

HIGH QUALITY ASSESSMENTS

Creating High Quality Assessments for a Bilingual Population

by

Leigh Sanna

Degree to be awarded December 2013

A thesis project submitted to the Department of Education and Human Development

of the State University of New York College at Brockport

in partial fulfillment of the requirements for the degree of a

Master of Science in Education

HIGH QUALITY ASSESSMENTS

Creating High Quality Assessments for a Bilingual Population

by

Leigh Sanna

APPROVED BY:

Advisor

Date

Director, Graduate Programs

Date

Table of Contents

Abstract.....	2
Chapter One: Introduction.....	3-14
Problem Statement.....	3
Significance of the Problem.....	3-8
Purpose.....	9
Rationale.....	9-11
Definition of Terms.....	11-13
Summary.....	13-14
Chapter Two: Literature Review	14-28
Chapter Three: Application.....	28-36
Chapter Four: Conclusions and Recommendations.....	37-40
Limitations.....	37-38
Recommendations.....	38-39
Conclusion.....	39-40
References.....	41-44
Appendix: Teacher Survey.....	45-46

Abstract

Lack of teacher training in test writing in both colleges and professional development programs has resulted in poorly designed assessments, particularly for culturally diverse and bilingual populations, that do not align with the high standards expected of standardized tests. Weak teacher-made assessments can result in improper student placement, ill-fitting lessons and interventions, and unnecessary frustration and confusion among test takers. These low quality tests can also serve as false predictors for student performance on standardized tests. The danger caused by skewed scores is such that if teachers do not have a reliable means by which to predict students' future success, many students will be left unprepared.

In an attempt to remedy the negative effects of poorly constructed and administered assessments, this research has created a professional development program through which teachers will be trained in writing culturally sensitive multiple-choice items and essay prompts that are aligned to the Common Core State Standards. This research and professional development program place a heavy emphasis on the importance of culturally sensitive language combined with well-crafted test items in order to insure that culturally and linguistically diverse students are provided an equal opportunity to demonstrate their knowledge on teacher-made tests. In addition, this researcher hopes that by improving the quality of teacher-made tests, educators can better prepare students for standardized tests, particularly those that are based on the new Common Core State Standards.

Creating High Quality Assessments for a Bilingual Population

Chapter 1

Problem Statement

Lack of teacher training in classroom test creation makes such tests incomparable to and incompatible with standardized tests. In order to make teacher-made tests worthwhile, they must be able to provide useful data that can inform instruction, student placement, and future assessment. If the data collected by teachers conflicts with the data gleaned from standardized tests, the classroom data will not serve as an indicator for student performance on state or nationwide tests. In addition, skewed teacher data can confuse student placement. Within the charter school used for this research, the achievement of students is based almost exclusively on teacher-made tests. These teacher-assigned scores appear on report cards and are discussed with parents as proof of student progress. In this charter school, standardized tests (specifically AIMSweb) are used to place children in appropriate intervention groups and classes as well as to evaluate the program implemented by the school.

Insufficient teacher training in test creation leads to inaccurate depictions of student knowledge. These imprecise measurements are then used to make important decisions regarding these students and can have negative effects on student learning, quality of instruction, and program success.

Significance of the Problem

This researcher conducted a five-question survey at a bilingual charter school in the city of Rochester. The charter school has an enrollment of approximately 385 students.

The student make-up is primarily Hispanic (54%) and black (45%); the remaining 1% of students are white. Students from low-income homes represent 95% of the charter school's student population, 6% are classified into Special Education, and 6% are classified as English Language Learners (Halsdorfer, 2012). This school's mission is to create students who are proficient in language arts, math, science and social studies while also learning to communicate in both English and Spanish. This researcher has taught at this charter school for 1.3 years as a primary school Spanish classroom teacher.

For this research, a random sampling of 11 teachers from the charter school was surveyed. The teachers ranged from Kindergarten to middle school. They included, more specifically, three special education teachers, one primary classroom teacher, two Spanish language arts teachers, three English language arts teachers, and two content teachers. Of these teachers, seven were from the elementary level: two teach at the Kindergarten level, two teach Grade 1, one taught Grades 1 and 2, one taught Grade 2, and one taught Grade 3. Of the remaining five teachers, one taught Grade 5, and the other four taught each grade in the middle school: Grades 6-8 (see Appendix for example of teacher survey).

The teachers responded to questions about their training in writing/creating assessments for use in their classrooms. Of the 12 respondents, five teachers cited that they have never received any training in writing assessments to use in their classrooms. The other seven teachers noted having received varying degrees of training in creation of assessments. On a scale of zero to four, zero being no training in assessment creation and four being specialized/extensive training, three teachers indicated "two", having received a moderate amount of training. Three teachers indicated "three", having received a

considerable amount of training, and one teacher indicated “four”, having received specialized/extensive training in writing assessments. These results show the mixed experiences that these teachers have had in their college preparation, pre-service training, and professional development in regards to the instruction they have received in creating assessments appropriate for their content and student population.

In the survey, teachers were asked to mark the top three challenges they confront when creating assessments. Of the options presented, the most frequently cited challenge was writing clear directions, which was chosen by seven teachers. The next challenge most often cited as important was writing culturally sensitive assessments, which was chosen by six teachers. Creating multiple-choice questions, writing clear essay prompts, and using appropriate language were each mentioned by five teachers as significant personal challenges. Only one teacher indicated that writing true/false questions was a difficulty of hers.

In order to find out what types of information or training would be most useful to this bilingual staff in regards to assessment creation, the survey asked that the teachers indicate all of the aforementioned qualities of test creation that they would be interested in learning more about, whether or not they had cited them as specific personal challenges. Seven teachers indicated they would like to learn more about writing culturally sensitive assessments, five wanted to know more about writing clear directions, four teachers stated they would like to learn about creating multiple-choice questions, and four teachers also cited wanting to know how to use more appropriate language. The

lesser-cited points of interest for teachers were learning how to write clear essay prompts and how to write true or false questions.

The final point on the survey asked teachers to reflect on what factors they believe can affect student performance on a test. They were asked to mark any and all factors that they thought were applicable. The option “wording of directions and questions” was marked 11 times as an aspect of assessments that can affect the performance of a student. “Culture of the student taking the test”, and “test format” (types of questions such as multiple choice, true/false, open-ended, etc.) were each also cited 11 times as having the potential to affect student performance. The least marked option was teacher bias, which was marked nine times by the teachers.

Although it is important to note whether or not teachers are prepared to create appropriate assessments, it is more significant to note how this lack of preparation is affecting students. For this research, data were collected from a teacher-created year-end math assessment for Grade 2, as well as from an AIMSweb standardized year-end math assessment for the same grade level. The tests measured the same Common Core math standards and are to be used together by classroom teachers to help assess student learning and place students in intervention groups for the upcoming school year. If the scores do not show similar results for each student, there is confusion as to whether the student has truly met the Grade 2 standards, and if they need to be placed in intervention for the next year. This researcher compared the scores from the teacher-made test and the AIMSweb test to find if there were any conflicting results. Through a comparison of the data collected from the Grade 2 class at the charter school, it was found that the students,

overall, performed higher on the teacher-made test than they did on the standardized AIMSweb test. As shown in Figure 1, the 11 lowest performing students on the AIMSweb test were considered “Below Average” or “Well Below Average”, and according to these results, will be placed in either Tier II or Tier III Intervention programs. These students earned scores that are considered failing. Of these 11 students, five passed the teacher-made test and are, therefore, considered “Average”. If the teacher-made tests were the only assessment used for placement, those five students would be considered benchmark and would receive no intervention services.

The scores of the 16 lowest performing students from AIMSweb were compared to the scores of these same students on the teacher-made test in terms of percentage. On average, these students performed 38% higher on the classroom test. This significant difference in performance on the two tests shows how teacher-made tests can give serve as a false indicator of student abilities and readiness. It also shows that the teacher-made tests are not properly aligned to the standards and rigor established in standardized tests.

As shown in Figure 2, the ten highest-performing students performed better on the teacher-made test than on the standardized test, as well. The final means of comparing data was to investigate the gap between the ten highest performing students and the 11 lowest performing students. On the AIMSweb test, there was a 39.6% gap between these two groups, whereas on the teacher-made test, there was only a 20.2% difference. Each of these comparisons led this researcher to the same conclusion; the teacher-made test was easier than the standardized test and therefore incompatible with the AIMSweb results.

Figure 1
Comparison of the Lowest-Scoring Students' Results on
Two Summative Mathematics Assessments

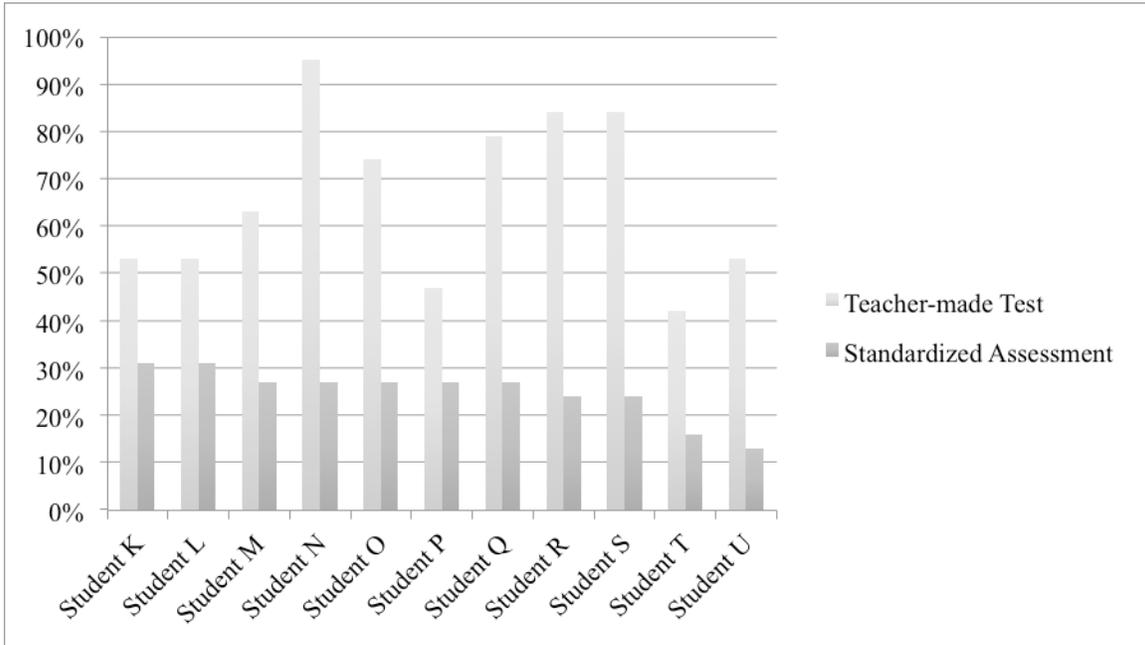
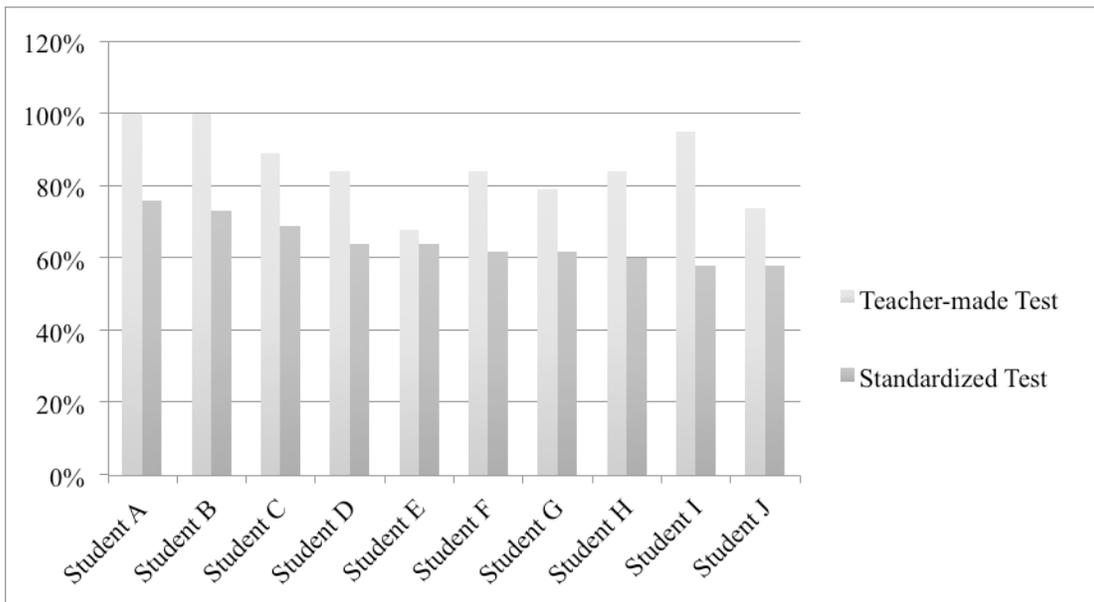


Figure 2
Comparison of the Ten Highest-Scoring Students' Results
on Two Summative Mathematics Assessments

Purpose



This research will find weaknesses in teacher-created tests and design a plan to help strengthen teachers' abilities in writing assessments so that such tests can be used in conjunction with standardized tests. This researcher sees the importance of aligning classroom tests with standardized tests so that educators are able to ensure the highest quality instruction and the most precise student placement possible.

By increasing teacher knowledge of creating classroom tests, all of the extensions of assessment will be affected. Such training can facilitate learning, better inform instructional practices, provide more accurate reports of student learning, and better place students in appropriate programs and levels.

Rationale

Teacher-made tests are an integral part of student learning, placement, and qualification. However, one significant problem in such evaluations is that the tests themselves are "often severely flawed or misused" (Burke, 2009, p.96). In 2009, Karen Burke cited the importance of training teachers in creating assessments, but had also noted the absence of this topic in teacher pre-service and in-service training. In her research, Burke found certain repetitive flaws in teacher-made tests including oversimplification of topics. She stated that many teacher-created questions focused on mere recall of facts and did not measure higher-level skills and processes. In 2011, Hussain Alkharusi reaffirmed these same issues. He claimed "findings from past and recent studies of classroom assessment have consistently expressed a concern about the adequacy of teachers' assessment skills" (p.40). He found that many teachers "did not have an adequate understanding of basic testing concepts such as item difficulty and

reliability” (Alkharusi, 2011, p.40). His study concluded that there is a need for closer attention to the topic of assessment and writing tests in teacher training.

Fray & Schmitt (2010) agree that even after research was conducted in the 1980s that highlighted issues with teacher-made assessments, teacher preparation programs did not improve. Their research states: “most college programs and state certification guidelines continued to have no explicit requirement that teachers were even trained in assessment” (Fray & Schmitt, 2010, p.115).

The diversity in bilingual schools and schools with high populations of second language learners create a need for further specialized training of teachers in regards to assessment. In 2007, TESOL (Teaching English to Speakers of Other Languages) created different characteristics to describe each of the TESOL standards. Standard 3: Curriculum, instruction, and assessment is characterized by the fact that “Teachers value, respect, and promote: D1- the diversity and individuality of students, [and] D2- the multiple ways in which students can demonstrate what they have learned” (Yturriago & Gil-Garcia, 2010, p.7). As the TESOL dispositions show, teachers are expected to respect student diversity and this respect and awareness need be reflected in teacher-created tests. The dispositions also show the importance of creating assessments that give all students the same opportunity to show what they have understood.

In an educational system where standardized tests are becoming increasingly common and important, why put such an emphasis on the tests that teachers create themselves? “Teachers place the highest information value on the tests they have constructed themselves and classroom assessment is perhaps the single most common

teacher professional activity” (Frey & Schmitt, 2010, p.108). Educational researchers also recommend that teachers use self-made tests in their classrooms, as teachers are the experts on the curriculum and their students (Frey & Schmitt, 2010). Although teachers create and use their own assessments in their classrooms, “they still rely on tests or items written by others about half the time” (Frey & Schmitt, 2010, p.115). This research intends to prove the necessity of training teachers in assessment so that they are better able to create quality tests aimed at a bilingual population, which also align to standardized assessments.

Definition of Terms

AIMSweb- an assessment used as a response to intervention in both literacy and math. It is “a complete web-based solution for universal screening, progress monitoring, and data management for Grades K-12” (NCS Pearson, Inc., 2012)

Adequate yearly progress (AYP)- declared by the state as a sufficient level of improvement made by schools and states in order to meet the regulations required by the No Child Left Behind Act. (Glossary of educational terms, n.d.)

Benchmark- a specific level designated as appropriate or “at grade level” student achievement. The benchmark details vary depending on subject and grade level. (Glossary of educational terms, n.d.)

Bilingual Education- a school program designed to increase content knowledge and language arts skills in both the child’s native language and a second language (Glossary of educational terms, n.d.)

Bloom’s Taxonomy- a hierarchy of levels of mental processes created by Benjamin

- Bloom. In classrooms, this taxonomy is used to help students move to higher levels of thinking, from basic recall to application and evaluation
- Charter school- a school funded by the city, but with exemption from many of the regulations that govern the city school district. They can be opened privately and run by individuals, teachers, parents, and other organizations. (Glossary of educational terms, n.d.)
- Common Core Standards- learning standards created to “provide a consistent, clear understanding of what students are expected to learn, so teachers and parents know what they need to do to help them” (Common Core State Standards Initiative, 2012)
- Dual Language- a strong form of bilingual education whose goal is to promote the attainment of two languages (Baker, 2011)
- English Language Learner (ELL)- any student who does not meet proficiency standards in English and who therefore qualifies as a second language learner and receives extra support. (Glossary of educational terms, n.d.)
- Formative Assessment- ongoing content assessment used to inform instructional practices and to subsequently increase student understanding of the current material (Glossary of educational terms, n.d.)
- Standardized test- “A test that is in the same format for all who take it. It often relies on multiple-choice questions and the testing conditions—including instructions, time limits, and scoring rubrics—are the same for all students”. Some accommodations are allowed for students that qualify. (Glossary of educational terms, n.d.)

Summative Assessment- used to assess students at the end of a topic or unit to see if they have met the objectives and standards. These scores are reported to families, administration, and communities.

TESOL- Teaching English to Speakers of Other Languages: teaching English as a second language

Tier II Intervention- an intervention level that requires progress monitoring of at-risk students used to create and evaluate interventions so that students can reach benchmark (NCS Pearson, Inc., 2012)

Tier III Intervention- an intervention level that requires intensive progress monitoring for the lowest performing students (NCS Pearson, Inc., 2012)

Summary

As is indicated by previous research and supported by the surveys and data collected for this current research, teachers are in need of further training in creating assessments to use in their classrooms. Teacher-made tests and standardized tests can only strive for a common goal if they yield comparable results, and it is through teacher preparation, training, and professional development that teachers can learn to better align their assessments to standardized tests in order to produce more similar scores.

There are multiple areas in which training and resources can help support and enhance teachers' test creation: writing clear directions, writing multiple-choice questions, writing culturally sensitive assessments, and using appropriate language. To address these topics, one must consider how each plays an important role in the outcomes of

assessments and contemplate the possible issues that lack of training in these areas could create.

Chapter 2

Literature Review

Assessment is a facet of education that, when implemented correctly, supports and helps further improve instruction, educational programs, and student attainment.

Assessments are used for various specific purposes, some of which extend far beyond the four walls of the classroom. According to Paul E. Newton (2007), different assessments and test results can be used to support learning and inform teacher instruction, place students in appropriate intervention levels, summarize student learning, serve as a predictor for future individual success, assess programs and schools, and serve as a form of “accountability to the public” (p.23). When teachers administer assessments, they must have specific purposes in mind for the results gleaned in order to make the assessments meaningful.

Some common self-reported uses for assessments by teachers were: re-teaching for better understanding of a topic, correcting misconceptions using the test questions as examples, and creating new but similar problems to ensure that the goals were met. Since teachers are provided with data that shows which students have mastered content and which have misconceptions, many reported having grouped students by their scores in order to better meet their individual needs (Frohbieter, Greenwald, Stecher, Schwartz, & National Center for Research, 2011).

Although assessments can be both formal and informal, this research will focus primarily on the former. Informal assessments, though important, are much more difficult to regulate as they include teacher observations, conferences, and oral questioning that take place during normal classroom instructional time (or center work). This researcher acknowledges the significance of informal assessments in classrooms in their ability to help a teacher immediately adjust his or her groups, focus, or method of instruction, but also understands that such observations cannot be used to report to administration, school boards, or families without other supporting data (Bailey, Little, Rigney, Thaler, Weideman, & Yorkovich, 2010). More structured or formal assessments tend to be those that are separate from normal instruction and work time and which also produce informative data.

Formative assessments tend to be informal and often include teacher observations, and conversations with students. Typically formative assessments are used to support the learning process and give the teacher feedback as to what topics have been covered completely and what topics need further support. The information provided by these assessments “show successes and failures for both teacher and students” (Bailey et al., 2010). These informal assessments must be focused in order to truly be an effective part of instruction and learning (Keeley, 2011). They “provide information teachers can use to change the way they teach” (Frohbieter et al., 2011, p.3). In this way, the results from such informal assessments become beneficial to the classroom because the information gained from these observations is used to tailor future lessons to the needs of whole and small groups (Bailey et al., 2010).

There are many ways to structure a formative assessment, which are useful in ensuring that there is variety in the classroom and that students are presented with multiple manners of displaying their knowledge. Some examples of such assessments are: homework, oral questioning, story retelling, self-assessment, practice quizzes, conferences, and daily classroom assignments (Bailey et al., 2010). Each of these different forms is intended for use in conjunction with others on a daily basis to check for student status and understanding. The goal of such assessments is to ensure that time is not being spent on classroom activities that are not needed or are not proving to be useful.

Unlike formative assessments, which foster learning, summative assessments summarize what learning has occurred and what level a student has achieved (Harlen, 2005). They provide judgments about the students and consequently about the programs and schools (Frohbieter et al., 2011). Summative assessments tend to be more formal, but can overlap with formative assessments. Some examples of summative assessment formats are: homework assignments, classroom activities, written reports, and tests. Considering this overlap, it is important to note that the primary difference in these two types of assessment is not the format, but rather their purpose (Bailey et al., 2010). The results of summative assessments are used for reporting, certification, student placement, and “monitoring the performance of teachers and schools” (Harlen, 2005, p.2). The results of these assessments let the teacher know if students have met state and classroom standards and if they are ready to move on to the next topic or level.

Assessments have become an integral part of education and the results serve as important evidence for teachers, administrators, parents, community members, state and

local governments, and students themselves. “Testing can aid in decisions about grouping students...diagnose what individual pupils know...help the teacher determine the pace of classroom activities...[and can be used to] to share information with boards of education, parents, and the general public” (Rudner & Schafer, 2002, p.2). Considering the diverse uses of assessment, it is essential that these assessments be well made, reliable, and validly interpreted so that the scores are credible.

For years, teachers have created their own tests to measure student understanding of language and content. “Teacher-made tests are written or oral assessments that are not commercially produced or standardized...[they are] design[ed] specifically for [each teacher’s] students” (Burke, 2009, p. 96). Such tests are not typically reviewed by anyone but the teacher himself or by the content coach for that particular subject. This means that the teacher is responsible for determining the content, style, and length of the test as well as its grading and score evaluation. In terms of format, such tests might include matching, labeling, fill in the blank, true or false questions, multiple-choice, short answer or essay questions. Due to the increasing complexity of assessments, “teachers spend as much as a third to a half of their professional time in classroom assessment activities” (Alkharusi, 2011, p.46). This time might include creating tests, grading, and using scores for any of the aforementioned purposes. Research states, “often, teachers are the best evaluators of their diverse students because they are most familiar with students’ idiosyncratic backgrounds and learning progress across time” (Gonzalez, 2012, p.293). Teacher-made tests allow for tailoring to specific classrooms and student needs as

long as the classroom teacher has been trained to create appropriate and authentic assessments that include these student differences.

Standardized tests differ from teacher-made tests in that they “use uniform procedures for administration and scoring” (Rudner & Schafer, 2002, p.21). With such tests, there is an attempt to “make the observation, materials, administration, and the scoring as uniform as possible” (Farah, 2013, p.209). This is done so that the same test can be given to students across the nation to produce comparable results. The uniformity of the process ensures that the scores themselves are being compared, without the interference of variables such as directions or grading style. Due to the ability to compare these scores across a large sample, standardized scores tend to be those that are reported to the public and the media (Burke, 2009).

The charter school used as a basis for this research reports standardized scores to the city and state in order to prove that students are making adequate yearly progress and are reaching state standards. These scores are generally accepted as more reliable because they are derived from large-scale tests that have been reviewed and piloted. These tests tend to be used for teacher accountability and student placement, and although the teachers are aware of the content required by the state, they are unaware of the questions that will appear on the test until it is administered (Rudner & Schafer, 2002). Preparations for standardized tests, therefore, differ from preparations made for teacher-created tests.

AIMSweb is an example of a standardized test used in schools across the nation. It is a “three-tier assessment, intervention, and progress monitoring system” (AIMSweb, 2005, p.13). This means that the tests are intended to separate students into three groups:

Benchmark (Tier I), Strategic (Tier II), and Intensive (Tier III). A student's classification determines the type of content-specific support he or she will receive (or not receive) to help him or her reach the benchmarks set for that subject and grade level. The Benchmark students are tested three times per year; the Strategic students are monitored throughout the year to evaluate intervention strategies. The final, Intensive tier includes progress monitoring alongside focused services. These students are continually observed, their intervention services are the most frequent, and their intervention groups include the smallest teacher to pupil ratio (AIMSweb, 2005).

The charter school used for this research uses AIMSweb to support and enhance its mathematics programs at the primary level. All students are tested at the beginning of the year, are placed into the appropriate intervention group and are monitored accordingly. The classroom teacher then uses these scores to create individualized intervention plans and set goals for the students. As these goals are reached, they are modified, and the tutoring group's focus increases in difficulty or changes topic. These AIMSweb test results are used for placement, monitoring, informing individuals' mastery of content knowledge, and to inform teacher instruction in both whole group and tutoring groups. In addition, these scores can be used as proof of students' abilities for things such as special education referrals or gifted/talented individualized instruction.

Teacher-made tests and standardized tests are both used to measure student knowledge throughout the school year. Since each is to be based on the state requirements (for this charter school, the Common Core Standards), the tests, ideally, should be comparable. The scores ought to show the same students as scoring at the

Benchmark level, and the same students should score in the below-average Strategic and Intensive intervention levels. Unfortunately, due to nuances in teacher-made tests, this is not always the case. The incompatibility of teacher-made and standardized tests creates conflict in student placement, retention, accountability, and instructional practice. If two tests on the same content produce inconsistent achievement results, confusion arises regarding scores are more accurate and which should be used officially.

It is the hope of universities and administrators that teachers are trained well enough in their field to create reliable tests that can be used alongside standardized tests, but there is often a lack of training. Many suggestions have been made concerning what teachers need to know in regards to assessments, including aligning assessments to research-based practices and state standards, but little is available to help teachers learn these skills (Rudner & Schafer, 2002; Yturriago & Gil-Garcia, 2010). Of 213 teachers surveyed by Hussain Alkharusi in his 2011 research, 133 claimed to have had no in-service assessment training and only 80 cited having received training. The teachers who had received in-service assessment training reported being more confident in the following assessment skills: “analyzing test items, communicating assessment results, writing test items, using performance assessment, and grading” (Alkharusi, 2011, p.46).

To help remedy this lack of training and resources, Rudner and Schafer (2002) compiled best research and practices in their book, *What Teachers Need to Know About Assessment*, in which they outline “fundamental concepts common to all assessments, essential classroom assessment concepts, [and] useful concepts and issues pertaining to district, state, and national assessment” (p.i). Their book provides teachers with basic

facts about assessments and gives advice on how to write multiple-choice test items and score assessments using rubrics (Rudner & Schafer, 2002).

Though assessment serves as an important tool in today's educational system, it is not without its shortcomings. Many assessments measure at the lowest level of Bloom's Taxonomy: knowledge; this level places an emphasis on concrete memorization and facts instead of higher-level thinking (Bailey et al., 2010). Directions and test questions themselves are often unclear, leading to unnecessary test-taker confusion (Jacobs, 2004). Another limitation of tests is that they are sometimes found to contain "biased language", or language that is favorable to certain cultural or socioeconomic groups (Saenz & Huer, 2003). Many issues including bias, unclear directions, and cultural sensitivity all address the same core concern: fairness. In most standardized tests, analyses are performed to identify any unfair tests or questions. There is an attempt to weed out items or qualities that may make the test easier or more difficult for specific groups of individuals (Camilli, 2013). Despite the attempts to create fair assessments, these same analyses are not performed with teacher-made tests, which leads to increased concern about the fairness of classroom assessments.

The faults of assessments can lead to scores that inaccurately represent student understandings of content and language. A more specific factor that can inadvertently affect student demonstration of achievement and create unfair testing is teacher bias. Teachers' evaluations can affect "students' motivation, self-confidence, and longer term school outcomes" (Van Ewijk, 2011, p.1045). The important role that teacher evaluation plays in students' educational careers highlights how detrimental bias can be. The burden

of such bias usually falls upon ethnic minority groups. “Research shows that ethnic minority students perform poorer in school when they are taught by teachers belonging to the ethnic majority” (Van Ewijk, 2011, p.1045). If the student belongs to this minority group, he or she may have a lower chance of performing at benchmark due to the types of questions asked or the language used. Questions that would be unfair to certain cultural groups are referred to “culturally insensitive” items. Such items may use stories or terms that are familiar to a certain culture and could be misinterpreted or cause confusion for a student who is of a differing culture (Camili, 2013). Group classifications that are typically considered in test fairness other than race/ethnicity are “social class, language, [and] urbanicity” (Camili, 2013, p.107).

There exists concern for fair assessment of student groups or individuals whose health, personalities, disabilities, etc. influence the way in which they demonstrate their understanding on assessments. The important focus in assessment of diverse groups is equity: giving students an equal opportunity to display their knowledge through individual accommodations or considerations used to level the playing field. The purpose of making accommodations is to ensure that the scores are comparable and that they do not reflect a student’s disability, culture, income level, or language dominance (Camili, 2013).

As classroom assessments have the advantage of serving a specific student population to which the teacher has daily access, training is required to help teachers incorporate and consider the multiculturalism of their classrooms when creating tests. “Contemporary research on the assessment of ELLs has revealed that the most important

tool for assessment is the evaluator's personality" (Gonzalez, 2012, p.293). Therefore, before teachers can be trained in the specifics of assessment, they must know how their own biases or understandings of other cultures play a role in their instruction and assessment. For teachers to be culturally sensitive and to act in such a way that reflects their understanding, "their education should foster the development of the necessary knowledge, skills and sensitivity" (Spinthourakis, Karatzia-Stavlioti, & Roussakis, 2009, p.267). This knowledge equips teachers with the tools to acknowledge, accept, and even incorporate these student differences in their classrooms. The understanding of student diversity is one of many factors essential to "help instructors write better tests—better in that they more closely assess instructional objectives and assess them more accurately" (Jacobs, 2004, p.1). This research is aimed at helping teachers approach this accuracy in assessment.

The multiculturalism of bilingual classrooms in particular creates a need for teachers to familiarize themselves with their students' language, culture and socio-economic status in addition to their learning style, strengths, and weaknesses. A teacher who is aware of the multicultural classroom and all its challenges recognizes that factors such as discrimination, poverty, home language, and exposure to the dominant language can affect student performance and achievement. In order to help prevent these factors from influencing test scores negatively, teachers must, in part, "link assessment to instruction and represent their [the students'] first language (L1) and native culture" (Gonzalez, 2012, p.291). In addition to being aware of and using student differences,

teachers also need training in intercultural sensitivity to help eliminate personal bias and to make teachers reflect on how their own beliefs affect their students (Gonzalez, 2012).

In order to be well-educated professionals who are able to write and administer assessments, teachers must know how to choose the format that will best fit the content and standard they intend to assess. After determining the most appropriate format, the educators must also be equipped with the tools to write each of the following types of questions: multiple-choice, fill in the blank, true or false, and essay questions. It is essential that teachers are able to create alternative assessments that allow students to show their knowledge in a manner separate from a paper and pen test. To ensure that teachers are able to vary assessments to accommodate for children who better perform on more hands-on tasks, they must also be trained in evaluating student understanding through authentic performance assessments (Rudner & Schafer, 2002). In addition to all of the aforementioned assessment formats, teachers can also use “classroom observations, interviews, rating scales, and rubrics” that represent the diversity of student experiences and strengths (Gonzalez, 2012, p.294).

Considering the vast amount of knowledge necessary to create and evaluate student assessments, it is important to consider whether teachers are adequately prepared for this immense task. According to a study conducted in Greece of pre-service teachers in regards to how they felt about their abilities in assessment after their pre-service training, “more than two-thirds of the students [pre-service teachers] consider their levels unsatisfactory” (Spinthourakis et al., 2009, p.274). A similar study conducted in Canada also revealed “many beginning teachers feel unprepared to assess students’ performance”

(Poeth, 2013, p.634). These findings are relevant in that they demonstrate a need for further training to fill the gaps in knowledge and understanding that these teachers have in regards to assessments. Teachers must feel confident in their abilities to create well-designed, organized, and concise test items that measure specific standards.

Though it is essential to create test questions that are both clear and correctly formatted, it is also important that teachers create prompts that elicit higher-level responses. Since it is a common error of teachers to design tests that are composed primarily of low-level recall questions, they must receive training on how to use Bloom's Taxonomy verbs to encourage students to demonstrate critical thinking. The inherent purpose of the taxonomy is to encourage questioning that reaches a variety of levels of thinking. Once teachers fully grasp the levels of the taxonomy, they are able to incorporate the levels into their test writing, promoting a deeper demonstration of student understanding (American Federation of, T., 2010, p.6).

The final consideration made regarding teacher preparation concerns the implementation of the Common Core State Standards. These standards, currently available for mathematics and literacy, reflect a dramatic increase in expectations for students in order to prepare them for college and careers. The standards aim to encourage higher-level thinking and problem solving alongside an in-depth understanding of content. The demanding standards necessitate further teacher training. One research work compiled the ideas of experts who helped create a list of recommendations for teacher professional development as it pertains to the Common Core State Standards. Of the suggestions made, the one that best aligns with this research is "Recommendation 3:

Design Common Core State Standards Professional Development Based on Features That Support Teacher Learning” (Marrongelle, Sztajn & Smith, 2013, p.205). This recommendation, as well as others mentioned in this article, concludes: “intensive, ongoing, and connected to practice...[and] focus on student learning and address the teaching of specific content” (Mattongelle et al., 2013, p.207). These high hopes for teacher training would help enable teachers to better present and assess student achievement in regards to the new standards.

One study from Bowling Green State University surveyed both elementary and middle school teachers to determine what they perceived as their weaknesses when following the Common Core Standards. The results of this study were used to create applicable training material (Bostic & Matney, 2013). The study found that elementary teachers cited a need for support in content development, specifically in Operations & Algebraic Thinking and Numbers & Operations-Fractions. This finding is applicable to this current research in that the professional development need aligns directly with the students’ state test performance. The low student performance and lack of teacher confidence in preparing students for test questions addressing CCSS reflect a need for teachers to align their own formative and summative assessments to future state tests. Better teacher-made tests reaching to these high standards will help ensure that teachers can predict their students’ performance on state-mandated standardized tests.

It is essential that teachers have a firm grasp on what the Common Core Standards ask of students in order to properly prepare students for current assessments, future grade levels, and, eventually, college. These standards have very clearly laid out what students

need to know at each level of mathematics and literacy. To ensure that each student has an opportunity to meet all of the Common Core Standards, teachers need to know which students are currently performing above, at, and below grade level so as to provide appropriate supports. In order to obtain such student performance data, classroom assessments are required at the beginning of the year, and throughout the entire school year as progress monitoring.

The state assessments will need to be redesigned to meet these new standards, as will classroom assessments. “The Common Core represents an opportunity to totally redesign assessment systems, using the standards and the college-ready goal as the guides” (Phillips & Wong, 2010, p.39) which will help keep track of and develop college and career-ready students. Teachers must be provided with examples of what Common Core aligned classroom assessments would look like so that they are working alongside the state goals and measurement systems towards a common learning goal (Phillips & Wong, 2010).

As has been established throughout this research, assessment is becoming an increasingly important aspect of education in today’s society. In order to facilitate the increase in assessment and its uses, teacher training programs must prepare teachers to align standards and best research with their assessments, to use assessments to improve instruction, and to utilize and report student scores appropriately (Poth, 2013, p.636). Training of teachers must prepare instructors to “assess critical thinking skills and learning potential in the various verbal and nonverbal forms in which they are represented across cultures and languages” (Gonzalez, 2012, p.294). With properly

equipped teachers, assessment scores will produce more accurate, and therefore valuable, results.

In order to ensure that educators are best measuring student knowledge, teachers must be well trained in all aspects of assessment: creation, grading, and interpretation of results. This research focuses on the creation aspect and will provide teachers with a test-writing manual to help guide them towards best practices in a user-friendly format. In the interest of serving the needs and wants of the survey group, the manual will focus on creating culturally sensitive tests appropriate to a dual language school with a diverse population.

Chapter 3

Application

The product of this research has two components; it includes a professional development program and a teacher handbook, both of which will support teacher creation of assessments for a bilingual population. Both the professional development and the manual were created in response to a teacher survey conducted at a bilingual charter school, which brought to this researcher's attention various needs that the teachers had in regards to assessment in their classrooms.

The professional development component of this product has been created to help train teachers in test writing through an initial workshop and two follow-up meetings. The first session will be whole group training, conducted with the entire teaching staff. The presentation pace and basic information will be provided by a PowerPoint presentation. The presenter will facilitate the discussions, questions, and activities.

The estimated time for Session I is 1 hour and 35 minutes. The presentation will begin with the teachers' self survey about their knowledge regarding creation of assessment for bilingual students. The self-assessment includes questions about how comfortable and prepared teachers feel to write high-quality, standards-aligned test items. There are also questions that require teachers to choose their personal challenges and beliefs about what can affect student test performance. The last two items on the survey ask that teachers describe what they know to be necessary for high-quality multiple-choice and essay test items. The teachers will be given approximately five minutes to complete the survey, which will then be collected and saved for use at Session 2.

To ensure that teachers understand the need for this professional development topic, the presenter will provide a brief synopsis of recent literature demonstrating a need for more preparation in test writing for the purposes of improving all facets of assessment. The group will discuss the extensive uses of assessment and the negative effects that poor assessments can have on students, teachers, families, and communities.

Following the quick literature review, the staff will be instructed on where to find the Common Core State Standards with a click-through of the links listed on the presentation. The teachers will also be shown how to find the verbs in the standards and will be encouraged to focus on the difficulty level established by such verbs. This will prove important when teachers write test items to address specific standards, as the complexity of the test item should match that of the standard.

In the next section, multiple-choice items will be discussed in terms of their parts, their uses and advantages, and the research-based guidelines for creation. Definitions are

provided for the “stem”, “distractors”, and “alternatives”. Before the teachers are shown how to write multiple-choice items, each pair will be given an envelope filled with cards. Each card will either contain a “Do” or a “Do not” in regards to writing such test items. Teachers will be asked to sort these cards into one of the aforementioned groups. Upon completion, the presenter will review the list of Dos and Do Not’s for writing high quality multiple-choice test items, using information gleaned from observations during the teacher sort to spur discussion about specific guidelines. For example, if the presenter finds teachers debating whether or not multiple-choice items should each have the same number of distractors, he or she will bring this topic to the group for discussion.

Teachers will then be provided with the most important pieces to be included in essay prompts: the role of the student, the audience they are addressing, the format of the essay, and the topic: whether it be content or literary. The presenter will give examples of each of the above parts of a prompt so that teachers know exactly how a complete prompt is constructed. In addition to discussing the pieces of an essay prompt, the presenter will discuss how to keep prompts culturally sensitive: through creating situations that all students are familiar with, or through providing choice in literary contexts.

Although previously mentioned, there will be a section of the PD dedicated specifically to culturally sensitive test items. Teachers will be instructed to focus on the standards that they are trying to assess and to ensure that there is no bias or cultural insensitivity in the test items that would prevent knowledgeable students from performing well. The teachers will also be trained on specific ways to write culturally sensitive test items.

The teachers will be given pairs of example multiple-choice items and essay prompts to review using the criteria previously presented. They will be given 15 minutes to find, in each pair, the high quality and low quality item, providing reasoning behind their decisions. After the teachers have made their decisions, the presenter will pass around copies of the Teacher Manual and will discuss the examples and their good and poor qualities, facilitating any questions or concerns that may arise.

At this point in the PD, the teachers will demonstrate their new knowledge. They will be provided with test writing worksheets to practice writing a multiple-choice test item and an essay prompt for their particular grade level and content area. The teachers will be encouraged to use their Teacher Manuals as a guide.

As a closing to Session 1, the teachers will share their test items in small groups and then choose what they think is the strongest multiple-choice item and the strongest essay prompt from their small group to share with the whole group. The presenter will comment on specific qualities of the test items that are shared out by the group. He or she will also remind teachers how to use the Teacher Manual on their own so that they will be able to prepare a collection of sample test items for Session 2.

Session 2: Creation Implementation is to be conducted at grade level meetings with the content and literacy coaches. The coaches will return the self-surveys from Session 1 to the teachers and will ask teachers to reflect on their previous knowledge of assessments for bilingual students. The teachers will be given another five minutes and a different colored pen to add or retract any information on their survey. The teachers and

coaches will talk about any changes that are shown on the survey and discuss how any new knowledge has affected the teachers' assessment in the classrooms.

The teachers will bring sample test items that they have created to share with the coaches. As a team, they will review some of the items, using the Teacher Manual checklists to ensure that they are reaching for the highest possible standards. Revisions will be made as needed and teachers will be asked to use these and other test items in future assessment and to save the student data from said assessments for use in Session 3.

Session 3 is to be conducted at grade level meetings, as well. At this time, the teachers must bring with them a self-made assessment along with the student scores from that assessment. The coaches will help the teachers find the most missed test items and the items that were most often answered correctly. As a team, the teachers and coaches will look at which students answered correctly or incorrectly, and will consider their affiliation with a specific group: gender, culture, language, etc. The goal in this session is to see if there is any correlation between specific test items and student profiles. From there, the teams can decide which of the test items may have contained bias, been unclear, or did not measure the designated standard. They can then also decide which test items produced accurate results and were of high quality. Any seemingly poor test items will be reviewed in detail to help decipher what made that test item poor, and will be either discarded or improved.

The manual will serve as an artifact and reference material for teachers to use post professional development. They will access the handbook as an online PDF document that they can print should they want a paper copy. The manual will consist of six umbrella

topics: Common Core State Standards, culturally sensitive assessments, creating multiple-choice assessments, writing essay prompts, flexibility in administration, and checking-in after the test. The aim of each section of the manual is to provide teachers with a brief overview of the varying facets of assessment. Each section includes a connection to the Common Core State Standards and examples of both a high and low quality test item along with explanations of their good and poor qualities. Where applicable, the manual provides checklists for teachers to use as a quick, yet complete, reference against which they can examine and scrutinize their own classroom tests with the goal of continuous improvement.

The Common Core State Standards section addresses four questions: 1) What are the Common Core State Standards? 2) Where can I find the standards? 3) How do these standards affect my assessment? and 4) How can I align my assessments to the core? The standards are included as the first section because they are the current driving force in education. The CCSS are what drive instruction, learning, resource selection and assessment; therefore, it is essential that teachers are able to align all classroom activity to these standards. There is a checklist provided for teachers to use as they are creating assessments to help them locate the standards and create test items that specifically address both the content and level of complexity established by the standards. An example is provided to present one question that accurately aligns with a specific CCSS, and one that does not align.

The next section is included to help teachers prepare tests that are culturally sensitive to all of the students in their classrooms. The main focus here is to establish

what it means to be culturally sensitive and how this perspective transfers into writing appropriate directions and using unbiased language. To provide a quick reference, there is a list of “Do’s” and “Do Not’s” in terms of language use on assessments, a good and poor example of a culturally sensitive test item, and an insert to help align language use with the Common Core. The primary objective of this section is to ensure that assessments are testing what they intend to test. For example, a mathematics test item must be worded so that each item is surveying the student’s ability to perform a given mathematical task, not to judge the student based on his or her reading level, language, or culture.

Teachers surveyed for this research also requested guidance in creating high quality multiple-choice questions. Although such items are easy to grade and can cover a vast amount of information in a short testing period, they are difficult to construct so that they accurately represent a student’s knowledge and understanding of the content or language. This section of the manual stresses the importance of aligning each item to the Common Core State Standards. It also provides a detailed checklist for teachers to use both during and after test creation to confirm that each item is formatted in such a way that it is highly likely to produce reliable results, even in a bilingual setting. As with each of the previous sections, good and poor examples are included that address different multiple-choice guidelines. This section is somewhat unique in that it is based on research data that had previously discovered which characteristics helped build strong items, and which caused interference. So although these test items are among the most difficult to create, they have the most streamlined, empirically based instructions for proper creation.

The manual then segues into a section on writing essay prompts and aligning the prompts to the Common Core. It also includes a checklist of what to include and consider in an essay prompt alongside a good and poor example. This section addresses the complexities involved in an essay test item as proof of content understanding. This test item format requires reading and writing skills beyond those typically required in content, and consequently may not be an appropriate way to measure content understanding for all students in a diverse classroom. This issue leads directly into the following section on maintaining flexibility in test administration.

Despite all training and preparation to create high quality, compatible assessments for a bilingual population, the manual acknowledges that there will always be gaps. In such classrooms, we expect and respect the diversity, understanding that diversity itself can be manifested differently depending on the school, year, family, and student. Flexibility in test administration must come into play to fill these gaps. This flexibility is the grey area that often exists in education, because no matter how seemingly perfect or how research-based an assessment may be, it will never be ideal for all teachers or all students all of the time.

“Checking in after the test” is a section that encourages teachers to reflect on the work that they have done in order to continue to make improvements, if necessary. It also provides teachers with some hints as to red flags for post assessment that may have negatively impacted the goals of their assessment, and consequently, their students’ performance. This session provides a time for teachers to find what *did* work on their

assessments, and to celebrate these successes as a breakthrough into the complex world of student assessment.

The professional development sessions and teacher manual are based on research on the various aspects of assessment: multiple-choice creation, essay prompts, and culturally sensitive language. The check list for multiple-choice creation is a compilation of both old and new research that focuses on the organization, language, and various parts of multiple choice stems and alternatives, as well as the means by which such questions can prompt higher-order thinking (Burton, Sudweeks, Merrill & Wood, 1991, Jacobs, 2004 , Rudner & Schafer 2002, Salend, 2011, Torres, Lopes, Babo, & Azevedo, 2011).

Furthermore, the professional development series and manual recommend that all essay prompts contain four basic parts: role, audience, format, and topic. Research states that all parts of an essay prompt must be clear and complete, so as to leave the student with an exact idea of what is expected (Jacobs, 2004, McMillan, 2007, Reiner, Bothell, Sudweeks, & Wood, 2002, Salend, 2011, and *Writing and grading*, 2012).

The overall purpose of the product of this research is to train teachers in the most important aspects of test writing. The professional development will guide teachers through the significance of creating quality assessments as well as the means by which such assessments must be generated. The guided sessions focus on the basics of assessment, but always incorporate the importance of cultural diversity and cultural sensitivity. The manual is a product of the professional development that will serve as a reference for teachers long after the three sessions have been completed.

Chapter 4

Limitations

This research serves the purpose of outlining basic assessment writing guidelines for teachers who deal with a diverse or bilingual population. Although informative, the scope of this study was not all-inclusive. This researcher acknowledges various limitations, which include size of the survey population and depth of research. Twelve teachers from the same bilingual school were chosen to participate in the survey. These teachers, therefore, receive the same (or very similar) professional development through the school, and many share the same local alma maters. These factors lessen the variability in the assessment and training experiences of the teachers.

Another limitation of this study is the pool of data used, which was considered when analyzing teacher-made tests versus standardized assessments. The data were collected from one mathematics teacher-made test and one mathematics standardized assessment from one Grade 2 class in a bilingual charter school. Although all the data from the 50 Grade 2 students were analyzed, only the scores of the highest and lowest students were discussed in this work.

One goal of this research is to prepare teachers to create quality assessments based on the Common Core State Standards. This research did not consider states and schools that are not teaching to these standards. Also, since the Common Core does not include science and social studies standards, this researcher chose to include New York State Standards in some of the examples, which are not applicable to teachers outside the limits of the state.

Due to these various limitations of the research, the conclusions made from this data analysis cannot be generalized for other populations.

Recommendations

This researcher recommends that further research be conducted on actual results of culturally relevant assessments and culturally biased assessments to further convey the inequalities in such tests. Also recommended is an analysis of the types of language and test items that cause unnecessary confusion to culturally and linguistically diverse students. Such information would prove the dangers of specific types of language or organization of test items. These data would also help inform teachers and other test writers of the types of questions and language to avoid in order to create clear, fair tests.

In addition, this researcher recommends the analysis of current standardized state tests for cultural bias, content specific language, and properly structured test items so that these test writers can also reflect on and improve their assessment writing skills. If we are to hold teachers to the highest standards in assessment creation, we must also do so with the assessments that have the highest stakes for students.

A further recommendation of this research is that teachers familiarize themselves with their student bodies, the standardized tests these students are required to take, and the content and language skills required by the CCSS. Using all of this aforementioned knowledge, teachers will be better equipped to use this teacher manual and benefit from this professional development. Such understanding will help teachers to personalize and tailor assessments to each classroom and educational setting. Teachers have the advantage of being able to serve their students in a way that standardized tests never will.

Therefore, teachers must take advantage of this opportunity to customize and adapt assessments to their specific students each year, as well as to research-based test writing standards in order to glean the most accurate and useful data possible.

Conclusion

This research makes certain conclusions regarding teacher-made tests and standardized tests. One such conclusion is that there is great diversity in the way teachers are prepared for creating and administering assessments, which produces inconsistencies in the types of assessments to which our students are susceptible. If every teacher in every classroom has a different idea about what qualifies as a high quality assessment, then the ways in which students are being trained for high stakes standardized tests vary greatly. The lack of teacher training also leaves students vulnerable to ignorance and bias of the test writer.

All students, especially culturally and linguistically diverse students, are among those at the mercy of poorly constructed test items. Students cannot perform their best or demonstrate their true knowledge if test items are not clear, concise, and based on previously taught state standards. It is the responsibility of teachers to create and administer assessments of only the highest quality to avoid improper placement, improper and ineffective lesson and tutoring plans, and student frustration, among other negative effects of poor test items.

To help prevent tests from inaccurately measuring and reporting student knowledge, this researcher concludes that teachers require additional training in creating and aligning assessments, particularly for culturally diverse populations. To remain

sensitive to these bilingual/diverse populations, teachers also need training in culture: diversity, understanding, and differences between the mainstream culture and the culture of their students. These types of understandings will enable teachers to tailor test items to their student body, as well as give them the ability to administer tests in such a way that allows flexibility when unforeseen hurdles arise.

References

- AIMSweb. (2005). *Special Education Technology Practice*, 7(4), 13.
- Alkharusi, H. (2011). Teachers' classroom assessment skills: Influence of gender, subject area, grade level, teaching experience and in-service assessment training. *Journal Of Turkish Science Education (TUSED)*, 8(2), 39-48.
- American Federation of, T. (2010). Quality classroom assessment techniques. classroom tips. *American Federation Of Teachers*.
- Bailey, J., Little, C., Rigney, R., Thaler, A., Weiderman, K., & Yorkovich, B. (2010). Assessment 101: Assessment made easy for first-year teachers. (Master's thesis). Retrieved from EBSCO.
- Baker, C. (2011). Foundations of bilingual education and bilingualism. Buffalo, NY: McNaughton & Gunn Ltd.
- Bostic, J., & Matney, G. (2013). Overcoming a common storm: Designing professional development for teachers implementing the common core. *Ohio Journal Of School Mathematics*, (67), 12-19.
- Burke, K. (2009). How to assess authentic learning. Thousand Oaks, CA: Corwin.
- Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1991). *How to prepare better multiple-choice test items: guidelines for university faculty*. Informally published manuscript, University Testing Services and The Department of Instructional Science, Brigham Young University, Provo, UT.
- Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research And Evaluation*, 19(2-3), 104-120.

- Common Core State Standards Initiative. (2012). *Implementing the common core state standards*. Retrieved from <http://www.corestandards.org/>
- Farah, Y. (2013). Through another's eyes: Modification or accommodation in standardized testing? *Gifted Child Today*, 36(3), 209-212.
- Frey, B. B., & Schmitt, V. L. (2010). Teachers' classroom assessment practices. *Middle Grades Research Journal*, 5(3), 107-117.
- Frohbieter, G., Greenwald, E., Stecher, B., Schwartz, H., & National Center for Research on Evaluation, S. (2011). Knowing and doing: What teachers learn from formative assessment and how they use the information. CRESST Report 802. *National Center For Research On Evaluation, Standards, And Student Testing (CRESST)*.
- Gonzalez, V. (2012). Assessment of bilingual/multilingual pre-k-grade 12 students: A critical discussion of past, present, and future issues. *Theory Into Practice*, 51(4), 290-296.
- Halsdorfer, J. (2012, August 14). *Eugenio maria de hostos charter school:annual report 2011-2012*. Retrieved from <http://www.emhcharter.org/index.html>
- Harlen, W. (2005). Teachers' summative practices and assessment for learning: Tensions and synergies. *Curriculum Journal*, 16(2), 207-223. doi:10.1080/09585170500136093
- Jacobs, L. (2004). How to write better tests: A handbook for improving test construction skills. Retrieved from http://www.indiana.edu/~best/write_better_tests.shtml
- Keeley, P. (2011). Formative assessment probes: With a purpose. *Science And Children*, 48(9), 22-25.

- Marrongelle, K., Sztajn, P., & Smith, M. (2013). Scaling up professional development in an era of common state standards. *Journal Of Teacher Education, 64*(3), 202-211.
doi:10.1177/0022487112473838
- McMillan, J. H. (2007). *Classroom Assessment: Principles and Practice for Effective Standards Based Instruction (4th ed.)*. New York: Pearson.
- Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment In Education: Principles, Policy & Practice, 14*(2), 149-170.
- NCS Pearson, Inc. (2012). *AimswEB*. Retrieved from aimsweb.com
- (n.d.) Glossary of educational terms. *School wise press*. Retrieved July 22, 2013, from <http://www.schoolwisepress.com/smart/dict/dict3.html>
- Phillips, V., & Wong, C. (2010). Tying together the common core of standards, instruction, and assessments. *Phi Delta Kappan, 91*(5), 37-42.
- Poth, C. (2013). What assessment knowledge and skills do initial teacher education programs address? A western canadian perspective. *Alberta Journal Of Educational Research, 58*(4), 634-656.
- Reiner, C. M., Bothell, T. W., Sudweeks, R. R., & Wood, B. (2002). *Preparing effective essay questions: A self-directed workbook for educators*. New Forums Press.
Retrieved from <http://testing.byu.edu/info/handbooks/WritingEffectiveEssayQuestions.pdf>
- Rudner, L., & Schafer, W. (2002). What teachers need to know about assessment. Washington, D.C.: National Education Association of the Unites States.
- Saenz, T., & Huer, M. (2003). Testing strategies involving least biased language

- assessment of bilingual children. *Communication Disorders Quarterly*, 24(4), 184-193.
- Salend, S. J. (2011). Creating student-friendly tests. *Educational Leadership*, 69(3), 53-58. Retrieved from <http://www.ascd.org/publications/educational-leadership/nov11/vol69/num03/abstract.aspx>
- Spinthourakis, J. A., Karatzia-Stavlioti, E. E., & Roussakis, Y. Y. (2009). Pre-service teacher intercultural sensitivity assessment as a basis for addressing multiculturalism. *Intercultural Education*, 20(3), 267-276.
- Torres, C., Lopes, A., Babo, L., & Azevedo, J. (2011). Improving Multiple-Choice Questions. *US-China Education Review B*, (6b), 1-11.
- Van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics Of Education Review*, 30(5), 1045-1058.
- Writing and grading essay questions*. (2012). Informally published manuscript, Center for Faculty Excellence, University of North Carolina, Raleigh, NC, Retrieved from <http://cfe.unc.edu/pdfs/FYC7.pdf>
- Yturriago, J., & Gil-Garcia, A. (2010). Preparing bilingual teacher candidates: A linguistic conundrum in a changing political landscape. Retrieved from EBSCO.

_____ Writing true or false questions

_____ Writing culturally sensitive assessments

5. Which of the following do you think can affect student performance on a test? **Check all that apply.**

_____ The wording of directions and questions

_____ Teacher bias

_____ Culture of the student

_____ Test format (the types of questions being used)