments are complex and cannot be adequately represented by a single survey.

Second, unfamiliar words or phrases, or ambiguous concepts may be difficult for some

individuals to appropriately respond. Finally, respondents may feel pressured to provide

socially desirable responses to sensitive questions and/or concepts that the question

represents.

A less common method of data collection is vignettes. Vignettes are

contextualized descriptions of classroom situations, which offer an alternative strategy

for examining instructional practice (Stecher et al., 2006). Finch (1987) defines vignettes

as "short stories about hypothetical characters in specific circumstances to whose

situation the interviewee is invited to respond" (p. 106). Because of the contextualized

nature of vignettes, they may be viewed by teachers as more realistic examples and

situations of classroom events. In addition, they are standardized so that teacher responses can be aggregated and compared (Ruiz-Primo & Li, 2002). However, these real-life prompts offer no guarantee that teachers' responses are accurate reflections of their own behaviors (Darling-Hammond, 2006b).

Examples of Studies Validating Different Measures of Instruction

Many professional development studies have relied on teacher self-reports by using a survey or an observation to measure teacher learning and teacher change (Sato, Wei, & Darling-Hammond, 2008). In the following section, I describe three studies that used a combination of these measures.

Study One. This study examined strategies to directly measure changes in science and mathematics assessment practices among National Board Certification candidates and other teachers who were not nationally certified (Sato et al., 2008). The researchers explored a variety of data sources using six dimensions of formative assessment: (a) views and uses of assessment; (b) range, quality, and coherence of assessment methods; (c) clarity and appropriateness of goals and expectations for learning; (d) opportunities for self-assessment; (e) modifications to teaching based on assessment information, and (f) quality and appropriateness of feedback to students. Classroom evidence of assessment practices using data—classroom artifacts, student surveys, teacher surveys, and teacher interviews—collected throughout the study, were aggregated into individual teacher data packets. In addition to using the video recordings to examine teacher assessment practices, part of the video recording process involved teachers providing written responses to questions about the videotaped lessons including the sources of the curriculum materials, intended learning goals, planned assessment

techniques, and their assessment on the success of the lesson. Teacher and student surveys were also planned to supplement the other data.

At the end of each year of the study, teachers were asked to complete an online survey that addressed their demographic characteristics, education, training, teaching contexts and participation in professional development opportunities. Additionally, teachers were asked to rate a list of 24 classroom-teaching practices based on how much emphasis they gave to various learning objectives and to report the frequency in which they implemented a variety of practices and assessments. Unfortunately, due to low return rates for the teacher surveys, those data were not reported in the study.

The student survey asked students to report on frequencies—of their participation, of their use of measurement tools, manipulatives, calculators, computers—and the extent to which their teachers' emphasized 33 different classroom activities. There were also four "free-response" questions that asked students to describe (a) how they knew what their teacher wanted them to learn in the class (goals for learning), (b) the kinds of discussions they had with their teacher (oral feedback), (c) the kinds of written comments their teacher made about their work (written feedback), and (d) whether they got to evaluate their own work or the work of other students (self- and peer-assessment).

Analyses of the free-response answers was thorough, beginning with a coding process that resulted in 71 categories—22 for goals, 25 for oral feedback, 22 for written feedback, and 2 for self- and peer-assessment. Further analyses were conducted on responses where more than 5% of categories were represented.

To establish inter-rater reliability, multiple raters independently scored a portion of the data packets across the three-year study. Results showed that 53% of overall

scores were exact matches (within .25 of a point) and 94% of scores were within 1 point

on a 5-point rubric (Sato et al., 2008). Spearman-Brown reliability estimates were

calculated for all data packets across the three years and showed reliability of .96 with

sub-score reliabilities ranging from .93 to .74. The results of the ratings indicated

consistent trends across the multiple data sources, illustrating an increase in formative

assessment practices for National Board Certified candidates who were involved in the

certification process.

Study Two. Darling-Hammond (2006b) conducted a similar study using multiple

data sources such as surveys, interviews, performance assessments, work samples, and

observations to evaluate program outcomes in the Stanford Teacher Education Program

(STEP). One of the measures, a survey, will be discussed here because of its relevance to

the current study.

To assess STEP graduates perceptions about their preparedness to teach and

collect information about their beliefs and practices, a graduate survey was developed

from a national study of teacher education programs by the National Center for

Restructuring education.

Factor analyses identified five factors from the responses to the graduate survey

(Darling-Hammond, 2006b). The factors closely mirrored the California Standards for

the Teaching Profession, supporting the validity of the survey in representing important

dimensions of teaching. Using a 5-point scale, teachers were asked how well STEP

prepared them to: (a) design curriculum and instruction, plan instruction and design

learning experiences for all students; (b) support diverse learners through engagement

and support for all students; (c) use assessments to guide learning and teaching; (d) create

a productive classroom environment, and (e) develop professionally as an educator.

Results from this survey were triangulated with a companion survey for principals and

superintendents measuring the same dimensions. The employer survey measured the

different teaching dimensions and resulted in ratings that were above 4 on a 5-point scale.

Ninety-seven percent of employers gave the program the top rating of 5 for teacher

preparedness. All employers said that they were likely to hire a STEP graduate because

they were well prepared. Similarly, responses from STEP graduates were positive, 87%

stated that they held teaching or other positions in education and over 80% of them

reported that they were engaged in practices aligned with the STEP program.

Study Three. Another study utilized two surveys to examine the effects of

different aspects of professional development on 454 teachers who participated in

GLOBE, an inquiry-based, earth-science, education program (Penuel et al., 2007). The

researchers focused on program effects on teachers' knowledge and ability to implement

the program. The first survey gathered information from 28 GLOBE partners about their

training programs. This survey consisted of 33 questions about different training

processes using the GLOBE program. The second survey was adapted by an instrument

developed by Garet et al. (2001) examining features of effective professional

development. The survey asked GLOBE teachers to report on their implementation of

the program and any barriers they experienced, which they felt inhibited program

implementation.

The GLOBE partner survey was validated using an expert process review, which

included two partner coordinators, an external researcher and the GLOBE administrator.

The reviewers were asked to examine items for relevance, appropriateness, and

significance for program improvement. Items that were deemed less relevant or significant were dropped and others were modified based on feedback from the reviewers.

The teacher survey did not involve extensive pilot testing because all questions used, were from earlier instruments. Confirmatory factor analyses were conducted following data collection to affirm that the constructs were measured reliably as well as to assess the discriminant validity of the measures.

Results from the first survey revealed that the professional development activities offered by GLOBE partners had a significant impact on teachers implementing the program (see Penuel et al., 2007 for complete results) and results for the teacher survey were associated with increased teacher knowledge and changes to science instructional practices.

Examples of Studies Validating the Measure Used in this Study

I developed three surveys for this study using the criteria described in the Standards Performance Continuum (SPC) rubric (see Appendix A), which has been the primary tool used in measuring implementation of the CREDE standards (Hilberg, Doherty, Epaloose, & Tharp, 2004) and one less frequently used measure, the Classroom Observation Schedule, observing student behaviors in classrooms where teachers used the CREDE standards (Waxman & Padron, 2004).

Doherty, Hilberg, Epaloose, & Tharp (2002) described three studies providing evidence of inter-rater reliability, concurrent validity, and predictive validity on the SPC. In the first study, pairs of observers made SPC scores using both live and videotaped observations of 24 public school teachers from an American Indian Pueblo in New

Mexico. Spearman rank-order coefficients and Kendall's *W*s ranged from .86 to .98 and

.80 to .88 on the five subscales, respectively. The total SPC score for each calculated

coefficient was .96.

The second study described by Doherty et al. (2002) examined the concurrent

validity of the SPC scores, measuring the degree to which the scores compared with other

instruments measuring the same constructs. Scores from two other observation measures,

the Teacher Roles Observation Schedule (TROS) and the Classroom Observation

Measure (COM) were compared with the SPC scores. The TROS measures the

instructional setting, the purpose of a particular lesson, and the content covered, whereas

the COM measures classroom characteristics and teacher and student behaviors. Results

indicated a positive correlation between the SPC and TROS (.44 and .33) and significant

relationship between the SPC and the COM (r = .54, $p < .001$).

Doherty et al. (2002) describe the third study, which examined the predictive

validity of the SPC, the degree to which its scores were related to external criteria. In this

study which involved Latino English-language learners, the criteria used were six

indicators from the SAT-9 year-end standardized test: comprehension, language, reading,

spelling, vocabulary, and overall achievement. Overall achievement was calculated by

averaging the scores across all subtests. Results indicated that the higher the use of the

CREDE standards by teachers during language arts instruction, the greater student

achievement on year-end standardized tests.

Results from the SPC rubric offer only a single score for each of the CREDE

standards for a teacher's performance on a particular day and may not reflect what they

do more generally over a longer period of time. Recognizing that professional

development experiences are part of a process and not an isolated event (Guskey, 1995,

2002; Loucks-Horsley et al., 1987), it follows that measuring changes throughout the

process is important. This study is grounded in the belief that professional development

is a process and all aspects, including the yearlong course and repeated measures design

are intended to provide new insights about teachers' participation in sustained

professional development opportunities and how the experience has influenced them.

 In this study, I used a multitrait-multimethod matrix to compare correlational data

across three measures; the practices and attitudes surveys and the video observation

ratings. Trait validity examines the extent to which a measure is related to the same

construct measured by a different method, rather than to measures of different constructs

assessed by the same method (Messick, 1995). The matrix is a convenient way to display

reliability and convergent validity results, which "forces the investigator to confront

simultaneously both convergent and discriminant evidence, or lack thereof" (Messick,

1995, p. 57). Because of the multiple measures used in this study and the multiple

constructs of the CREDE standards, the multitrait-multimethod matrix afforded an

organized approach for displaying a large amount of data.

 In this chapter there are three main points that are worth reemphasizing before

continuing on to the methods section, because they were in part what this study was built

upon. At the policy level, NCLB has a direct impact on what states, schools, and teachers

must do to insure that equity and academic proficiency are accessible to all learners. At

the school and teacher levels there is a plethora of literature on professional development

research, much of which examined teachers' knowledge, practices and beliefs. Finally, in

terms of measurement, it is important that research questions drive methodologies and

that measures are deemed reliable based on evidence and theoretical rationales.

Chapter 3

Methods

This study used a repeated measures evaluation design and multitrait-multi-method matrix (Campbell and Fiske, 1959) to analyze the extent in teachers' change over time during their participation in a yearlong professional development course by, and across measures. The measures were repeated throughout the duration of the course and included three surveys completed by teachers, a video-rating rubric used by two coders, and survey responses from CREDE expert instructors on one occasion. Participants included K-12 teachers and CREDE expert instructors.

Participants

K-12 Teachers. Twenty teacher participants were recruited for this study in 2008-2009. Ten teachers were first year participants learning about the CREDE standards, and five teachers were participants enrolled in a second-year professional development course in which they were learning about two additional CREDE standards, modeling and student directed activity, while continuing to refine their knowledge and practice of the five CREDE standards, and engaging in their own action research. The five continuing teachers were referred to as Hui 1, or Group 1, and the ten, first-year teachers were referred to as Hui 2. All of the teacher participants worked in schools that served a student population comprised of 22% to 74% Native Hawaiians. The grade levels that these teachers served ranged from kindergarten to Grade 12.

Of the 10 Hui 2 teachers, 3 were selected from the Leeward coast of Oʻahu, 4 were from the island of Hawaiʻi, and 3 were from the island of Molokaʻi. Hui 1 included

3 teachers from the Leeward coast and 2 teachers from Maui. Four of the five Hui 1

teachers taught elementary school and all five were Native Hawaiian.

The teacher participants came from 12 different public schools (with multiple

teachers at some of the schools) including one teacher who worked in the Hawaiian

Language Immersion Program[5]. Table 2 presents teachers' background information.

There were 3 males and 12 females ranging in age from 25-50 years old, with 1 to 12

years of teaching experience. Seven of the 15 teachers were born and raised in Hawai'i,

and 8 moved to Hawai'i from other places. Eleven of the participants were elementary

school teachers, and 4 were high school teachers.

For project recruitment purposes, we prioritized teacher participants who wanted

to learn about the teaching strategies, were willing to commit to the yearlong course, and

were from a school or complex area where multiple teachers were interested in

participating. The reason for the last criteria was to promote teacher collaboration and

support throughout the professional development program. Almost 50% of the selected

participants had master's degrees, and one had a doctoral degree. Thus, the participants

were not randomly selected, and in addition to this, they were representative of an

academically advanced group of teachers who might have been more open to exploring

new teaching strategies than other teachers. Teachers who completed the yearlong

professional development course received a stipend upon their completion of the

program.

CREDE Expert Instructors. Participants also included four CREDE experts

with 7 to 35 years of teaching and coaching experience using the CREDE professional

---

[5] The State established the immersion program to promote the study of Hawaiian culture, language and history.

Table 2

Teacher Participants Background Information

ID	Gender	Age	Years of Teaching Experience	Highest Degree Earned	Years Living in Hawaiʻi	Current Grade Level	School Context
* 2	F	28	4	Bachelors	Born/raised	K	Pre-K – Grade 5
* 4	F	28	1	Bachelors	Born/raised	K	Pre-K – Grade 5
* 6	F	29	7	Bachelors	Born/raised	9-12	Grades 7 – 12
* 7	F	25	3	Bachelors	Born/raised	1	Pre-K – Grade 6
* 10	F	31	4	Bachelors	Born/raised	1	Pre-K – Grade 6
12	F	45	4	Bachelors	7	11	Grades 9 – 12
14	F	50	8	Doctoral	22	4	K – Grade 6
15	M	50	11	Bachelors	42	9-12	Grades 9 – 12
16	F	26	6	Masters	4	K-2	Pre-K – Grade 5
17	F	38	2	Masters	Born/raised	K	K – Grade 7
18	F	36	7	Masters	5	4-5	K – 12
19	M	36	12	Masters	5	5	Pre-K – Grade 5
22	F	27	5	Bachelors	Born/raised	2	Pre-K – Grade 6
24	F	33	10	Masters	3	2	Pre-K – Grade 6
25	M	34	8	Bachelors	0	9-10	Grades 9 – 12

* Hui 1 teachers

development model. All expert instructors are female and their ages at the time of the

study were 71, 64, 44, and 33, respectively. Two of the experts served as coaches, retired

K-12 teachers, and were part of the original KEEP study, the third was the principal

investigator of the broader study and Hui 1 instructor, and the fourth was a teacher

educator.

Instruments

Table 3 presents the multiple data sources used in this study, which included four

surveys (Teacher Background, Teaching Practices, Attitudes Toward Teaching, and

Classroom Vignette) and video observational ratings using the Classroom Observation

Rubric (COR).

Table 3

Data Sources and Method Used for Measuring Teachers' Practices, Attitudes, and
Knowledge of the CREDE Five Standards For Effective Pedagogy

Data Source	Method	Teacher Participants	No. of Participants	No. of Times Collected
Teacher Background Survey	Questionnaire	2, 4, 6, 7, 10, 12, 14, 15, 16, 17, 18, 19, 22, 24, 25	15	1
Teaching Practices Survey	Questionnaire	2, 4, 6, 7, 10, 12, 14, 15, 16, 17, 18, 19, 22, 24, 25	15	4
Attitudes Toward Teaching Survey	Questionnaire	2, 4, 6, 7, 10, 12, 14, 15, 16, 17, 18, 19, 22, 24, 25	15	3
Classroom Vignette Survey	Questionnaire	2, 4, 6, 7, 10, 12, 14, 15, 16, 17, 18, 19, 22, 24, 25	15	4
Video Observations	COR	7, 12, 14, 15, 16, 18, 19, 22	8	4

Teacher Background Survey. The Teacher Background survey (see Appendix B) asked teacher participants to provide demographic data such as gender, age, university course work, years of teaching experience, certificates and degrees earned, subject and grade levels taught, in addition to other background information.

Teaching Practices (TP) Survey. The practices survey was developed using the SPC rubric and Indicators Criteria[6]. The survey was comprised of 34[7] items; 5 items measuring Joint Productive Activity, 5 items measuring Language and Literacy Development, 7 items measuring Contextualization, 12 items measuring Complex Thinking, and 8 items measuring Instructional Conversation. The Teaching Practices survey (see Appendix D) was designed to measure teachers' behaviors and practices using the CREDE standards and assess how they changed over time. For example, Item 1 states, "I organize lessons so that students can work together in small groups." This item represents JPA and a teacher has a choice of selecting one of four responses: "never;" "1-2 times per week;" "3-4 times per week," or "daily."

Attitudes Toward Teaching (ATT) Survey. The attitudes survey (see Appendix E) was developed using the SPC rubric and Indicator Criteria. The survey was comprised of 33[8] items that mirrored those on the Teaching Practices survey, but used a different

---

[6] See Appendix C. Indicators are also available on the CREDE Website, http://crede.berkeley.edu/research/crede/standards.html

[7] The item total was equal to 37; however, five of the items—3, 4, 15, 17, and 24—measured two of the Five CREDE Standards (e.g., Joint Productive Activity and Instructional Conversation or Complex Thinking and Instructional Conversation).

[8] Some of the items were dropped and four remaining items—3, 5, 14, 16—measured two of the Five CREDE Standards (e.g., Joint Productive Activity and Language and Literacy Development or Joint Productive Activity and Instructional Conversation or Complex Thinking and Instructional Conversation).

response scale. There were 3 items measuring Joint Productive Activity, 5 items measuring Language and Literacy Development, 6 items measuring Contextualization, 11 items measuring Complex Thinking, and 6 items measuring Instructional Conversation. The items on the Attitudes Toward Teaching survey were designed to measure teachers' attitudes and beliefs about the CREDE standards and to assess how they changed over time. For example, Item 1 on the Attitudes Toward Teaching survey states "It is important for me to organize lessons so that students can work together in small groups" with response options of, "strongly agree," "somewhat agree," "somewhat disagree," or "strongly disagree." This item was developed to assess how important a teacher believed organizing lessons around small group activities was for students.

Classroom Vignette Survey. The vignette survey (see Appendix F) was designed to measure teachers' knowledge of the CREDE standards. There were a total of 17 items on the survey, three items designed to represent each of the five CREDE standards and two items that did not reflect any of the standards. The latter two items were included to illustrate a variety of instructional practices, including classrooms where the CREDE standards were absent. Respondents were asked to identify the "best standard," with the option to select a second standard if they chose, to represent each scenario and the level of implementation of the same standard ranging from "not observed" to "enacting." For example, item 12 stated the following:

> Fourth grade students attending a rural public school have been learning about the natural food chain in their science class. Today the teacher asked students to work together in small groups to classify parts of an owl pellet (p. 4).

The correct response for the "best standard" that represents this item was Joint Productive

Activity and the level of implementation was "2" for "developing." According to the

SPC, at the developing level "the teacher and students collaborate on a joint product in a

whole-class setting, or students collaborate on a joint product in pairs or small groups."

Classroom Observation Rubric (COR). The Classroom Observation Rubric

(see Appendix G) was based on the Standards Performance Continuum (SPC), a

validated measure of the CREDE standards (Doherty et al., 2002). The SPC was the

original rubric used to measure teacher enactment of the CREDE standards. The SPC

rubric and the corresponding indicators were the sources used to develop the three

surveys for this study. Part way through the video rating process, the SPC was revised

from a zero to four-point scale, "not observed, emerging, developing, enacting," and

"integrating" to a six-point scale on the newly developed COR (Luning, 2011).

Procedures

Teacher Surveys. The three surveys—practices, attitudes and vignette—were

collected on multiple occasions during the professional development course. The

practices and vignette surveys were each collected four times and the attitudes survey

was collected three times (see Table 3). All teachers completed the Background survey

once at the beginning of the professional development course.

Video recordings of Teacher Instruction. Teacher participants were videotaped

on four occasions during the year, and they selected when their recordings would take

place. They were recorded for a full class period and wore a lavaliere microphone during

each session. Whoever was assigned to videotape the class followed a list of pre-

established protocol to insure that the recording process would be minimally disruptive for the teacher and students.

Two trained coders rated teachers' videotapes using a two-step process. First, each coder assigned a score between 0 (not observed) and 5 (exemplary) for each CREDE standard. Next, the coders met to discuss discrepancies between their individual codes, until they reached consensus on final scores.

CREDE Expert Instructor Surveys. At the beginning of the project, four CREDE expert instructors completed the practices, attitudes, and classroom vignette surveys on one occasion. Because of their many years of experiences implementing the standards in their own classrooms and teaching them in other educational settings, they were assumed to be knowledgeable content experts on the CREDE standards and their responses assisted in distinguishing patterns of good and bad items across the three survey measures.

Data Analysis

I used multiple methods for analyzing data in this study to examine the relationships among teachers' instructional practices, attitudes toward teaching and learning, knowledge about the CREDE standards and video recorded performances. Each of the following analyses corresponds to one of the six research questions and the final section describes the multiple roles I played in the study.

1. Analysis of Consistency of CREDE Expert Instructors Ratings on their Use of, Attitudes Toward, and Knowledge of the CREDE Standards. Inter-rater reliability results were calculated for the practices and attitudes surveys using Cronbach's alpha, and the vignette survey was analyzed using percent agreement.

2. Validity Evidence of Teachers' Practices, Attitudes, and Performances

Implementing the Five CREDE Standards. I examined the relationships between
teacher self-report scores on the two surveys measuring teachers' use of the standards in
their practices (Teaching Practices Survey), their attitudes about using the standards
(Attitudes Toward Teaching Survey) and their observed practice scores using the COR.
A multitrait-multimethod matrix is provided (see Table 4) with correlation coefficients
from the surveys and the COR. The matrix provides a framework to determine how
adequate the surveys were in relation to the COR scores as measures of the CREDE
standards.

The multiple traits are the CREDE standards and the multiple methods are the
practices and attitudes surveys and the video ratings. The fundamental notion is that if a
construct is adequately defined and measured, different methods of measuring the
construct should converge and give the same information.

The simplified, schematic matrix reflects five traits—in this case, the CREDE
standards—Joint Productive Activity (JPA), Language and Literacy Development (LLD),
Contextualization (CTX), Complex Thinking (CT), and Instructional Conversation (IC).
For illustrative purposes, the matrix includes two methods, Method 1, the Teaching
Practices survey and Method 2, the video ratings. Results from a third method, the
Attitudes Toward Teaching survey, will be included in the multitrait-multimethod matrix
in the results section. The solid shaded areas represent reliability coefficients along the
main diagonal of the super matrix for each of the traits (CREDE standards). Typically,
all of the numbers in this diagonal in a correlation matrix would reflect a 1.00
representing a perfect correlation between a variable and itself. In this matrix, the

numbers measure internal consistency using Cronbach's alpha. The cells with the grids along the main diagonal of the heterotrait-heteromethod rectangle represent convergent validity coefficients—the convergence of different methods on the same construct. The non-shaded areas in the rectangle represent discriminant validity coefficients— correlations between different constructs measured by the same or different methods (Shavelson, et al., 1976).

Table 4

Schematic Representation of Areas of a Multitrait-multimethod Matrix

Traits		Method 1 Teaching Practices Survey					Method 2 Video Ratings				
		JPA_1	LLD_1	CTX_1	CT_1	IC_1	JPA_1	LLD_2	CTX_2	CT_2	IC_2
Method 1 Teaching Practices Survey	JPA_1	$r_{xx'}$									
	LLD_1	$r_{LLD,JPA}$	$r_{xx'}$								
	CTX_1	$r_{CTX,JPA}$	$r_{CTX,LLD}$	$r_{xx'}$							
	CT_1	$r_{CT,JPA}$	$r_{CT,LLD}$	$r_{CT,CTX}$	$r_{xx'}$						
	IC_1	$r_{IC,JPA}$	$r_{IC,LLD}$	$r_{IC,CTX}$	$r_{IC,CT}$	$r_{xx'}$					
Method 2 Video Ratings	JPA_2	r_{xy}	$r_{JPA,LLD}$	$r_{JPA,CTX}$	$r_{JPA,CT}$	$r_{JPA,IC}$	$r_{xx'}$				
	LLD_2	$r_{LLD,JPA}$	r_{xy}	$r_{LLD,CTX}$	$r_{LLD,CT}$	$r_{LLD,IC}$	$r_{LLD,JPA}$	$r_{xx'}$			
	CTX_2	$r_{CTX,JPA}$	$r_{CTX,LLD}$	r_{xy}	$r_{CTX,CT}$	$r_{CTX,IC}$	$r_{CTX,JPA}$	$r_{CTX,LLD}$	$r_{xx'}$		
	CT_2	$r_{CT,JPA}$	$r_{CT,LLD}$	$r_{CT,CTX}$	r_{xy}	$r_{CT,IC}$	$r_{CT,JPA}$	$r_{CT,LLD}$	$r_{CT,CTX}$	$r_{xx'}$	
	IC_2	$r_{IC,JPA}$	$r_{IC,LLD}$	$r_{IC,CTX}$	$r_{IC,CT}$	r_{xy}	$r_{IC,JPA}$	$r_{IC,LLD}$	$r_{IC,CTX}$	$r_{IC,CT}$	$r_{xx'}$

The non-shaded areas contain the correlations among the different standards using one method, the practices survey.

The heterotrait-monomethod matrix is located in the lower right corner of the super matrix. This submatrix contains information on the different relationships among the five traits using Method 2, the video ratings. The main diagonal (shaded boxes) contains inter-rater reliability coefficients. These coefficients were calculated for the two raters who used the COR.

As shown in Table 5, the heterostandard-heteromethod matrix contains information on the relationships among the CREDE standards as measured by two methods, Methods 1, the Teaching Practices survey and Method 2, the video ratings. The cells with the grids on the diagonal in this submatrix represent convergent validity coefficients. These coefficients should be statistically significant, of practical value, and greater than the correlations in their corresponding rows and columns. These are the correlations that bear directly on the research question. That is, they provide the correlation between different measures for the same standard.

Table 5

Submatrix for Methods 1 and 2

		Method 1 Teaching Practices Survey				
		JPA_1	LLD_1	CTX_1	CT_1	IC_1
Method 2 Video Ratings	JPA_2	$r_{xx'}$	$r_{JPA,LLD}$	$r_{JPA,CTX}$	$r_{JPA,CT}$	$r_{JPA,IC}$
	LLD_2	$r_{LLD,JPA}$	r_{xy}	$r_{LLD,CTX}$	$r_{LLD,CT}$	$r_{LLD,IC}$
	CTX_2	$r_{CTX,JPA}$	$r_{CTX,LLD}$	r_{xy}	$R_{CTX,CT}$	$r_{CTX,IC}$
	CT_2	$r_{CT,JPA}$	$r_{CT,LLD}$	$r_{CT,CTX}$	r_{xy}	$r_{CT,IC}$
	IC_2	$r_{IC,JPA}$	$r_{IC,LLD}$	$r_{IC,CTX}$	$r_{IC,CT}$	r_{xy}

3. **Analysis of Changes in Teaching Practices on the Use of the CREDE Standards**. The Teaching Practices survey measured the extent to which teachers used the CREDE standards during their instruction. I examined the trends of the standards over four time periods (August, November, February, and May). Trend analyses were performed using repeated measures ANOVA to examine changes in teachers' self-reported instructional practices for each of the standards during their participation in the CREDE professional development course. I controlled for three covariates from the Teacher Background Survey to establish whether there were any relationships between teachers' use of the CREDE standards in their classroom and their years of teaching experience, certification, and placement as an elementary or secondary teacher.

4. **Analysis of Changes in the Teachers' Attitudes on the Use of the CREDE Standards**. The Attitudes Toward Teaching Survey measured the degree to which teachers agreed or disagreed with using the CREDE standards in their instruction. I examined the trends of their attitudes toward the standards over three time periods (August, February, and May; data were not collected in November). Trend analyses were performed using repeated measures ANOVA to examine changes in teachers' attitudes about using each of the standards during their participation in the professional development course. The covariates I controlled for from the Teacher Background Survey were the same as those used with the Teaching Practices Survey, years of teaching experience, certification, and placement as an elementary or secondary teacher.

5. **Analysis of Videotaped Scores as Measured using the Classroom Observation Rubric (COR)**. Two trained coders watched the teachers' video recorded instruction and assigned a score for each of the CREDE standards, reflecting the extent to

which the observed standard was being implemented during the lesson. The two coders then met to compare their scores, discuss any differences, and reconcile a final agreed upon single score. The COR uses a zero to six-point scale, "not observed, emerging, developing, advancing, enacting," and "exemplary." For each of the standards, the inter-rater reliability was calculated using intraclass correlation. Trend analyses using repeated measures ANOVA was used to examine changes in teachers' observed practices during their participation in the professional development course.

6. Analysis of Changes in Teachers' Knowledge about the Use of the CREDE Standards. Data from the Classroom Vignette survey were analyzed over the four time periods using Cronbach's alpha and percentages to determine how reliable the survey captured teachers' knowledge of the CREDE standards. Each item was developed using the criteria listed in the Standards Performance Continuum rubric, measuring various degrees of implementation of each CREDE standard.

Situating Myself within the Study. I served multiple roles in this study ranging from teacher recruiter, to course instructor, program manager, and educational researcher, and therefore, it is important to acknowledge my relationship with the teacher participants from the beginning, to end of this study. Recognizing my role as instructor, to reduce biases throughout the data collection process, some alternative measures were established for proctoring the three surveys, as well as videotaping teachers' instruction. For the surveys, another CREDE staff member administered and collected each survey and followed a script explaining the purpose of the survey and basic instructions for completing them. The videos were recorded by either a professional videographer, or by a trained CREDE staff member.

My relationship with all of the teachers was a partnership. My expectations of teachers, in terms of the course and assignments, were equitable and consistent. However, as I observed teachers' efforts over time in completing assignments, meeting deadlines, and actively participating in the professional development course, my relationship with those teachers who were more engaged became richer than my relationship with those who were less engaged.

I appreciated the dual roles of the participants as in-service teachers and university students enrolled in the course. For the local teachers, as is often customary within the Hawaiian culture, it was important for them to understand *where* I was from, where I grew up, which high school I graduated from, etc. And although no one ever explicitly asked me the question directly, there were a few times when a certain look that a teacher gave me made me feel as though they were asking, "*What do you know about Hawaiian children*," which in my mind was both a relevant and fair question for them to ask.

My Hawaiian ancestry dates back five generations, however, my fair skin and green eyes reflect more of my English, Scotch, Irish, French, German heritage, rather than my Hawaiian, Chinese, Portuguese, or Tahitian ancestry. For local teachers, there seemed to be greater interest in the place where I grew up, the school I attended and what experiences I had involving Native Hawaiian education, compared with teachers who were new to the islands, who seemed less interested in this information and somewhat more interested in my academic credentials. For the local teachers, knowing that I was part-Hawaiian, that I attended public schools and that I had a child who graduated from a public school, was perhaps more of a cultural acceptance, which enhanced what they

knew about my academic background, more so than for the teachers who were not from the islands. Of all of the roles I played in this study, the greatest potential bias resulted from my dual role as instructor and researcher. My first interest was to assist teachers in understanding the CREDE standards so they could use the strategies in their classroom. However, I was also interested in measuring the extent to which teachers reported their practices, attitudes, and understandings of the CREDE standards through the surveys. During the study, teachers may have felt that they had to respond to the surveys with correct answers because they would receive a grade for the course, but because of the lapse in time from when the study was conducted and the data analyzed and summarized, my biases as instructor were reduced.

Chapter 4

Results

In this chapter I present the results of my analyses addressing the six questions underlying this research: (a) To what extent did CREDE expert instructors agree on the measurement of the CREDE standards for the practices, attitudes, and vignette surveys? (b) How did teachers' responses on the practices and attitudes surveys compare with results from teachers' video recorded instruction? (c) To what extent did teachers' instructional practices change over time based on a survey that was designed to measure teachers' performance using the CREDE standards in the classroom? (d) To what extent did teachers' attitudes change over time based on a survey that was designed to measure teachers' beliefs about teaching and learning using the CREDE standards in the classroom? (e) To what extent did teachers' instructional practices change over time based on results from their videotaped lessons using the Classroom Observation Rubric? (f) To what extent did teachers' reliably respond to the survey designed to measure teachers' understandings of the CREDE standards?

Research Question 1: To what extent did CREDE expert instructors agree on the measurement of the CREDE standards for the practices, attitudes, and vignette surveys?

CREDE Expert Instructors. As shown in Table 6, reliability results varied across the standards on the practices and attitudes surveys. On the practices survey, there was no reliability for Language and Literacy Development and the reliability for Complex Thinking was .82. There was no reliability for Joint Productive Activity or Instructional Conversation on the attitudes survey and results for the other standards were also low.

Responses were also calculated using percent agreement among the four experts

for the practices and attitudes surveys. To calculate the percentages, the total number of

disagreements among the four experts was calculated to determine the difference between

items where there were perfect matches (i.e., all experts had the same response), to one,

or two teachers having a different response[9]. For example, for Contextualization on the

practices survey, there were seven items measuring this standard. Four of the items were

perfect matches, two items had three matching experts' responses, and one item had only

50% agreement among the experts. There were a total of three disagreements across the

experts that totaled 89 percent agreement for this standard.

Table 6

Reliability among the CREDE Expert Instructors on the Teaching Practices and Attitudes Toward Teaching Surveys

CREDE Standard	Teaching Practices		Attitudes Toward Teaching	
	Reliability	Percent Agreement	Reliability	Percent Agreement
JPA	.69	.75	0	.94
LLD	0	.88	.67	.92
CTX	.27	.89	.74	.89
CT	.82	.86	.51	.70
IC	.46	.79	0	.71

Reliability results and percent agreement were also calculated for experts on the

vignette survey (see Table 7). Again, reliability results varied across the standards;

however in general, percent agreement results were consistent among the raters.

As shown in Table 8, on the Classroom Vignette survey all four experts

responded with the "correct" answers to Items 1, 2, and 12 (Items 1 and 12 are Joint

Productive Activity and Item 2 is Contextualization). Three of the four

---

[9] There were no items that produced four different responses from the experts.

Table 7

Reliability and Percent Agreement for CREDE Expert Instructors on the Classroom Vignette Survey

CREDE Standard	Classroom Vignette	
	Reliability	Percent Agreement
JPA	.50	.83
LLD	0	.75
CTX	0	.91
CT	.75	.75
IC	.88	.75

Table 8

Vignette Expert Instructor Responses

Item	Expert 1	Expert 2	Expert 3	Expert 4	CORRECT
1	JPA	JPA	JPA	JPA	JPA
2	CTX	CTX	CTX	CTX	CTX
3	LLD	LLD	IC	LLD	CTX
4	MOD	CT	JPA	LLD	LLD
5	CTX	CTX	BLANK	CTX	NONE
6	CT	IC	IC	IC	IC
7	SDA	LLD	LLD	JPA	LLD
8	CT	LLD	BLANK	CTX	NONE
9	LLD	LLD	LLD	BLANK	LLD
10	CT	IC	IC	IC	IC
11	JPA	IC	IC	JPA	IC
12	JPA	JPA	JPA	JPA	JPA
13	MOD	JPA	CT	LLD	JPA
14	CT	CT	CT	CTX	CT
15	IC	CTX	LLD	CTX	CTX
16	MOD	IC	IC	JPA	IC
17	CT	CT	CT	JPA	CT

experts responded correctly to Items 6, 9, 10, 14 and 17. For Items 3 and 5, three experts

responded with the same answer, but the answer given was not the standard that the items

were developed to illustrate. For example, Item 3 was developed to measure

Contextualization, however, three experts believed that the vignette illustrated Language

and Literacy Development and one expert believed that the scenario illustrated

Instructional Conversation. The same applied to Item 5. Although it was designed as a

non-CREDE standard classroom situation, three experts believed it was an example of Contextualization.

Research Question 2: How did teachers' responses on the practices and attitudes surveys compare with results from teachers' video recorded instruction?

Multitrait-Multimethod Matrix. The second research question asked whether teachers' responses on the practices and attitudes surveys and video ratings by trained coders who used the COR, produced similar results across the multiple measures and time periods. Specifically, this research question addressed whether surveys of teachers' practices and attitudes might serve as surrogates for video observation ratings in classrooms where teachers use the CREDE standards. Data results from the vignette survey were not included in the multitrait-multimethod analyses.

In the following multitrait-multimethod matrix (Table 9), the reliability coefficients and correlation results are presented for three methods: the Teaching Practices survey, the Attitudes Toward Teaching survey, and video ratings across four times periods.

Results are shown for Time 4 because it represents the time when it would be expected that teachers knew the most, having participated in the yearlong course. There are three methods reflected in each major column of the multitrait-multimethod matrix: Method 1, Teaching Practices survey; Method 2, Attitudes Toward Teaching survey and Method 3, Video Ratings. The same methods are repeated in the rows. Within each method, there are five subcategories (JPA, LLD, CTX, CT, and IC) that represent the CREDE standards. Results are presented for each submatrix. There are a total of six

submatrices, which are comprised of 5x5 sets of cells for, and across, each method, beginning with the three main reliability diagonals for the three methods.

In the upper left corner of the multitrait-multimethod matrix, the entries in the reliability diagonal of the heterotrait-monomethod submatrix represent coefficients for Method 1, the teaching practices survey, using Cronbach's alpha for each of the standards. In general, these results are acceptable but the lower coefficients suggest more items are needed if used in the future. Reliability for Complex Thinking is the highest at .824, followed by Joint Productive Activity at .734. Reliability for the remaining three standards are marginally acceptable: .636 (LLD), .686 (CTX) and .663 (IC), respectively.

The second heterotrait-monomethod submatrix, located in the middle of the super matrix, contains information on the different relationships among the standards using Method 2, the attitudes survey. Reliabilities using Cronbach's alpha are highest for Contextualization at .901, acceptable for Joint Productive Activity at .770 and Complex Thinking (.767), but unacceptable for Language and Literacy Development (.357) and Instructional Conversation (.370).

The last of the three main heterotrait-monomethod submatrix diagonals report inter-rater reliability coefficients, and in this case, coefficients calculated for the two raters who used the COR. Results indicated that raters were consistent in scoring at time four, ranging from the lowest reliability (using a single rater) of .824 for Language and Literacy Development, to the highest reliabilities of .976 (Instructional Conversation) and .977 (Joint Productive Activity).

The next set of diagonals examined across the three methods was validity. The first diagonal compared Method 1 (practices) and Method 2 (attitudes) and the second

diagonal compared Method 2 (attitudes) and Method 3 (video). Validity results in both heterotrait-heteromethod submatrices did not reflect convergent validity for Methods 1 and 2, or Methods 2 and 3, for any of the standards.

The final heterotrait-heteromethod submatrix examined the relationships among the CREDE standards as measured by two methods, Methods 1, the Teaching Practices survey and Method 2, the video ratings. These results represent convergent validity coefficients. Of the five standards, only Joint Productive Activity produced statistically significant results at .883.

Research Question 3: To what extent did teachers' instructional practices change over time based on a survey that was designed to measure teachers' performance using the CREDE standards in the classroom?

Teaching Practices. The Teaching Practices survey was designed to measure teachers' self-reported performance in the classroom. To address this question, I examined the influences of the possible covariates affecting teachers' attitudes as well as the changes over time using repeated measures ANCOVA. The covariates included teachers' experience levels, certification, and placement as an elementary or secondary teacher. Results were not included in the study because none of the covariates turned out to be statistically significant predictors of survey proportional scores. Accordingly, the data were analyzed by repeated measures ANOVA.

Table 9

Multitrait-Multimethod Matrix Using Cronbach's Alpha for Practices and Attitudes Surveys and Intraclass Correlations for Video Ratings Time 4

Traits		Method 1 Teaching Practices Survey					Method 2 Attitudes Toward Teaching Survey					Method 3 Video Ratings				
		JPA	LLD	CTX	CT	IC	JPA	LLD	CTX	CT	IC	JPA	LLD	CTX	CT	IC
Method 1 Teaching Practices Survey	JPA	.734														
	LLD	.488	.636													
	CTX	.228	.632*	.686*												
	CT	.654**	.625**	.494	.824*											
	IC	.340	.633*	.648**	.832**	.663										
Method 2 Attitudes Toward Teaching Survey	JPA	.005	-.012	.215	.086	-.224	.770									
	LLD	.057	-.635*	-.196	-.177	.232	.319	.357								
	CTX	.481	-.329	-.639*	-.006	.337	-.524*	.141	.901							
	CT	.299	-.156	-.390	-.243	.193	-.188	.259	.232	.767						
	IC	.201	-.414	-.574*	-.249	.233	.476	-.244	-.444	-.423	.370					
Method 3 Video Ratings	JPA	.883*	.641	.000	.277	.647	.142	-.065	.411	.093	-.379	.977**				
	LLD	.490	.533	-.267	.635	.385	-.463	-.106	.522	.304	-.773*	.612	.824**			
	CTX	.111	-.181	-.301	.104	-.313	-.540	-.318	.075	-.608	-.309	.000	.167	.857**		
	CT	.549	.404	-.090	-.039	.129	-.267	-.275	.387	.526	-.268	.354	.722*	-.289	.706**	
	IC	.105	.256	.171	.074	-.542	.045	.367	.818*	.219	-.223	.589	.481	-.385	.500	.976**

*p<.05; **p<.01

The reliability of the proportional scores[10] was examined using Cronbach's alpha (see Table 10). Reliability results increased from Time 1 to Time 4 for three of the standards across time periods: Joint Productive Activity (Time 1 was .601, Time 4 was .734), Language and Literacy Development (Time 1 was .491, Time 4 was .636), and Contextualization (Time 1 was .649, Time 4 was .686). Reliabilities for the other two standard proportional scores decreased from Time 1 to Time 4 on the Teaching Practices survey. Complex Thinking was consistently high across the 4 time points, at .77 or above. For Times 3 and 4 reliabilities averaged around .70, and were at an acceptable level, except for Instructional Conversation.

Table 10

Cronbach's Alpha Coefficients of Teaching Practices Survey

Standard	Time 1	Time 2	Time 3	Time 4
JPA	.601	.223	.684	.734
LLD	.491	.707	.659	.636
CTX	.649	.721	.773	.686
CT	.883	.830	.770	.824
IC	.696	.327	.510	.663

No substantial mean change was observed over the four time periods for Joint Productive Activity, Contextualization, Complex Thinking, or Instructional Conversation suggesting that experience and training did not produce teacher change over time or reflect significant improvement (see Table 11).

Figure 1 shows the trend of mean scores over the four time points for each of the standards. Means appear to increase over time for Language and Literacy Development, Complex Thinking and Instructional Conversation.

---

[10] These scores were calculated by taking teachers' scores for each time period (e.g., Time 1, Joint Productive Activity, total score equal to 13) and dividing it by the number of items (n_i=5), multiplied by the number of scale responses (n_{pts}=4): (13)/(5 x 4) = 13/20 = 0.65.

Table 11

Descriptive Statistics Means (M) and Standard Deviations (SD) of Proportional Scores for Teaching Practice Survey

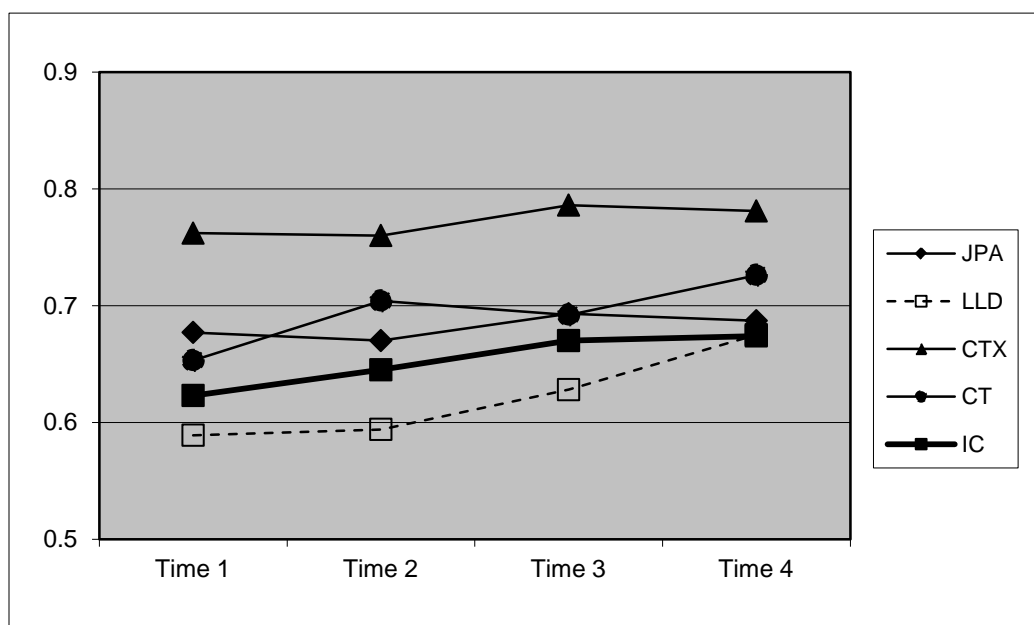| Standard		Time 1	Time 2	Time 3	Time 4
JPA	M	.677	.670	.693	.687
	SD	.110	.110	.133	.079
LLD	M	.589	.594	.628	.675
	SD	.117	.131	.112	.104
CTX	M	.762	.760	.786	.781
	SD	.079	.125	.121	.110
CT	M	.653	.704	.692	.726
	SD	.114	.119	.093	.103
IC	M	.623	.645	.670	.674
	SD	.084	.091	.091	.062

Figure 1. Mean Proportion Scores for Each CREDE Standard on the Teaching Practices Survey on Four Occasions

In the following tables, the results of repeated measures ANOVA are presented for each standard to determine whether teachers' mean scores on the Teaching Practices survey changed over time. Judging from the descriptive data and contrary to what was

expected, there was reason to believe I would not see improvement except perhaps on the Language and Literacy Development and Complex Thinking standard.

Joint Productive Activity. Table 12 presents the results of repeated measures ANOVA with Joint Productive Activity. As shown, Time was not statistically significant (using Huynh-Feldt), implying that the teachers' mean responses on their practices for Joint Productive Activity did not reliably change throughout the professional development period.

Table 12
Repeated Measures ANOVA with Joint Productive Activity

Source	SS	df	MS	F
Between				
Teachers	209.933	14	14.995	
Within				
Time	1.933	3	.644	.451
Error	60.067	42	1.430	
Total	271.933	59		

Language and Literacy Development. Results of repeated measure ANOVA with Language and Literacy Development are presented in Table 13. Significant results were observed, implying that the teachers' practices using Language and Literacy Development changed, on average, over time. Subsequent post-hoc analyses using the Least Significant Difference (LSD) test revealed that the average Language and Literacy Development score at Time 2 was significantly higher than at Time 1, Difference = -1.933, $p = .001$, and the average at Time 4 was significantly higher than at Time 1, Difference = 1.133, $p = .023$.

Contextualization. As shown in Table 14, time (Huynh-Feldt) was not significant. Consistent with the mean differences in Table 5, teachers' Contextualization practices mean scores did not significantly change over the four time periods.

Table 13

Repeated Measures ANOVA with Language and Literacy Development

Source	SS	df	MS	F
Between				
Teachers	357.833	14	25.560	
Within				
Time	40.583	2.897	14.007	7.269**
Error	78.167	40.563	1.927	
Total	476.583	57.460		

**p<.01

Table 14

Repeated Measures ANOVA with Contextualization

Source	SS	df	MS	F
Between				
Teachers	383.933	14	27.424	
Within				
Time	6.183	3	2.061	.557
Error	150.067	42	3.573	
Total	540.183	59		

Complex Thinking. The teachers' Complex Thinking mean scores changed

significantly over the time periods. As shown in Table 15, time was significant.

Subsequent post-hoc analyses with the LSD test revealed that the average Complex

Thinking score at Time 2 was significantly higher than at Time 1, Difference = -2.467,

p = .039, and the average at Time 4 was significantly higher than at Time 1, Difference =

-3.533, p = .020.

Instructional Conversation. The teachers' average scores for Instructional

Conversation did not change reliably over the professional development period. As

shown in Table 16, time (Huynh-Feldt) was not statistically significant.

Table 15

Repeated Measures ANOVA with Complex Thinking

Source	SS	df	MS	F
Between				
Teachers	1156.1	14	82.579	
Within				
Time	98.733	2.780	35.519	4.117*
Error	335.767	38.916	8.628	
Total	1590.6	55.696		

*p<.05

Table 16

Repeated Measures ANOVA with Instructional Conversation

Source	SS	df	MS	F
Between				
Teachers	237.857	13	18.297	
Within				
Time	24.357	3	8.119	2.452
Error	129.143	39	3.311	
Total	391.357	55		

Research Question 4: To what extent did teachers' attitudes change over time based on a survey that was designed to measure teachers' beliefs about teaching and learning using the CREDE standards in the classroom?

Data from the Attitudes Toward Teaching survey were used to address this question. I examined the influences of the possible three covariates affecting teachers' attitudes as well as the changes over time using repeated measures ANCOVA. The covariates included teachers' attitudes about the CREDE standards and teachers' experience levels, certification, and placement as an elementary or secondary teacher. None of the covariates were statistically significant and the results of these analyses are

not reported. Accordingly, trend analyses for each of the standards were performed using repeated measures ANOVA.

Attitudes Toward Teaching. Reliability results for the Attitudes Toward Teaching survey using Cronbach's alpha are shown in Table 17. Looking at the reliability coefficients at Time 1 and Time 4, Joint Productive Activity (Time 1 was .732, Time 4 was .770), Contextualization (Time 1 was .802, Time 4 was .901) and Complex Thinking (Time 1 was .751, Time 4 was .767), scores improved from the beginning of the school year to the end. For Language and Literacy Development, there was a slight decrease in the alpha scores from Time 1 to Time 4, while scores for Instructional Conversation were significantly lower from the first administration to the fourth.

Table 17

Cronbach's Alpha Coefficients of Attitudes Toward Teaching Survey

Standard	Time 1	Time 2	Time 3	Time 4
JPA	.732	-	.872	.770
LLD	.411	-	.284	.357
CTX	.802	-	.772	.901
CT	.751	-	.635	.767
IC	.740	-	.760	.370

Means and standard deviations of the attitude scores are presented in Table 18 for each standard over the three time periods (data were not collected at Time 2). In general, results for each of the CREDE standards reflect little mean change from Time 1 to 4. Time was not a statistically significant predictor of change in teachers' mean attitude scores over the three time periods for Joint Productive Activity, Contextualization, Complex Thinking, or Instructional Conversation. Only Language and Literacy

Development showed statistically significant mean change, although inconsistently and not in the predicted, increasing direction.

Table 18

Descriptive Statistics Means (M) and Standard Deviations (SD) of Proportional Scores for Attitudes Toward Teaching Survey

Standard		Time 1	Time 3	Time 4
JPA	*M*	.677	.688	.646
	SD	.164	.189	.148
LLD	*M*	.509	.425	.472
	SD	.082	.084	.073
CTX	*M*	.302	.305	.289
	SD	.072	.091	.077
CT	*M*	.417	.420	.397
	SD	.077	.074	.079
IC	*M*	.733	.713	.663
	SD	.167	.163	.092

Figure 2 shows the range of mean proportional scores on the Attitudes Toward Teaching survey at Times 1, 3 and 4.

Tables 19 through 23 present repeated measures ANOVA results for each standard to determine whether teachers' mean scores on the Attitudes Toward Teaching survey changed over the professional development period.

Joint Productive Activity. As shown in Table 19, the time (Huynh-Feldt) was not significant, implying that the teachers' Joint Productive Activity attitude scores did not change over the professional development period.
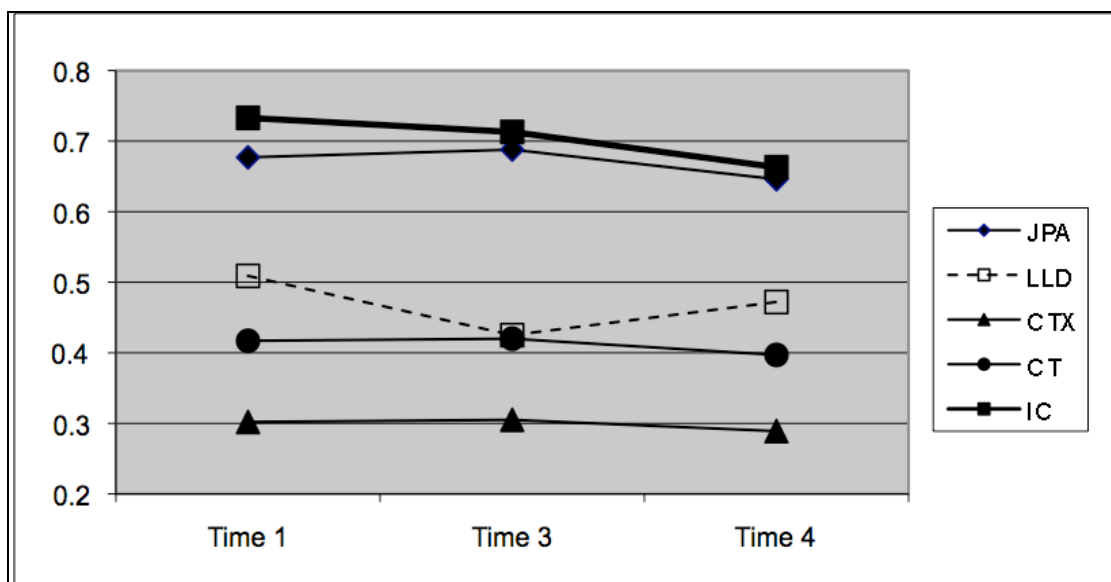
Figure 2. Mean Proportion Scores for Each CREDE Standard on the Attitudes Toward Teaching Survey on Four Occasions

Table 19

Repeated Measures ANOVA with Joint Productive Activity

Source	SS	df	MS	F
Between				
Teachers	135.917	15	9.061	
Within				
Time	2.167	2	1.083	.709
Error	45.833	30	1.528	
Total	183.917	47		

Language Literacy Development. As shown in Table 20, time (Huynh-Feldt) was significant, suggesting that the teachers' Language and Literacy Development attitude scores changed over the professional development period. Subsequent post-hoc analyses using Least Significant Difference revealed that the average Language and Literacy Development attitude scores at Time 3 was significantly lower than at Time 1, Difference= -1.688, *p* = .001, and the average at Time 4 was significantly higher than at

Time 3, Difference = .938, p = .047. Consequently, the change seemed inconsistent and

not in predicted direction.

Table 20

Repeated Measures ANOVA Table with Language and Literacy Development

Source	SS	df	MS	F
Between				
Teachers	72.583	15	4.839	
Within				
Time	22.875	2	11.438	8.210**
Error	41.792	30	1.393	
Total	137.25	47		

**p<.01

Contextualization. As shown in Table 21, the time (Huynh-Feldt) was not

significant, implying that the teachers' attitude scores for Contextualization did not

change over time.

Table 21

Repeated Measures ANOVA with Contextualization

Source	SS	df	MS	F
Between				
Teachers	138.00	15	9.200	
Within				
Time	1.292	2	.646	.660
Error	29.375	30	.979	
Total	168.667	47		

Complex Thinking. As shown in Table 22, the time (Huynh-Feldt) was not

significant, implying that the teachers' attitudes scores for Complex Thinking did not

change over the professional development period.

Table 22

Repeated Measures ANOVA with Complex Thinking

Source	SS	df	MS	F
Between				
Teachers	329.452	13	25.342	
Within				
Time	11.752	2	5.881	2.219
Error	68.905	26	2.650	
Total	410.109	41		

Instructional Conversation. As shown in Table 23, the time (Huynh-Feldt) was not

significant, implying that the teachers' attitudes scores for Instructional Conversation did

not change over the professional development period.

Table 23

Repeated Measures ANOVA with Instructional Conversation

Source	SS	df	MS	F
Between				
Teachers	206.976	13	15.921	
Within				
Time	13.905	2	6.952	1.806
Error	100.095	26	3.850	
Total	366.80	44		

Research Question 5: To what extent did teachers' instructional practices change over

time based on results from their videotaped lessons using the Classroom Observation

Rubric?

   Video Ratings. The inter-rater reliability was used to assess the consistency of

the video ratings. As shown in Table 24, in general, results reflect that coders were

consistent in their individual ratings, with the exception of Contextualization at Time 2

(-.240) and Time 3 (.311) and Complex Thinking (.500) at Time 3.

Results from the descriptive statistics, shown in Table 25, reflect little mean change from Time 1 to Time 4. Time was not a statistically significant predictor of change in teachers' mean observed video scores over the four time periods, except for Instructional Conversation. Figure 3 shows the range of mean scores of the video ratings using the COR.

Table 24

Intraclass Correlations for Two Coders on Video Ratings

Standard	Time 1	Time 2	Time 3	Time 4
JPA	.903	.971	.786	.977
LLD	.881	.706	.832	.824
CTX	.940	-.240	.311	.857
CT	.935	.949	.500	.706
IC	.957	.968	.938	.976

Table 25

Descriptive Statistics of Video Ratings

Standard		Time 1	Time 2	Time 3	Time 4
JPA	M	3.00	3.13	4.00	3.00
	SD	.926	1.46	1.77	1.51
LLD	M	3.63	3.63	4.25	4.00
	SD	.916	.916	1.17	.926
CTX	M	2.00	2.00	1.88	2.00
	SD	1.07	1.31	.835	.926
CT	M	2.88	3.00	3.63	3.00
	SD	.991	1.31	.744	1.07
IC	M	1.00	.875	3.00	1.00
	SD	1.42	1.64	1.93	1.60

In the following tables, the repeated measures ANOVA results are presented for each standard. These scores were used to determine whether teachers' mean scores changed in their observed practice during the professional development period. The

scores included in these analyses are the reconciled scores between coders, not their individual scores.

Joint Productive Activity. As shown in Table 26, time was not statistically significant (using Huynh-Feldt), implying that teachers' observed behaviors did not change over time.

Language and Literacy Development. Results for Language and Literacy Development, as reflected in Table 27, show that there were no significant changes in teachers' performances over time (Huynh-Feldt).

Table 26

Repeated Measures ANOVA with Joint Productive Activity

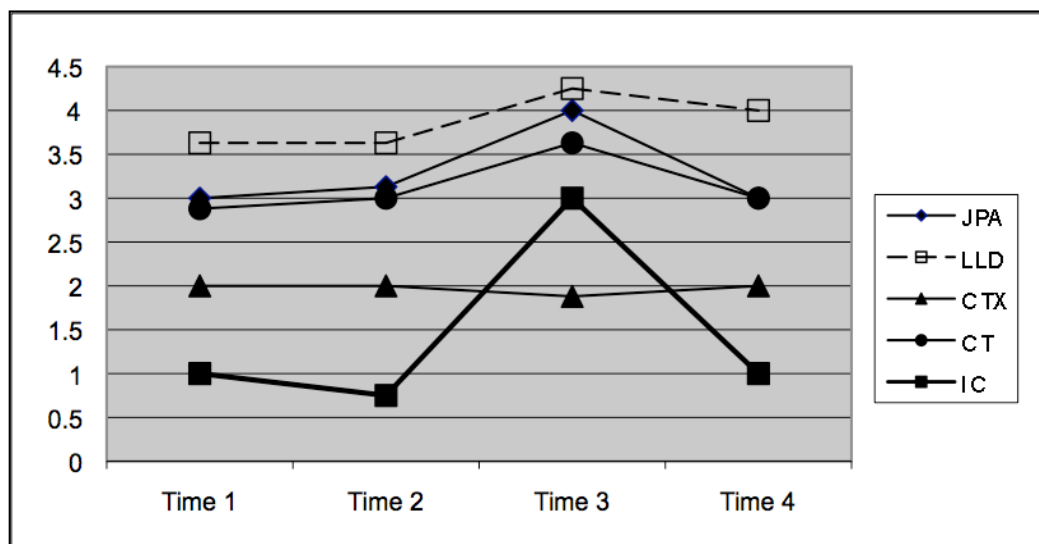| Source	SS	df	MS	F
Between				
Teachers	45.719	7	6.531	
Within				
Time	5.594	3	1.865	2.976
Error	13.156	21	.626	
Total	64.469	33		

Figure 3. Observed Video Ratings for Each CREDE Standard on Four Occasions

Table 27

Repeated Measures ANOVA with Language and Literacy Development

Source	SS	df	MS	F
Between				
Teachers	14.500	7	2.071	
Within				
Time	2.250	3	.750	1.235
Error	12.750	21	.607	
Total	29.5	33		

Contextualization. As shown in Table 28, the time (Huynh-Feldt) was not significant, implying that teachers observed behaviors for Contextualization did not change over the four time periods.

Table 28

Repeated Measures ANOVA with Contextualization

Source	SS	df	MS	F
Between				
Teachers	16.375	7	2.339	
Within				
Time	5.125	3	1.708	.827
Error	43.375	21	2.065	
Total	64.875	33		

Complex Thinking. As shown in Table 29, the time (Huynh-Feldt) was not significant, implying that the teachers observed video recording did not change over time.

Table 29

Repeated Measures ANOVA with Complex Thinking

Source	SS	df	MS	F
Between				
Teachers	16.000	7	6.531	
Within				
Time	2.750	3	.917	1.305
Error	14.750	21	2.286	
Total	33.5	33		

Instructional Conversation. As shown in Table 30, time (Huynh-Feldt) was significant, suggesting that the teachers' video ratings for Instructional Conversation changed over the professional development period. Subsequent post-hoc analyses using LSD revealed that the average Instructional Conversation scores at Time 2 were significantly higher than Time 1, Difference = -.875, *p* = .041 and the average at Time 3 was significantly higher than at Time 1, Difference = -.125, *p* = .041.

Table 30

Repeated Measures ANOVA with Instructional Conversation

Source	SS	df	MS	F
Between				
Teachers	32.719	7	4.674	
Within				
Time	25.094	3	8.365	3.978**
Error	44.156	21	2.103	
Total	101.969	33		

**p<.01

Research Question 6: To what extent did teachers reliably respond to the survey designed to measure teachers' understandings of the CREDE standards?

Classroom Vignette. Teachers' understanding of the CREDE standards was measured using the Classroom Vignette survey. Reliability coefficients were calculated for each of the standards over the four points in time for the Vignette Survey (see Table 31). These reliabilities were low, even at Time 4 when teachers would be expected to have mastered the standards and hence their responses consistent. Results indicate that other than Joint Productive Activity at Time 1 (.691) and Time 4 (.731), there was little internal consistency of items on the Classroom Vignette survey.

Table 31

Cronbach's Alpha for Classroom Vignette Survey

Standard	Time 1	Time 2	Time 3	Time 4
JPA	.691	.010	.348	.731
LLD	.432	.212	.000	.000
CTX	.468	.420	.492	.499
CT	.210	.423	.000	.075
IC	.337	.446	.000	.287

Alternatively, the survey results were analyzed by calculating the percent of teachers who answered each item correctly. Seven of the items were dropped due to inconsistent responses from teachers. For example, I dropped Items 5 and 8, which none of the teachers answered with the intended correct response. The items for which at least 70% of teachers responded correctly are included in Table 32. For example, at Time 1, of the 10 items at least 70% of teachers responded correctly to items 1, 2, 14, 15, 16 and 17. Likewise, at Time 2, at least 70% of teachers responded correctly to items 1, 2, 6, 12 and 17. Item responses for Time 3 mirror Time 2, with the exception of one additional item, 16. By Time 4, all remaining items except for two (6 and 9) had at least 70 % correct responses.

Table 32

Items with Partially or Fully Correct Responses Greater than Seventy Percent

Items	Time 1	Time 2	Time 3	Time 4
1	1	1	1	1
2	2	2	2	2
6	.	6	6	.
9	.	.	.	.
10	.	.	.	10
12	.	12	12	12
14	14	.	.	14
15	15	.	.	15
16	16	.	16	16
17	17	17	17	17

The percentages for each item response across the four time periods are included in Table 33 below. The greatest percentages of correct responses occurred at Time 4.

Combined Reliabilities Between K-12 Teacher Participants and CREDE Expert Instructors. In Table 34, reliability results are presented from the practices, attitudes and vignette surveys at Time 4 for teacher participants and the CREDE expert instructors.

Table 33

Percent of Partially or Fully Correct Responses Greater than Seventy Percent

Items	Time 1	Time 2	Time 3	Time 4
1	76.5	88.2	88.2	73.3
2	94.1	94.1	100.00	100.00
6	.	75.0	73.3	.
9	.	.	.	.
10	.	.	.	80.0
12	.	88.2	70.6	73.3
14	82.4	.	.	73.3
15	70.6	.	.	73.3
16	70.6	.	76.5	86.7
17	82.4	70.6	70.6	86.7

Table 34

Cronbach's Alpha Coefficients for Teachers and Intraclass Correlations for CREDE Experts at Time 4

	Teaching Practices		Attitudes Toward Teaching		Classroom Vignette	
	Teachers	CREDE Experts	Teachers	CREDE Experts	Teachers	CREDE Experts
JPA	.734	.772	.770	.977	.731	.726
LLD	.636	.962	.357	.921	0	.568
CTX	.686	.971	.901	.426	.499	.675
CT	.824	.881	.767	.862	.075	.567
IC	.663	.780	.370	.818	.287	.703

Chapter 5

Discussion

There were two purposes for this study. One purpose was to examine whether teachers' self-reported responses on surveys about their practices, attitudes, and understandings of the CREDE standards were consistent. The second purpose was to compare results from these three measures to those from a video observation rubric used in past CREDE research.

CREDE Expert Instructors

There were two sets of reliability results from the three surveys, one set from the teacher participants and the other from CREDE expert instructors. Before discussing the teachers' results, presented in the multitrait-multimethod matrix, I will briefly discuss the survey responses from CREDE experts whose responses were fairly consistent. One strategy to determine the validity of an instrument is to ask "experts" to help judge the degree to which a particular measurement instrument seems to measure what it is supposed to (Tashakkori & Teddlie, 1998). Tashakkori and Teddlie refer to this as "judgmental validity." In general, the experts were able to identify the CREDE standards, particularly on the practices and attitudes surveys, with higher intraclass correlation coefficients than the teacher participants. For some items, all four experts, or at times three of them, were able to accurately identify items measuring the different standards on the surveys. The combination of their many years of experiences (a) using the CREDE standards in their own practices, (b) teaching the standards to others, and (c) maintaining positive attitudes about the standards, suggest that time and experience play a role in teachers use of, attitudes toward, and understanding of the standards. This

combination of knowledge and skill may explain why the experts had better results on the classroom vignette survey compared to the teacher participants. Unlike the practices and attitudes surveys, the vignette survey was not a scaled measure and required pedagogical and subject-matter knowledge, in this case, the CREDE standards. These conclusions about CREDE experts are consistent with Borko and Putnam (1995) who emphasized teacher knowledge and Shulman (1987, 2004), who described pedagogical knowledge and subject-matter knowledge as important elements of good instruction.

Convergent Validity Results

Convergent validity results in the heterotrait-heteromethod submatrix between Method 1, practices survey and Method 3, video ratings, were significant for Joint Productive Activity at Time 4.

Of the CREDE standards, Joint Productive Activity may be the most familiar to teachers. Whether in an early childhood environment (Yamauchi & Kuwahara, 2008), a middle school (Gillies, 2004), or high school classroom (Slavin, 1987), the idea of teachers' working closely together with students, through collaboration, cooperative learning, or small group activities, is a strategy that is comfortable for many teachers. Reliability results across all three survey measures—practices, attitudes, and vignettes— were acceptable at Time 4 for Joint Productive Activity for both teacher participants' and expert instructors' responses. One reason for this may be because this standard is the easiest to identify through observation, considering that by definition, it implies "group work." The alternative surveys either explicitly, or subtly implied group work. On the practices survey, Item 1 states, "I organize lessons so that students can work together in small groups." On the attitudes survey, the parallel item states, "It is important for me to

organize lessons so that students can work together in small groups." Finally, Item 1 on the vignette survey states:

> Ms. Takei is a 3rd grade teacher who has 24 students in her classroom. Many of them are mixed-ethnicities, coming from different cultural backgrounds. For today's mathematics lesson Ms. Takei has asked her students to work together in pairs, having student take turns drilling each other on multiplication and division problems using flashcards (p. 1).

In general, both the teacher participants and CREDE expert instructors were able to identify the Joint Productive Activity standard on the surveys, and the raters were able to do so using the COR. Considering the small sample size of teacher participants, the convergent validity results for this standard are encouraging.

Non-Convergent Validity Results

The overall results presented in the multitrait-multimethod submatrix, examining the relationship between the practices and attitudes surveys and video ratings did not reflect convergent validity.

The results from the heterotrait-heteromethod submatrix between Methods 1, practices and 3, video ratings, indicated that three of the correlations were not significant and that the survey did not capture the same constructs as the video observation rubric. The lack of significant findings may be due to a number of reasons.

First, the results may be due to a method effect, indicating that each instrument was tapping different domains and teachers were not asked the same question as the observers. The practices survey asked teachers to report on their overall teaching behaviors, whereas, the video ratings were from behaviors that teachers used during a

specific lesson. Responses on the practices survey asked teachers to "best describe" their

current teaching practices using the CREDE standards, ranging on a scale from "never"

to "daily." In contrast, the COR asked raters to score teachers' use of the CREDE

standards on the observed videotaped lesson by level, "not observed, emerging,

developing, advancing, enacting," and "exemplary." Some research suggests that

although videotaped lessons can be helpful in modifying teachers' behaviors, there is also

evidence that behaviors tend to change when the camera is removed from the classroom,

such that teachers' revert back to the practices that they were most comfortable using

(Hawkins & Heflin, 2011). Other research suggests that although observations are often

expensive and require specialized training, they are accepted as reliable (Koziol & Burns,

1986). In contrast, teacher self-reports are a cost-effective method for assessing

classroom behaviors, although results may be less reliable.

In addition to differences between methods, the time at which the data were

collected may partly explain the reason for the low convergent validity. Koziol and

Burns (1986) refer to this problem as a "mismatch of observation schedules" (p. 205).

The mismatch occurs when teachers are asked to report on their practices either before, or

after observers visited their classroom. In this case, teachers reported on general

practices, which may not have matched the practices observed during their videotaped

lesson.

Along these same lines, teacher participants were not given the option of when

they wanted to complete the surveys (all three surveys were administered simultaneously

to all teachers), whereas they selected when they wanted to be videotaped within a given

timeframe However, just as self-selecting a time to be videotaped poses biases, teacher

self-reporting on the surveys and the concern for social desirability presents other biases. Teachers were given approximately 15-20 minutes to complete the surveys, compared with potentially hours of preparation time for their videotaped lessons.

These methodological and time issues and also self-reported and self-selected biases across Method 1 and Method 3 are possible reasons for why convergent validity results were low for four of the CREDE standards.

Results from the heterotrait-heteromethod submatrix between Method 1, practices and Method 2, attitudes, indicated negative convergent validity for three of the standards. There are a number of reasons for this. The first consideration is social desirability. For example, Item 6 on the attitudes survey states, "In-class writing assignments are an important part of my regular instruction." Item 11 states, "I believe that it is important for me to focus my instruction on students' linking prior knowledge with new facts and concepts." It is unlikely that a teacher, particularly one who voluntarily signed up for a course advertised as "Professional Development Course for Teachers of At-Risk Native Hawaiian Students," would "disagree," or "strongly disagree" with either of these statements. Yet, despite agreeing with the statements, teachers responded that they do not regularly use these strategies. Tashakkori and Teddlie (1998) refer to this type of reactivity as "acquiescence bias" or agreement, or "positivity bias," when respondents tend to agree with items, especially those on the positive side of the response option.

This argument leads to another possible reason for the negative convergent validity results across the two methods. Despite teachers' positive attitudes toward the different standards, they may not have been quite sure how to implement them. Teachers' may not have known what each standard looked like in practice, because they

did not have sufficient opportunities to observe classrooms relevant to their own instructional contexts. Teachers were shown video samples of CREDE classrooms during the first workshop; however, the videos were selected because they were examples of high levels of enactment on all or most of the CREDE standards. Therefore, the benefits for a high school teacher observing a CREDE elementary classroom, or vice versa, may not have been as helpful as observing a classroom teaching the same subject or at the same instructional level.

Finally, results from the heterotrait-heteromethod submatrix between Method 2, attitudes, and Method 3, video ratings, were not significant. These results were not expected to be significant because teachers reported their attitudes on the survey, which measured the extent to which they agreed or disagreed with various strategies of teaching and learning. The COR, however, was designed to measure the extent to which teachers used the strategies.

Discriminant Validity

In terms of discriminant validity measuring theoretically different constructs, results between the practices and video methods were consistent and did not highly correlate with one another across measures. However, discriminant validity results on the practices survey were at significant levels. These results suggest that the teachers did not easily differentiate among the CREDE standards. Another possibility is that the standards measure one underlying dimension. The video ratings, with the exception of Complex Thinking and Language and Literacy Development reflect little discriminant validity, which we would expect from coders who had been trained to use the COR. The attitude survey produced results that were not significant, except for the negative

correlation between Contextualization and Joint Productive Activity, which I cannot explain.

Reliabilities

Teacher Participants and Coders. In general, reliability results are highest on the video ratings, then on the practices survey, and finally on the attitudes survey. There is one exception for Contextualization on the attitudes survey, which was the highest at .901. As mentioned earlier, the purpose of the course, and the program that funded it, was to improve educational opportunities for Native Hawaiians. Arguably a good starting place is with the idea of connecting what students are learning in school, with what they already know from their homes. One of the criteria for enrolling in the course was that teachers were working in schools with high concentrations of Native Hawaiian students. Therefore, it is unlikely that a teacher would not respond favorably to any of the items reflecting Contextualization, because it was likely something that led them to apply for the program.

Similar Findings Across Instruments. Mean scores on the practices and attitudes surveys increased from Time 1 to Time 4 for Joint Productive Activity and Contextualization. However, video rating results indicated no change in mean scores for Joint Productive Activity (3.0) and Contextualization (2.0) at Time 1 and Time 4.

As previously mentioned, Contextualization was emphasized at the beginning of the recruitment process. Then, during the first full day workshop, there were many discussions around this standard, as well as samples of videotaped lessons illustrating what this and the other standards looked like in practice. As mentioned above, most teacher participants applied to the professional development course because they were

interested in learning alternative strategies for working with culturally diverse students,

particularly, Native Hawaiians, which may help explain the high reliability for the

attitudes survey. For the practices survey, the reliability results by Time 4 were

acceptable, but still not as high as those on the attitudes survey. Of all of the reliability

results for the video ratings, the two lowest correlations were for Contextualization at

Times 2 and Time 3. These low results may be artificial and due to the method of

calculating intraclass correlations. They could also be because most lessons were

between 45 minutes and one hour, spread out four times during the year. So depending

on where teachers were on the day of their videotaping within a particular lesson or unit,

a deliberate connection may not have been observed. Unless a teacher specifically began

by referring to an earlier lesson and perhaps reviewed [on camera] what students' learned

already, with what they already know, teachers may not have received high ratings for

this particular standard.

There were four items that were correctly answered by both teachers and experts

on the Vignette survey at Time 1[11]. Seventy percent of teachers and all of experts

identified two items, one that measured Joint Productive Activity and the other measured

Contextualization. Additionally, 70% of teachers and 75% of experts correctly answered

items that measured Complex Thinking and another that measured Joint Productive

Activity. The teachers completed the vignette, as well as the other three surveys

(practices, attitudes, and background) at the end of the second day of their first

introductory workshop. They were asked not to use the SPC, but answer as best as they

---

[11] Time 1 was compared because the CREDE expert instructors only completed the
surveys on one occasion.

could without using the rubric. Part of the reason for their correct responses may be because they were immersed in discussions about the standards for two days.

Results reflecting change over time were not consistent across measures. Language and Literacy Development and Complex Thinking were significant on the practices survey and Instructional Conversation was significant using the COR. As previously discussed, neither Joint Productive Activity nor Contextualization produced significant results. Instead, the remaining three standards were significant. These results were consistent with the argument that Joint Productive Activity and Contextualization were teaching strategies about which teacher participants were already familiar. It might also be that teachers believed they were the "easiest" in terms of modifying instructional practice, compared to the other standards.

Joint Productive Activity at the enacting level is indicated by the teacher and a small group of students collaborating on a joint product. This criterion does not establish a level of quality of the discussion, just that the teacher and students are working together. For Contextualization, the teacher integrates the new activity/information with what students already know from home, school, or community. By comparison, enacting levels for the three standards that produced significant results were more content specific and required the teachers' assistance. For example, Language and Literacy Development at the enacting level requires the teacher to design and enact instructional activities that generate language expression and development of content vocabulary and assists student language use or literacy development through questioning, rephrasing or modeling. Likewise, Complex Thinking at the highest level requires the teacher to design and enact

challenging activities with clear standards and performance feedback, and assists the development of more complex thinking.

Finally, to establish an Instructional Conversation at the enacting level a teacher must design and enact an instructional conversation with a clear academic goal; listen carefully to assess and assist student understanding; and question students on their views, judgments, or rationales. Student talk must occur at higher rates than teacher talk. The three standards that were significant are consistent with the NCLB mandates and what many schools, especially those in restructuring are working toward, which is for teachers to assist students in becoming proficient on state academic achievement standards and academic assessments (NCLB Act, 2001).

Limitations

The first limitation of this study is the small sample size, which affected the statistical power of the analyses. There were only fifteen teachers with complete data sets across the three surveys, and only seven of these with completed video ratings.

Another limitation is that teachers were not trained to the same degree as the two coders who worked toward establishing inter-rater reliability. Except for their workshop video presentations, which included the opportunity for the other teachers to "rate" their video, the teachers did not have the opportunity to establish inter-rater reliability on measures of the standards.

Finally, although the items were developed from the Standards Performance Continuum and the Indicators Criteria, there was evidence that some items were not well defined. The most obvious were on the vignette survey. For example, Item 5 was developed to reflect a non-CREDE standard, but was almost unanimously identified as

Contextualization by both teachers and CREDE experts. Another example, Item 3 on the same survey was developed to measure Contextualization; however, three experts believed that the vignette illustrated Language and Literacy Development, one expert believed that the scenario illustrated Instructional Conversation, and the item was dropped from the vignette analyses due to the low reliability results.

Future Research

Of the three surveys developed for this study, the practices survey may warrant further research. The item construction and scales should be more closely examined, based on findings in this study. One possibility would be to have teachers complete the survey on the same day as they are videotaped and change the scale to represent the same observed video recording.

Another consideration based on the expert survey responses is for a CREDE expert to use the survey on the same day of observation, and also have the teacher complete the survey and then compare the two findings, using the method as more of a formative assessment process. Later, these results could be compared with the video ratings to see how closely they aligned to produce summative findings, such as how much have teachers changed over time using the standards.

Another consideration is to spend greater time training teachers to understand each of the Standards using relevant video examples and also providing more opportunities to use the CREDE rubric and determine the reliability of teachers' self-ratings. With these considerations in mind, designing a study using a control and experimental group to identify practices that CREDE professional development participants may already be using in their classrooms, but may not have named them,

could guide further research in strengthening the CREDE professional development model.

Finally, in future workshops, perhaps pausing the video after each enacted standard, followed by a discussion about what was observed might provide a stronger visual example for teachers to reflect back upon on what the standards each standard looked like in practice.

My Reflections

Teaching is a complex process that involves many components and therefore, teacher change is not always easy to measure. In this study, results from the surveys and the video ratings indicated that teachers' practices did not significantly change during their participation in the yearlong professional development course. Because the video ratings were based on the modified version of the SPC, a validated measure of the CREDE standards, these results—which do not contradict the teachers' self-reported survey responses—have led me to wonder why teachers' ratings did not improve over time. Were some of the teachers' already implementing some of the standards at Time 1? Are four 1-hour videos collected over a one-year period an accurate, representative sample reflecting teachers' change over time? Finally, might the professional development model be better suited for younger children than for high school students?

Based on the mean score of the video rating results, there was little change in teachers' performance using the standards from Time 1 to Time 4. For Joint Productive Activity at Time 1, the mean score was 3.00 (advancing), which was also the mean score at Time 4. The same pattern was reflected for Contextualization (mean score of 2.00, developing) and Instructional Conversation (mean score of 1.00, emerging) at Times 1

and 4. This supports the idea that teachers' were in fact implementing the standards to some degree at the beginning of the school year, which leads to the next question, why did teachers' performance not change?

Teachers' mean scores did change at Times 2 and 3; some mean scores increased, others decreased, although ultimately by Time 4, the scores ended up where they began. These results suggest that it may take longer than one year for teachers to change their practices. This project, funded by the Native Hawaiian Education Act and similar to many other federally funded programs, offer 3-year contracts. For many of these projects, Year 1 is a planning and preparation year, Year 2 is program implementation, including data collection, and Year 3, a combination of data analyses and efforts toward improving continued research within a particular domain. Results from this professional development study support the notion that in order to accurately examine teacher change, it may be necessary to collect data over multiple years, rather than a single year.

Finally, there may be significant differences between high school teachers and those who work in early childhood and elementary schools. First, early childhood and elementary school teachers tend to teach all subjects and the same group of students on a given day, compared to high school teachers who teach one subject and multiple classes (i.e., many students) throughout the day. Teacher preparation programs also vary, with high school teachers' concentrating on their specialized subject area—mathematics, science, language arts, social studies—compared with early childhood and to some extent, elementary school teachers' whose training not only focuses on content, but also on appropriate instructional strategies based on children's developmental stages. For this

reason, school context and teacher preparation programs may influence teachers' abilities to implement the CREDE standards.

References

AERA, APA, & NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education) Joint Committee on Standards for Educational and Psychological Testing (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Benham, M. K. P., & Heck, R. H. (1998). *Culture and educational policy in Hawai'i: The silencing of native voices*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Berliner, D. C. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-21.

Berliner, D. C., & Biddle, B. J. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley.

Borko, H., & Putnam, R. T. (1995). Expanding a teacher's knowledge base: A cognitive psychological perspective on professional development. In T. Guskey, & M. Huberman (Eds.), *Professional development in education: New paradigms and practices* (pp. 35-65). New York, NY: Teachers College Press.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*, 3-15.

Campbell, D. T., & Fiske D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 54*, 297-312.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C. P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26*(4), 499-531.

Cash, J. (2007). The "s" word: Ten years later. *Leadership, 37*(1), 8-12.

Cohen, D. K. & Barnes, C. A. (1993). Conclusion: A new pedagogy for policy. In D. Cohen, M. McLaughlin, & J. Talbert (Eds.), *Teaching for understanding: Challenges for policy and practice* (pp. 240-275). San Francisco, CA: Jossey-Bass.

Common Core State Standards. (2011). *The Standards*. Retrieved from http://www.corestandards.org/the-standards.

Dalton, S. D. (1998). *Pedagogy matters: Standards for effective teaching practice*. Santa Cruz, CA: University of California, Center for Research on Education, Diversity & Excellence.

Dalton, S. D. (2008). *Five standards for effective teaching: How to succeed with al learners*. San Francisco, CA: Jossey-Bass.

Darling-Hammond, L., & Ball, D. L. (1997). *Teaching for high standards: What policy-makers need to know and be able to do*. Philadelphia, PA: Consortium for Policy Research in Education.

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement. *Education Policy Analysis Archives*, *North America*. Retrieved from http://epaa.asu.edu/ojs/article/view/392/515.

Darling-Hammond, L. (2005). *Professional development schools: Schools for developing a profession*. New York, NY: Teachers College Press.

Darling-Hammond, L. (2006a). *Powerful teacher education: Lessons from exemplary programs*. San Francisco, CA: Jossey-Bass.

Darling-Hammond, L. (2006b). Assessing teacher education: The usefulness of multiple

    measures for assessing program outcomes. *Journal of Teacher Education*, *57*(2),

    pp. 120-138.

Demmert, W. G., & Towner, J. C. (2003). *A review of the research literature*

    on the influences of culturally based education on the academic performance of

    Native American students. Northwest Regional Educational Laboratory, Portland

    Oregon. Retrieved from http://www.nwrel.org/indianed/cbe.pdf.

Demmert, W. G., Grissmer, D., & Towner, J. C. (2006). A review and analysis of the

    research on Native American students. *Journal of American Indian Education -*

    Special Edition, *45*(3), 5-23.

Desimone, L. M. (2009). Improving impact studies of teachers' professional

    development: Toward better conceptualization and measures. *Educational*

    Researcher, 38(3), 181-199.

Doherty, R. W., Hilberg, R. S., Epaloose, G., & Tharp, R. G. (2002). Standards

    performance continuum: Development and validation of a measure of effective

    pedagogy. *Journal of Effective Pedagogy*, *96*(2), 78-89.

Doherty, R. W., & Hilberg, R. S., Pinal, A., & Tharp, R. G. (2003). Five standards and

    student achievement. *Journal of Research and Practice, 1*(1), 1-24.

Education Government. (2010). *Nine states and the District of Columbia win second*

    round race to the top grants. Retrieved from

    http://www.ed.gov/news/press-releases/nine-states-and-district-columbia-win-

    second-round-race-top-grants.

Elmore, R. F., & McLaughlin, M. W. (1988). *Steady work: Policy, practice, and the*

reform of American education. The RAND Corporation, National Institute of Education. Santa Monica: CA.

Engstrom, M. E., & Danielson, L. M. (2006). Teachers' perceptions of an on-site staff development model. *A Journal of Educational Strategies, Issues, and Ideas, 79*(4), 170-173.

Estrada, P. (2004). Patterns of language arts instructional activity and excellence in first and fourth grade culturally and linguistically divers classrooms. In H. Waxman, R. Tharp & R. Hilberg (Eds.) *Observational research in U.S. classrooms: New approaches for understanding cultural and linguistic diversity* (pp. 122-143). Cambridge, MA: Cambridge University Press.

Farmer, J., Hauk, S., & Neumann, A. M. (2005). Negotiating reform: Implementing process standards in culturally responsive professional development. *High School Journal*, *88*(4), 59-71.

Finch, J. (1987). The vignette technique in survey research. *Sociology, 21*, 105-114.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., &Yoon, K. S. (2001). What makes professional development effective? Results from a national sample or teachers. *American Educational Research Journal, 38*(4), 915-945.

Gillies, R. M. (2004). The effects of cooperative learning on junior high school students during small group learning. *Learning and Instruction 14(*2), 197-213.

Goldhaber, D. D., & Brewer, D. J. (1997). Evaluating the effect of teacher degree level on educational performance. In W. Fowler (Ed.), *Developments in school finance* (pp. 197-210). Washington, DC: U. S. Department of Education, National Center for Education Statistics.

Goldhaber, D. D. (2002). The mystery of good teaching. *Education Next 2*(1), 50-55.

Guskey, T. R. (1995). Professional development in education: In search of the optimal

mix. In T. Guskey & M. Huberman (Eds.), *Professional development in*

education: New paradigms and practices (pp. 114-131). New York, NY:

Teachers College Press.

Guskey, T. R. (2002). Professional development and teacher change. *Teachers and*

Teaching: Theory and Practice, 8(3/4), 381-391.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in the

public schools. *Journal of Economic Literature, 24*(3), 1141-1178.

Hanushek, E. A. & Raymond, M. E. (2005). Does school accountability lead to

improved student performance? *Journal of Policy Analysis and Management,*

24(2), 297-327.

Hao, R. N. (1988). Research and development model for improving reading instruction.

In S. Samuels & P. Pearson (Eds.), *Changing school reading programs:*

Principles and case studies (pp. 95-118). Newark, Delaware: International

Reading Association.

Hao, R. N., & Hartley-Forsyth, P. (1993). Reading and writing across the preservice

curriculum. *Kamehameha Journal of Education, 4*, 97-104.

Hawai'i Department of Education. (2009). *Continuous improvement: Building for the*

future. Superintendent's 20ᵗʰ Annual Report. State of Hawai'i: Author.

Hawai'i Educational Policy Center (2008). *Report to the legislature on senate*

concurrent resolution 56 S.D.1 requesting the Hawai'i Educational Policy Center

to report on the retention and change in assignment of teachers within the

department of education. Retrieved from http://www.hawaii.edu
/hepc/pdf/Reports/SCR_56_HEPC_Interim_Report_0108.pdf.

Hawkins, S., & Heflin, L. J. (2011). Increasing secondary teachers' behavior-specific praise using video self-modeling and visual performance feedback intervention. *Journal of Positive Behavior Interventions, 13*(2), 97-108.

Hilberg, R. S., Tharp, R. G., & DeGeest, L. (2000). The efficacy of CREDE's standards-based instruction in American Indian mathematics classes. *Equity and Excellence in Education, 33*(2), 32-39.

Hilberg, R. S., Doherty, R. W., Epaloose, G., & Tharp, R. G. (2004). The Standards Performance Continuum: A performance-based measure of the standards for effective pedagogy. In H. Waxman, R. Tharp, & R. Hilberg (Eds.), *Observational research in U. S. classrooms: New approaches for understanding cultural and linguistic diversity* (pp. 48-71). Cambridge, UK: Cambridge University Press.

Jorgensen (Zevenbergen), R., Grootenboer, P., Niesche, R., & Lerman, S. (2010). Challenges for teacher education: The mismatch between beliefs and practice in remote indigenous contexts. *Asia-Pacific Journal of Teacher Education, 38*(2), 161-175.

Kahumoku, W., & Kekehio, W. (2010). The Kahua Induction Program: Systemically supporting teachers through culturally relevant, place-based, and community mentor strategies. *Hulili: Multidiscilpinary Research on Hawaiian Well-Being 6,* 213-236.

Kanaʻiaupuni, S. M., & Ishibashi, K. (2003, June). *Left Behind? The status of*

Hawaiian students in Hawai'i's public schools. (No. PASE Report 02-03:13, Honolulu, HI: Kamehameha Schools.

Kana'iaupuni, S. M. & Kawai'ae'a, K. (2008). E lauhoe main a wa'a: Toward a Hawaiian indigenous education teaching framework. *Hulili: Multidiscilpinary Research on Hawaiian Well-Being 5,* 67-90.

Kana'iaupuni, S., Ledward, B., & Jensen, U. (2010). *Culture-based education and its relationship to student outcomes*. Honolulu: Kamehameha Schools, Research & Evaluation. Retrieved from http://www.ksbe.edu/spi/PDFS/CBE_relationship_to_student_outcomes.pdf.

Kekahio, W. (2007). Native Hawaiians in public schools: Implications of AYP in predominantly Native Hawaiian schools. Honolulu, HI: Kamehameha Schools.

Koziol, S. M., Jr., & Burns, P. (1986). Teachers' accuracy in self-reporting about instructional practices using a focused self-report inventory. *Journal of Education Research, 79*(4), 205-209.

Ladson-Billings, G. (1995a). But that's just good teaching! The case for culturally relevant pedagogy. *Theory Into Practice, 34*(3), 159-165.

Ladson-Billings, G. (1995b). Multicultural teacher education: Research, practice, and policy. In J. Banks & C. Banks (Eds.), *Handbook of research on multicultural education* (pp. 747-759). New York: Macmillan.

Lemlech, J., Hertzog-Foliart, H., & Hackl, A. (2005). The Los Angeles professional practice school: A study of mutual impact. In L. Darling-Hammond (Ed.), *Professional development schools: Schools for developing a profession* (pp. 156-175). New York: Teachers College Press.

Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from www.tqsource.org/publications/ practicalGuide.pdf.

Loucks-Horsley, S., Harding, C. K., Arbuckle, M. A., Murray, L. B., Dubea, C., & Williams, M. K. (1987). *Continuing to learn: A guidebook for teacher development.* Massachusetts and Ohio: The Regional Laboratory for Educational Improvement of the Northeast Islands and The National Staff Development Council.

Luning. R. (2011). *Classroom Observation Rubric: Development and validation of a measure of effective pedagogy.* Unpublished manuscript. Honolulu, HI: University of Hawaiʻi.

Mayer, D. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, *21*(1), 29-45.

Mabry, L., Poole, J., Redmond, L., & Schultz, A. (2003). Local impact of state testing in Southwest Washington. *Education Policy Analysis Archives, 11*(22), 1-37.

McCarthy, B. (1981). *The 4MAT system: Teaching to learning styles with right/left mode techniques.* Barrington, Ill: Excel Incorporated.

McCarthy, B. (1985). What 4Mat Training Teaches Us about Staff Development? *Educational Leadership*, *42*(7), 61-68.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), p 741.749.

Miller, L., & Silvernail, D. L. (2005). Wells Junior High School: Evolution of a professional development school. In L. Darling-Hammond (Ed.), *Professional development schools: Schools for developing a profession* (pp. 28-49). New York: Teachers College Press.

Milner, H. (2011). Culturally relevant pedagogy in a diverse urban classroom. *Urban Review, 43*(1), 66-89.

Mullens, J., & Kasprzyk, D. (1996). Using qualitative methods to validate quantitative survey instruments. In *1996 Proceedings of the section on survey research methods* (pp. 638-643). Alexandria, VA: American Statistical Association.

Native Hawaiian Education Council, with Ka Haka ʻUla O Keʻelikōlani College of Hawaiian Language, University of Hawaiʻi Hilo. (2002). *Nā honua mauli ola: Hawaiʻi guidelines for culturally healthy and responsive learning environments*. Honolulu: Author.

News Release. *Department of Education*. Retrieved from http://lilinote.k12.hi.us/STATE /COMM/DOEPRESS.NSF/a1d7af052e94dd120a2 561f7000a037c/a3d4d3f562c248900a25735d00084e49.

No Child Left Behind Act of 2001: *Improving basic programs operated by educational agencies, 20 U.S.C. 6311*. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf.

Office of Hawaiian Affairs (1994). *The Native Hawaiian data book.* Honolulu, HI: Author.

Penuel, W., Fishman, B., Yamaguchi, R., & Gallagher, L. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal, 44*(4), pp. 921-958.

Pukui, M. K. (1983). *ʻŌlelo noʻeau: Hawaiian proverbs and poetical sayings.* Honolulu, HI: Bishop Museum Press.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*(1), 4–15.

Quint, J. (2011). *Professional development for teachers: What two rigorous studies tell us.* Retrieved from http://www.mdrc.org/publications/603/full.pdf.

Reyhner, J., Gilbert, W., & Lockard, L. (Eds.). (2011). *Honoring our heritage. Culturally appropriate approaches for teaching indigenous students.* Flagstaff, AZ: Northern Arizona University.

Ruiz-Primo, M. A., & Li, M. (2002). *Vignettes as an alternative teacher evaluation instrument: An exploratory study.* Paper presented at the 2002 annual meeting of the American Educational Research Association, New Orleans (April 1-5).

Santoro, N. (2009). Teaching in a culturally diverse context: What knowledge about ʻself' and ʻothers' do teachers need? *Journal of Education for Teaching, 35*(1), 33-45.

Sato, M., Wei, R. C., & Darling-Hammond, L. (2008). Improving teachers' assessment

practices through professional development: The case of national board certification. *American Educational Research Journal, 45*(3), pp. 669-700.

Saunders, W., O'Brien, G., Lennon, D. & McLean, J. (1998). Making the transition to English literacy successful: Effective strategies for studying literature with transition students, in R. Gersten & R. Jimenez (Eds.) *Promoting learning for culturally and linguistically diverse students* (pp. 99-132). Monterey, CA: Brooks Cole Publishers.

Saunders, W., & Goldenberg, C. (1999). The effects of instructional conversations and literature logs on limited- and fluent-English proficient students' story comprehensions and thematic understanding. *Elementary School Journal, 99*(4), 277-301.

Sawchuk, S. (2010). No-proof-positive training approaches. *Education Week, 30*(11), 4-5.

Scher, L. & O'Reilly, F. (2009). Professional development for K-12 math and science teachers: What do we really know? *Journal of Research on Educational Effectiveness, 2*(3), 209-249.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46*(3), 407-441.

Shavelson, R. J. & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research, 51*, 455-498.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1) 1-22.

Shulman, L. S., & Grossman, P. (1988). *Knowledge growth in teaching: A final report to the Spencer Foundation* (Technical Report of Knowledge Growth in a Profession Research Project). Stanford, CA: School of Education, Stanford University.

Shulman, L. S. (2004). *The wisdom of practice*. San Francisco, CA: Jossey-Bass.

Singh, M. (2011). *Long-term effects of Native Hawaiian students' early academic achievement under the No Child Left Behind legislation: A multilevel cohort analysis.* Unpublished doctoral dissertation, University of Hawai'i.

Slavin, R. E. (1987). Cooperative learning and the cooperative school. *Educational Leadership, 45*(3), 7-14.

Spillane, J. P. (1999). External reform initiatives and teachers' efforts to reconstruct practice: The mediating role of teachers' zones of enactment. *Journal of Curriculum Studies, 31,*143-175.

Stecher, B., Le, V., Hamilton, L., Ryan, G., Robyn, A., & Lockwood, J. (2006). Using structured classroom vignettes to measure instructional practices in mathematics. *Educational Evaluation and Policy Analysis, 28*(2), 101-130.

Stumbo, C., & McWalters, P. (2011). Measuring effectiveness: What will it take? *Educational Leadership, 68*(4). 10-15.

Swanson, C. B. (2008). Grading the states. *Education Week, 27*(18), 36-60.

Takenaka, C. (1995). *A perspective on Hawaiians. A report to the Hawai'i Community Foundation*. Honolulu, HI: Hawai'i Community Foundation.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology, combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Teacher Quality. (2008) Hawaiʻi Department of Education. Retrieved from

http://doe.k12.hi.us/teacherquality/index.htm.

Tharp, R. G. (1982). The effective instruction of comprehension: Results and

descriptions of the Kamehameha Early Education Program. *Reading Research Quarterly, 17(*4), 503-527.

Tharp, R. G., Jordan, C., Speidel, G. E., Au, K. H., Klein, T. W., Calkins, R. P., Sloat,

K., & Gallimore, R. (1984). Product and process in applied developmental

research: Education and the children of a minority. In M. Lamb, A. Brown, & B.

Rogoff (Eds.), *Advances in developmental psychology* (Vol. 3). Hillsdale, NJ:

Lawrence Erlbaum Associates.

Tharp, R. G., Estrada, P., Dalton, S. D., & Yamauchi, L. A. (2000). *Teaching*

transformed: Achieving excellence, fairness, inclusion, and harmony. Boulder,

Co: Westview.

Tharp, R. G. (2006). Four hundred years of evidence: Culture, pedagogy and Native

America. *Journal of American Indian Education, 45* (2) 8-25.

Tharp, R. G., & Dalton, S. (2007). Orthodoxy, cultural compatibility, and universals in

education. *Comparative Education, 43*(1), 53-70.

Tharp, R. G., Jordan, C., Speidel, G. E., Au, K. H., Klein, T. W., Calkins, R. P., Sloat, K.

C. M., and Gallimore, R. (2007). Education and Native-Hawaiian children:

Revisiting KEEP. *Hulili*, 4(1), 269-305.

Thomas, S. & Heck, R. (2009). *Hawaiian cultural influences in education multilevel*

statistical analyses. Report delivered to Kamehameha Schools. Honolulu, HI:

Kamehameha Schools.

Thomas, S. & Kearney, J. (May 2008). Teachers working in culturally diverse

    classrooms: Implications for the development of professional standards for teacher

    education. *Asia-Pacific Journal of Teacher Education*, *36*(20), 105-120.

Tillema, H. H., & Imants, J. G. M. (1995). Training for the professional development

    of teachers. In T. Guskey & M. Huberman (Eds.), *Professional development

    in education: New paradigms and practices* (pp. 135-150). New York:

    Teachers College Press.

Tyack, D. & Cuban, L. (1995). *Tinkering toward utopia: A century of public school

    reform.* Harvard University Press.

U. S. Census Bureau. (2010). *2010 Census Data.* Retrieved on from

    http://2010.census.gov/2010census/data/.

U. S. Department of Education. (1996). *Building bridges: The mission and principles

    of professional development.* Washington, D.C.: Author.

U. S. Department of Education (2008). *Native Hawaiian Education Act: Greater

    oversight would increase accountability and enable targeting of funds to areas

    with greatest need.* Report to Congressional Requesters. GAO-08-422. US

    Government Accountability Office, Washington, DC 20548.

U. S. Department of the Interior & United States Department of Justice (2000). *From

    mauka to makai: the river of justice must flow freely: report on the reconciliation

    process between the federal government and Native Hawaiians.* Washington,

    D.C.: Author.

Valenzuela, A., Prieto, L., & Hamilton, M. P. (2005). No child left behind and minority

    youth: What the qualitative evidence suggests. *Anthropology and Education*

Quarterly, 38(1), 1-8.

Vogt, L. A., Jordan, C., & Tharp, R. G. (1987). Explaining school failure, producing school success: Two cases. *Anthropology & Education, 18*(4), 276-286.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological psychological processes.* Cambridge, MA: Harvard University Press.

Waxman, H. C., & Padron, Y. N. (2004). The uses of the Classroom Observation Schedule to improve classroom instruction. In H. Waxman, R. Tharp & R. Hilberg (Eds.) *Observational research in U.S. classrooms: New approaches for understanding cultural and linguistic diversity* (pp. 172-96). Cambridge, MA: Cambridge University Press.

Yamauchi, L. A. (2003). Making school relevant for at-risk students: The Wai'anae High School Hawaiian Studies Program. *Journal of Education for Students Placed At Risk, 8*(4), 379-390.

Yamauchi, L. A., Wyatt, T. R., & Taum, A. K. H. (2005). Making meaning: Connecting school to Hawaiian students' lives. *Hulili: Multidisciplinary Research on Hawaiian Well-Being, 2*, 161-178.

Yamauchi, L. A. & Kuwahara, R. H. (2008). Joint productive activity: Collaboration that builds new understandings. *Young Children 63*(6), 34-38.