Journal of Education and Learning; Vol. 7, No. 1; 2018
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

Identifying New Jersey Teachers' Assessment Literacy as Precondition for Implementing Student Growth Objectives

Victoria Prizovskaya[1]

[1] Elizabeth Public Schools District, Elizabeth, New Jersey, USA

Correspondence: Victoria Prizovskaya, 411 Cynthia Court, Princeton, NJ, 08540, USA. E-mail: prizvi1@verizon.net

Received: October 4, 2017 Accepted: October 24, 2017 Online Published: October 30, 2017

doi:10.5539/jel.v7n1p184 URL: http://doi.org/10.5539/jel.v7n1p184

Abstract

The Student Growth Objectives are assessments created locally or by commercial educational organizations. The students' scores from the Student Growth Objectives are included in teacher summative evaluation as one of the measures of teacher's effectiveness. The high amplitude of the requirements in teacher evaluation raised a concern of whether New Jersey public school teachers were competent in assessment theory to effectively utilize the state mandated tests. The purpose of this quantitative study was to identify New Jersey teachers' competence in student educational assessments. The researcher measured teachers' assessment literacy level between different groups based on subject taught, years of experience, school assignment and educational degree attained. The data collection occurred via e-mail. Seven hundred ninety eight teachers received an Assessment Literacy Inventory survey developed by Mertler and Campbell. Eighty-two teachers fully completed the survey (N=82). The inferential analysis included an independent-sample t test, One-Way Analyses of Variances test, a post hoc, Tukey test and Welch and Brown-Forsythe tests. The results of this study indicated teachers' overall scores of 51% on entire instrument. The highest overall score of 61% was for Standard 1, Choosing Appropriate Assessment Methods. The lowest overall score of 39% was for Standard 2, Developing Appropriate Assessment Methods. The conclusion of this study was that New Jersey teachers demonstrated a low level of competence in student educational assessments. In general, the teacher assessment literacy did not improve during the last two decades.

Keywords: assessment literacy, student assessment, teacher evaluation

1. Introduction

In attempt to reinforce educators' accountability for students learning the federal and state educational agancies mandate school districts to include the student achievement data in teacher evalautation. The number of school districts that in some way incorporate the students' outcomes from the assessmetnts in teacher evaluation is growing across the states. In 2011, New Jersey Department of Education (NJDOE) adopted ACHIEVENJ, educators' evaluation and support system under the Teacher Effectiveness and Accountability for the Children of New Jersey (TEACHNJ) policy. The ACHIEVENJ model consists of three components: Student Growth Percentile (SGP), Student Growth Objectives (SGO) and a few classroom observations by the school administrator using one of the rubrics approved by the NJDOE (New Jersey Department of Education [NJDOE], 2015). The SGPs are the student scores from the standardized tests and are available only for teachers teaching tested grades or subjects such as mathematics and English. The SGOs are student assessments designed locally, by educators, or by commercial organizations and implemented by teachers (Riordan, Lacireno-Paquet, Shakman, Bocala, & Chang, 2015). The students' outcomes from the SGP and SGO tests are included in teacher summative evaluation. The inclusion of these components provides three measures of teacher effectiveness.

The high amplitude of requirements in teacher evaluation raised a concern of whether the New Jersey public school teachers were competent in assessment theory to effectively utilize the state mandated tests, especially the SGO assessments. The rational for this inquiry was that designing any type of assessment is a complex process. An educator who is undertaking this task needs to know the fundamental principles of assessment theory. The poorly developed and incorrectly implemented assessments produce unreliable results and inaccurate inferences about teaching and learning. To address this concern the researcher posted a question: What is the level of competency in student educational assessments of New Jersey public school teachers? To advance the prior

studies the comparison of teachers' competence level occurred among different groups of teachers. The following research questions facilitated the development of this study:

- What is the statistical comparison of assessment literacy level between teachers from high and low achieving public schools in the state of New Jersey?
- What is the statistical comparison of assessment literacy level between elementary, middle and high public school teachers in the state of New Jersey?
- What is the statistical comparison of assessment literacy level between groups of teachers who taught 0-4 years, 5-10 years, 11-20 years, and more than 21 years?
- What is the statistical comparison of assessment literacy level between tested and nontested teachers?
- Does a statistically significant difference exists in the level of assessment literacy between groups of teachers based on level of education attained?

The study followed quantitative nonexperimental methodology. The data collection occurred via e-mail by utilizing Assessment Literacy Inventory (ALI) developed by Mertler and Campbell (2005). The inferential analysis included an independent-sample *t*-test and One Way Analysis of Variance (ANOVA) test at α=0.05 level. Eighty two (N=82) teachers from demographically different schools participated in this study forming a purposive sample size. The applications of this study may lead to informed teacher professional development course, improved administrative decisions pertaining to SGO tests and district wide system of student assessments. On a big scale this study may influence the development of teacher evaluation policy in the state of New Jersey. The most important change may occur in teachers' assessment practices.

2. Review of the Literature

Preparing students for life, career, and college became a mission for public schools in the United States. The Smarter Balanced Assessments (SBA) consortium and the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium assembled a common set of K-12 assessments in core academic subjects to create a pathway to college and careers readiness for all students (Herman & Linn, 2013; Rentner & Kober, 2014). The school participation in PARCC or SBA testing compelled classroom teachers to raise students' scores on standardized tests in accordance with the federal and state policies. The assessments became catalysts for improving students' achievements and bonds between curriculum and instructions. Simultaneously, teacher assessment literacy became a provision for effective instructions. As a result, the federal and state educational agencies pressured school districts to adopt and implement evaluation systems that measure teacher effectiveness based on students' performance on standardized tests (Baker, Oluwole, Green, & Preston, 2013).

2.1 The Students Tests Scores in Teacher Evaluation

The last school improvement initiative, Race to the Top (RTTT), encouraged school districts across the country to include students' achievements on standardized tests in teacher evaluation for the exchange of RTTT's federal monies (Mathis, 2011; Onosko, 2011). The USDOE allocated 350 million dollars to support the states in which teachers' evaluation was based on 50% of student achievements' scores (Onosko, 2011). To execute the federal requirements statisticians developed the Value Added Models (VAM) to measure teachers' effectiveness. In VAMs, the students' previous tests scores used to predict the future scores on assumption that students perform approximately the same each year. The difference between predicted and current scores considered as a teacher or school contribution into students' learning (Gitomer, 2011; Groen, 2012; Marder, 2012).

The other statistical approach to estimate teacher effectiveness is SGP which is similar to VAMs. The SGP measures the student academic growth from one year to the next compared to students with a similar performance history from across the state academic peers (NJDOE, 2015). The utilization of the SGP scores occurs in the following order: The statisticians assign a percentile rank to each student; teachers receive the percentile ranks of all students they taught this year; each teacher receives a score for the year as a median value of the percentile ranks of her or his students (Gill, English, Furgeson, & McCullough, 2014). The SGP scores are available only for teachers who teach mathematics and English from grade three to eight. During these years, students are taking the state standardized tests in mathematics and English.

The inclusion of students' scores in teacher evaluation brought a big resonance into research and practice. Two polarized understanding regarding teacher evaluation exist in educational community. Some stakeholders believe that the students' achievements justify how well teachers perform (Hanushek & Haycock, 2010). The others think that teachers should not be evaluated based on students' tests scores because they do not have control of many factors affecting students' learning (Gitomer, 2011).

The allies of VAM argued that focusing on students' achievement gains helps to eliminate ineffective teachers from low-performing schools (Hanushek & Haycock, 2010). According to Darling-Hammond (2014) the elimination of 5% to 10% of ineffective teachers every year will increase student academic achievements, and the United States will catch up with the high performing nations. The opponents of the VAMs argued that firing teachers will not solve the problems in teacher evaluation and it will not raise the students' scores on standardized tests. Hendrickson (2012) speculated that, in Finland, which has one of the strongest education system in the world, educators pay a little attention to evaluating teachers. Instead, the Finnish educators devote more resources to developing collaborative relationship and collective learning among colleagues to promote student learning (Hendrickson, 2012).

The substantial research on VAMs and SGP raised questions about the validity of the statistical formulas used to estimate teacher's effectiveness. One of the questions was that whether the VAM like estimates accurately measure contributions of special education or English Language Learners teachers (Baker et al., 2013). Gitomer (2011) argued that the VAM does not support the theory of teaching and learning because the changes in scores do not explain what teacher did or did not do to improve students learning. Gitomer argued that the good example of controversy is in fact that 10 points gain in the lower scale is equivalent to 10 points gain on the higher scale. The student with the baseline score of 95% may not show increase and teacher of such student was estimated to be ineffective based on VAM formulas (Gitomer, 2011). The variables other than teacher alone influenced the VAMs estimate (Baker et al., 2013). According to Baker et al. the VAMs produced a different rating for teachers compared to other evaluation instruments. The student mobility and missing data compromised VAM's outcomes (Baker et al., 2013). These facts disrupted the link between the teacher and student. The student prior knowledge, enriched summer classes, private tutor, family background and socio-econimic status brought unfairness to the VAM's outcomes (Baker et al., 2013). Finally, the nonrandom student assignment to teacher, the classroom composition and school functioning added biases into statistical formulas (Marder, 2012; Onosko, 2011).

The SGO assessment is an alternative way to measure teachers' contribution to students learning, especially for teachers teaching nontested grades or subjects (Riordan et al., 2015). The SGOs are "academic goals for different groups of students that are aligned to the state standards and can be tracked using objective measures" (NJDOE, 2015, p. 3). The SGO development varies from district to district and state to state. The SGO tests may be developed by utilizing local resources or by using commercial educational organizations. The state or school districts decide what type of the SGO to use for teacher evaluation. The SGO developed by teacher was the most common type. According to Lacireno-Paquet, Morgan and Mello's (2014) study twenty-three states mandated teachers to individually develop the SGO tests. In general, the SGO implementation begins with the developing or selecting appropriate assessments followed by preassessment or diagnostic tests (Gill et al., 2014). Based on the preassessments results teacher sets the learning targets for the entire class, or group of students or individually for each student (Gill et al., 2014). Than teacher chooses measures to evaluate each student or group of students' proficiency level. At the end of the school year, teachers administer the post SGO test to measure students' growth (Gill et al., 2014).

2.2 The SGO Assessments in Teacher Evaluation

The architects of the SGO asserted that this test perfectly supplies data of students' growth due to teacher factor. The first onset of the SGO in the state of New Jersey began in 2011, yet it is not fully understood the validity and reliability of this test or the way schools utilize this assessment. Gill et al. (2014) believed that the locally developed SGOs compromise the comparability and reliability of the test. In effort to validate the SGO test, Hu (2015) conducted a quantitative study in Charlotte-Mecklenburg, North Carolina. One hundred and fifty nine schools participated in Hu's study yielding to 18,800 teachers.

To note that, the teacher evaluation in North Carolina is similar to ACHIEVENJ. Both evaluations include the SGO test and the classroom observation by the school principal. The difference between two evaluations is that the ACHIEVENJ includes the SGP scores while the North Carolina evaluation includes the VAM scores.

Hu's (2015) goal was to find correlation between the quality SGO and VAM's scores. Hu hypothesized that the VAM and SGO scores should be similar in estimating teachers' effectiveness. Hu found 67% positive correlation and 33% negative correlation between the VAM and quality SGO across years and grades in mathematics, and 73% positive correlation and 27% negative correlation across the grades and years in reading. The student race and ethnicity were significant predictors in the models for both mathematics and reading across the grades and years (Hu, 2015). The class size as a factor varied across the grades (Hu, 2015).

Hu (2015) stated that after controlling for the extraneous variables such as class size, prior student achievements and background characteristics the higher quality SGO corresponded to higher teachers' VAM scores in mathematics and reading. No statistically significant findings of the relationship between the VAM and SGO attainment status existed. Hu explained this fact that some effective teachers did not have skills in designing quality SGO, or some ineffective teachers were skilled in developing quality SGO. Hu suggested that the school districts should not use VAM or SGO to make high stake decisions about teacher practices because there were many factors influencing instructions and learning.

Pollins (2014) explored SGO as a process and its implication on teachers and school administrators. Pollins found positive impact of the SGO on elementary and middle school teachers' practices in Rhode Island public schools. According to Pollins, teachers and administrators agreed that the SGO increased collaboration among colleagues and opened dialogue about students' learning and common assessments. Additionally, Polins reported that teachers encountered obstacles during the SGO process. The Rhode Island teachers stated that they rarely used quality assessments for the SGO purposes, and teachers needed directions of how to create or choose student assessments for the SGO (Pollins, 2014). Likewise, Pollins suggested not to use the outcomes from the SGO for the decisions related to teacher retention, pay and evaluation.

The theory behind the SGO tests is effective teaching through the quality assessments. The SGOs are assessments with multiple roles: To communicate to students learning goals and expectation, to provide students feedback, to help teachers to monitor students learning, to adjust instructions and to measure students' academic growth (Lacireno-Paquet et al., 2014; Gill et al., 2014; NJDOE, 2015). The instructional planning represents the SGO's main role. The inclusion of students' outcomes from the SGO assessments in teacher evaluation conflicts with the SGO's primarily role.

2.3 Teacher Evalaution in the State of New Jersey

In the state of New Jersey, ACHIEVENJ teacher evaluation consists of three measures: The SGP scores, the SGO scores, and the classroom observation by the school principal using one of the teacher evaluation models approved by NJDOE. According to NJDOE (2015), the Danielson Model was the most popular in the state of New Jersey. One hundred and thirty six out of 571 school districts in the state of New Jersey utilized Danielson Model. The second most popular evaluation was Stronge Teacher and Leader Effectiveness Performance System, 65 school districts in the state of New Jersey utilized this model (NJDOE, 2015). The third most popular was Marzano Causal Teacher Evaluation Model, 53 school districts in the state of New Jersey used Marzano Model (NJDOE, 2015). The Danielson Model has a 4-tiered rubric: Highly effective, effective, partially effective and ineffective. The scores of a highly effective teachers are in the range from 3.5 to 4; the scores of effective teachers are in the range from 2.65 to 3.49; the scores of a partially effective teachers are in the range from 1.85 to 2.64 and for ineffective teachers the score are in the range from 1.0 to 1.84 (NJDOE, 2015). The SGO score has a 4-tiered rubric with the same range in scores for four levels of performance. Teachers without the SGP score receive a summative score which combines 20% of the SGO and 80% of the classroom observations. Teachers with the SGP scores receive a summative score of 10% of the SGP median, 20% of the SGO and 70% of classroom observations.

Callahan and Sadeghi (2015) conducted a statewide study to investigate three phenomena: New Jersey teachers' perceptions of ACHIEVENJ, the level of communication between teachers and administrators, and the availability, frequency and effectiveness of the professional development opportunities. Callahan and Sadeghi reported that teachers perceived ACHIEVENJ model as unfair that does not accurately evaluates their teaching abilities. According to Callahan and Sadeghi teachers reported that the number of classroom observations increased and resulted in increased professional dialog about instructions, students' assessments, and students learning. Furthermore, teachers stated that the quality of the observations decreased because administrators spent more time entering the evidences and information into laptops than actually observing teachers (Callahan & Sadeghi, 2015). In 2014, 56% of teachers wanted more professional development related to their areas of need, and only 5% of teachers reported that they were satisfied with the training they received (Callahan & Sadeghi, 2015). According to New Jersey teachers the implementation of the ACHIEVENJ did not address poor practice, excellent teachers were not recognized, novice teacher did not receive support, and professional learning was not tailored to teacher's needs (Callahan & Sadeghi, 2015). In support of the SGO method the NJDOE advocates that a thoughtfully developed and collaboratively implemented SGOs improve the quality of discussion about student growth, learning and instructions. The SGO method increases teacher engagement in the assessment practices, enriches teacher knowledge of curriculum standards and it fosters teacher leadership (Gill et al., 2014; NJDOE, 2015).

The induction of a new measure in teacher evaluation intensified the role of educational assessments in the process of improving learning outcomes. The teacher assessment literacy became a policy consideration. Popham (2011) underlined two reasons for teachers to become an assessment literate: To understand how the accountability assessments determine educator's professional quality, and to understand how assessments improve students learning and instructions. The modern schools, according to Popham need educators who are competent in theory of assessments, and effectively use assessments to make instructional and administrative decisions.

2.4 Teacher Assessment Literacy

Teacher assessment literacy continues to conquer public interest. The stakeholders in education want to know how teachers utilize assessments to evaluate students' learning. Educational specialists defined an assessment literate teacher as an expert who is able to select or develop and administer different types of assessments, use the data from the assessments to inform instructional decisions, and to communicate the assessments results to students and their parents (Mertler & Campbell, 2005; Popham, 2011; Stiggins, 2002). Measuring teacher assessment literacy shortly began after the committee of Teaching Profession the American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), and National Education Association (NEA) developed the Standards for Teacher Competence in Educational Assessment of Student in 1990 (American Federation of Teachers [AFT], National Council on Measurement in Education [NCME], & National Education Association [NEA], 1990).

Plake et al. (2005) used the Standards to develop Teacher Assessment Literacy Questionnaire (TALQ) instrument. Five hundred and fifty-five teachers from 98 school districts in 45 states completed the TALQ survey. Plake et al. reported teachers' score as 66% on overall instrument. According to Plake et al. teachers answered 23 items out of 35 correctly. In 2003, Mertler and Campbell (2005) replicated Plake et al.'s study using TALQ. Two hundred and twenty undergraduate preservice teachers participated in the study. Mertler and Campbell reported that teachers answered 21 items out of 35 correctly yielding to 60% of overall teachers' score. Mertler and Campbell believed that TALQ's questions were difficult and lengthy to read. Mertler and Campbell modified TALQ into Classroom Assessment Literacy Inventory (CALI) and later to Assessment Literacy Inventory (ALI).

In 2013, Davidheiser (2013) measured 180 high school teachers' assessment literacy in the state of Pennsylvania reporting that teachers answered less than 25 items out of 35 correctly. Simultaneously, Perry (2013) measured 14 teachers' and 32 principal's assessment literacy in the state of Montana. Perry reported that on average teachers answered 22 items out of 35 correctly while principals answered 21 items out of 35 correctly on the same instrument.

Barone (2012) investigated Albany Middle and Albany High schools teachers' assessment literacy before the professional development and after as measured by ALI. Barone concluded that as teachers became more knowledgeable in assessments their correct scores on ALI, on average, increased from 73% to 77%. Barone reported a strong correlation between two administrations of the test as 0.97 ($r = 0.97$, $p < 0.05$). Barone attributed the increase in teachers' scores to the effect of the professional development.

The research in the field of education produced a substantial knowledge related to preservice teachers' assessment literacy. Siegel and Wissehr (2011) explored the assessment literacy of 11 preservice teachers enrolled in secondary science methods course of their teacher preparation program. Siegel and Wissehr found that during the methods course preservice teachers identified 19 assessments tools, described their advantages and disadvantages, and evidently demonstrated how to align instructional goals to specific assessments. Siegel and Wissehr concluded that the knowledge of the assessment tools attained by teachers does not constitute its realization in the future classrooms. The working conditions and reality of the classrooms may impede application of the acquired knowledge.

Wallace and White (2015) investigated how secondary mathematics preservice teachers' perspectives of the assessment practice evolved during one year of reform based preparation programs. Six preservice teachers in the state of California participated in the study. The findings of Wallace and White's study showed that the perception of the assessment practices of preservice teachers evolved through three different stages. As course of the program continued to unfold the teachers' viewpoints progressed from the test oriented to task oriented and to tool oriented (Wallace & White, 2015). According to Wallace and White during the test oriented stage teachers used the limited number of assessments only for one purpose to evaluate and grade students' work. During this stage teachers utilized criterion referenced tests which included strictly procedural items with no connections between the concepts (Wallace & White, 2015). During the task oriented stage teachers began to recognize that the test could be utilized for different purposes. The test became criterion referenced and student referenced with

some connections between the concepts (Wallace & White, 2015). During the tool oriented stage teachers realized that the main purpose of the assessment is to improve teaching and learning. The assessments significantly transformed from the traditional, with only one purpose to measure, to reform-based for learning purposes. The format of the tool oriented tests included items that require students to apply exploration, reasoning and analysis to solve the problems (Wallace & White, 2015).

Odo (2015) believed that developing teachers' assessment literacy requires a significant attention during teacher preparation programs. According to Odo, teachers' familiarity with the assessment assortment, available for their use, improves instructions in the diverse classrooms. Teachers' understanding of fundamental characteristics of the assessments such as validity, reliability and bias will help teachers to interpret the standardized tests proliferating in public schools (Odo, 2015). The results of this study suggested that teacher education related to student assessments should continue after teachers enter classrooms and should be given a thoughtful attention during teacher entire career.

3. Theoretical Foundation

The contemporary teacher evaluation models are based on Professional Teaching Standards developed by the National Board for Professional Teaching Standards (NBPTS) organization in 1987 (Darling-Hammond, 1999). The NBPTS's policy statement, *What Teachers Should Know and Be Able to Do*, clearly communicates to stakeholders professional standards for teachers. The Professional Teaching Standards provide a common language to educators to discuss teaching and learning. The authors of the NBPTS formulated five core propositions as follows:

- Teachers are committed to students and their learning.
- Teachers know the subjects they teach and how to teach those subjects to students.
- Teachers are responsible for managing and monitoring student learning.
- Teachers think systematically about their practice and learn from experience.
- Teachers are members of learning communities.

Parallel to the Standard for Teaching Profession the AFT, NCME and NEA committee declared that good teaching occurs through the effective methods of assessing students learning (AFT, NCME, & NEA, 1990). In 1987 the AFT, NCME, and NEA committee developed the Standards for Teacher Competence in Educational Assessment of Student (AFT, NCME, & NEA, 1990). The goal was to establish standards that will guide educators in designing and implementing student assessments, in identifying needs for professional development and designing professional development for inservice teachers (AFT, NCEM, & NEA, 1990). The AFT, NCEM, and NEA committee suggested to incorporate the standards in teacher training and certification program before standards are included in teacher evaluation systems. The AFT, NCEM, and NEA committee formulated the Standards for Teacher Competence in Educational Assessment of Student as follows:

- Standard 1: Teacher should be skilled in choosing assessment methods appropriate for instructional decisions.
- Standard 2: Teacher should be skilled in developing assessment methods appropriate for instructional decisions.
- Standard 3: Teachers should be skilled in administering, scoring, and interpreting the results of both externally-produced and teacher-produced assessment methods.
- Standard 4: Teacher should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
- Standard 5: Teacher should be skilled in developing valid pupil grading procedures that use pupil assessments.
- Standard 6: Teachers should be skilled in communicating assessments results to students, parents, other lay audience, and other educators.
- Standard 7: Teacher should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

The Standards for Teaching Profession and the Standards for Teacher Competence in Educational Assessment of Student served as a theoretical framework for this study. The student educational assessments and teacher assessment literacy was discussed within the scope of the standardized teacher evaluation. The Danielson Model

for Teaching and Learning served as a theoretical foundation for this study because almost every domain of teacher practices in Danielson Model includes the assessment element as an effective instructional method. All domains in Danielson Model are rooted in Professional Teaching Standards and reflect the Standards for Teacher Competence in Educational Assessment of Student. According to Danielson (2013), teachers need to know how to design quality assessments that have the ability to provide wide range of evidences of students learning.

The Danielson Model consists of four domains and 72 elements of teacher practices: Domain 1 Planning and Preparation, Domain 2 Classroom Environment, Domain 3 Instruction, and Domain 4 Professional Responsibilities (Danielson, 2013). The model incorporates four tier-rubric to rank teacher performance in each domain: Unsatisfactory, Basic, Proficient, and Distinguished (Danielson, 2013).

4. Methodology

This study applied non-experimental quantitative methodology and followed causal-comparative design. The research's questions, data collection, sample size, instrumentation and variables determined the method and design. Each research question had the following hypothesis:

RQ1: What is the statistical comparison of assessment literacy level between teachers from high and low-achieving public schools in the state of New Jersey?

Ho: No significant statistical difference exists in assessment literacy level between teachers from high and low achieving public schools in the state of New Jersey.

Ha: A significant statistical difference exists in assessment literacy level between teachers from high and low achieving public schools in the state of New Jersey.

RQ2: What is the statistical comparison of assessment literacy level between elementary, middle and high public school teachers in the state of New Jersey?

Ho: No significant statistical difference exists in assessment literacy level between elementary, middle and high public school teachers in the state of New Jersey.

Ha: A significant statistical difference exists in assessment literacy level between elementary, middle.

RQ3: What is the statistical comparison of assessment literacy level between groups of teachers who taught 0-4 years, 5-10 years, 11-20 years, and more than 21 years?

Ho: No significant statistical difference exists in assessment literacy level between groups of teachers who taught 0-4 years, 5-10 years, 11-20 years, and more than 21 years.

Ha: A significant statistical difference exists in assessment literacy between groups of teachers who taught 0-4 years, 5-10 years, 11-20 years, and more than 21 years.

RQ4: What is the statistical comparison of assessment literacy level between tested and nontested teachers?

Ho: No significant statistical difference exists in assessment literacy level between tested and nontested teachers.

Ha: A significant statistical difference exists in assessment literacy level between teachers of tested and nontested subjects.

RQ5: Does a statistically significant difference exists in the level of assessment literacy between groups of teachers based on level of education attained?

Ho: No significant statistical difference exists in assessment literacy level between groups of teachers based on level of education attained.

Ha: A significant statistical difference exists in assessment literacy level between groups of teachers based on level of education attained.

The inferences about the group differences, stated in research questions, required statistical analysis based on theory of probability. The hypothesis testing approach assisted the analysis of each question (Bock, Velleman, & DeVeaux, 2010). The null hypothesis for each question stated that there is no difference in assessment literacy level between the means among the groups of teachers ($\mu 1 = \mu 2$) while the alternative hypothesis stated that there is a difference in assessment literacy level between the means ($\mu 1 \neq \mu 2$, two-tailed). The null hypothesis was rejected when the probability of occurrence of the value for the null hypothesis was less than 5% (at $\alpha = 0.05$) concluding that there was an evidence that the statistically significant differences in groups means exist in population.

The other reason for the quantitative approach was the data collection method. In quantitative studies the data is gathered by structured instruments (Johnson & Christensen, 2014). The data collection for this study occurred by utilizing ALI instrument, developed by Mertler and Campbell (2005), specifically for quantitative analysis. Furthermore, the researcher of this study collected data from a large population of 789 school teachers which is another trait of quantitative approach. The sample size of 82 is considered to be a large enough to compute sample's statistics that accurately reflect the population parameters (Bock et al., 2010). The other trait of the quantitative method was the nature of variables. The dependent variables were quantitative: Teachers' composite scores on Standards. Finally, this study may be replicated by the future researchers which is considered as an important characteristic of a quantitative methodology (Bock et al., 2010; Johnson & Christensen, 2014).

4.1 Population and Sample

The general population of this study was 139,699 certified school teachers employed in 694 operating school districts in the state of New Jersey (NJDOE, 2015). The target population combined school teachers from three school districts and one high school. The school districts and high school received pseudonyms as, SD#1, SD#2, SD#3 and HS#4 to maintain confidentiality regarding participation.

Until 2015 in the state of New Jersey the students' outcomes from two standardized tests, High School Proficiency Assessment (HSPA) and New Jersey Assessment of Skills and Knowledge (NJASK) in elementary and middle schools, determined the school district performance. At the time of the study, based on HSPA and NJASK results, the SD#1 and SD#2 were suburban high achieving school districts and SD#3 and HS#4 were low achieving urban schools. The SD#1 employed 201 teachers, the SD#2 employed 335 teachers, the SD#3 employed 175 teachers, and the HS#4 employed 69 teachers. The target population was 798 participants. In total, 169 teachers, 21%, responded to survey. Only 82 (N = 82) fully completed responses defined the purposive sample size.

Sullivan and Feinn (2012) suggested calculating the effect size to estimate reasonable sample size before the study is carried out. The Cohen's d value for the independent-sample t tests and ANOVA tests aided the analysis of the effect size. According to Sullivan and Feinn, Cohen classified effect sizes as small (d = 0.2), medium (d = 0.5), and large (d > 0.8). To estimate an effect size, Sullivan and Feinn suggested to pilot the study, or use the estimates from the similar work published by other researchers, or use the minimum difference that are considered important by experts.

The utilization of G*Power 3.1 allowed to calculate the minimum sample size required to produce statistically significant results. Base on anticipated effect size d = 0.45 and confidence level of 95% (α = 0.05) the Power Analysis for the F-tests was 0.9541217 for the total sample size of 81 participants. The utilization of the Statistical Package for the Social Sciences IBM SPSS Statistics Base 21 (SPSS) software facilitated analysis of collected data. The applications of the descriptive statistics aided the description of the study sample. The applications of inferential statistics expedited the answers to the study questions.

4.2 Data Collection Method

During the winter and spring of 2016, 165 superintendents of public school districts received invitations to participate in this study. The researcher made phone calls to school superintendent's offices following alphabetical order in which districts were listed on the NJDOE publicly open website. Three superintendents and one high school principal agreed to participate in the study. In May 2016, 798 public school teachers from participating schools received an online survey, ALI (Mertler & Campbell, 2005) via school e-mails. Two weeks later the researcher e-mailed a reminder to teachers to complete the survey. The access to the survey was opened for five weeks.

4.3 Instrumentation

The members of the American Educational Research Association (AERA) reviewed ALI instrument in Montreal in 2005 and concluded that the instrument is a reliable tool to measure teacher assessment literacy (Mertler & Campbell, 2005). After the AERA review the educational researchers widely used ALI instrument. Hamilton (2014) conducted a quantitative study to investigate to what extent teacher assessment literacy as measured by ALI associated with the teacher knowledge of Curriculum Based Measurement (CBM). The CBM is another research based instrument measuring educators skills in assessing students' knowledge of curriculum taught in the classroom. The study took place in elementary schools in Rhode Island. Hamilton found positive significant correlation between two instruments (r = 0.505, p < 0.01). Hamilton reported that teachers with the high scores on ALI tended to have high scores on the CBM, and vice versa. Hamilton concluded that the two instruments measured the same construct.

After testing ALI on 152 preservice teachers, Mertler and Campbell (2005) reported instrument reliability, *rKR*20, as 0.75, the mean difficulty as 0.64, and mean item discrimination as 0.32. These values, according to Mertler and Campbell described ALI as a reliable instrument from the psychometric point of view. The administration of Cronbach Alpha test provides the measure of the instrument's internal consistency, the extent to which all items in the instrument measure the same construct or concept (Tavakol & Dennick, 2011). Davidheiser (2013) reported the Cronbach Alpha coefficient for the ALI survey as 0.824 on 35 items. For this study the Cronbach Alpha was 0.772 on 35 items. For results see Figure 6 in Appendix P. Tavakol and Dennick (2011) suggested that the value of Cronbach Alpha should be in the range between 0.70 and 0.90.

The ALI instrument consists of two sections. In the first section participants provided information regarding the number of years teaching, subject matter and grade level taught, and educational degree attained. The second part of the survey had five scenarios followed by seven questions. Teachers had to read each scenario and answer 35 questions. The 35 questions in ALI instrument reflected Seven Standards of Teacher Competence in Student Educational Assessments. Table 1 demonstrates an alignment between the Standards and ALI items.

Table 1. Alignment of Standards with ALI Items

Standards for Teacher Competence	ALI Items Numbers
Standard 1 Choosing Appropriate Assessment Methods	Items 1,8,15,22,29
Standard 2 Developing Appropriate Assessment Methods	Items 2,9,16,23,30
Standard 3 Administering, Scoring, and Interpreting the Results of Assessments	Items 3,10,17,24,31
Standard 4 Using Assessments Results to Make Decisions	
Standard 5 Developing Valid Pupil Grading Procedures	Items 4,11,18,25,32
Standard 6 Communicating Assessments Results	Items 5,12,19,26, 33
Standard 7 Recognizing Unethical or Illegal Practices	Items 6, 13, 20,27, 34
	Items 7,14,21,28,35

5. Results and Discussions

5.1 Research Question 1

What is the statistical comparison of assessment literacy level between public school teachers from high and low achieving schools in the state of New Jersey?

On average, teachers from high achieving schools performed better on every standard compared to teachers from low achieving schools. Figure 1 shows New Jersey teachers' composite score based on school assignment.

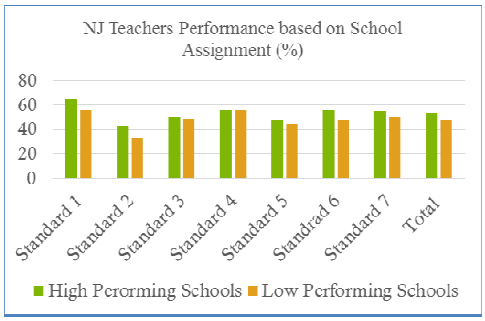

Figure 1. NJ teachers' performance based on school assignment

Based on descriptive statistics, the high achieving district, SD#1, had the highest average score of 55%. The SD#3, the low achieving district, demonstrated the next highest score of 54%. The lowest average score of 42% was for high school teachers from the low achieving school, HS#4. The high achieving district, SD#2, demonstrated the score of 52% for the entire instrument.

The pattern in performance related to standards emerged as follows: Regardless of the school assignment, the highest average score was for Standard 1. The Standard 1 refers to teacher skills and knowledge in selecting assessment methods pertinent for instructional decisions. Danielson (2013) defined a distinguished teacher as an expert who utilizes different methods to assess students' learning and who uses assessments outcomes to design instructions. Importantly to note, that the SGO process requires teacher to select a high quality student assessments. On average, teachers' score for Standard 1 was 61% (M = 0.61, SD = 0.23). The high achieving school district, SD#1, showed the highest average score of 71% (M = 0.71, SD = 0.24). The lowest average score of 56% occurred for HS#4 (M = 0.56, SD = 0.28) and SD#3 (M = 0.56, SD = 0.16). The average score for Standard 1 for SD#2 was 59% (M = 0.59, SD = 0.21).

Regardless of the school assignment, the lowest average score was for Standard 2. To note, that the SGO process requires teachers to develop a quality assessments. The SD#2 demonstrated the highest score of 44% (M = 0.44, SD = 0.23). The next highest average score of 42% was for SD#1 (M = 0.42, SD = 0.22). The HS#4 demonstrated the lowest score of 25% (M = 0.25, SD = 0.26). The SD#3 score was 41% (M = 0.41, SD = 0.21). For results see Table 2 in Appendix A. The inferential analysis did not demonstrate evidences that the statistically significant differences exist between the teachers form high and low achieving schools.

In comparison to previous studies, New Jersey teachers demonstrated different results. Perry (2013) reported that the highest performance was for Standard 4 (M = 4.07; maximum possible score = 5). The lowest performance was for Standard 7 (M = 1.29; maximum possible score = 5). Davidheiser's (2013) findings related to Standard 2 were parallel to the findings of this study. Davidheiser reported that the lowest performance occurred for Standard 2 (M = 0.57; maximum possible score = 1). Davidheiser reported the highest performance for Standard 7 (M = 0.79; maximum possible score = 1).

5.2 Research Question 2

What is the statistical comparison of assessment literacy level between elementary, middle and high public school teachers in the state of New Jersey?

On average, middle and elementary school teachers outperformed their peers from high schools. Figure 2 demonstrates New Jersey teachers' composite score on standards based on grade level taught.

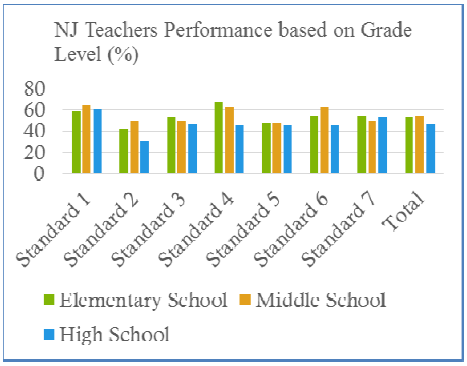

Figure 2. NJ teachers' performance based on grade level

The pattern in teachers' performance has emerged as follows: The middle school teachers performed higher on Standards 1, 2 and 6. The average score on Standard 1 was 65% (M = 0.65, SD = 0.22). The average score on Standard 2 was 49% (M = 0.49, SD = 0.25). The average score for Standard 6 was 63% (M = 0.63, SD = 0.26). The elementary school teachers performed higher on Standards 3 and 4. The average score on Standard 3 was 53% (M = 0.53, SD = 0.20). The average score on Standard 4 was 67% (M = 0.67, SD = 0.23). The high school teachers demonstrated the lowest scores for every standard accept for Standard 7. The average score for Standard 7 was 53% (M = 0.53, SD = 0.27). For results see Table 3 in Appendix B.

Inferential analysis demonstrated an evidence that the statistically significant differences between the teachers from elementary, middle and high schools exist for Standard 2. The significance, p values at α = 0.05 level for ANOVA test was 0.019 (p = 0.019, $p < 0.05$). The significance, p value for the post hoc, Tukey's HSD at α = 0.05 level was 0.016 (p = 0.016, $p < 0.05$). The Tukey's HSD test detected the differences between the middle and high school teachers. The middle school teachers performed higher compared to high school teachers. For

Standard 2, the group mean for middle school teachers was 0.49 (M = 0.49, SD = 0.25) and the group mean for high school teachers was 0.31 (M = 0.31, SD = 0.22). For results see Table 8 in Appendix G.

Again, Standard 2, refers to teachers' knowledge and skills to develop quality assessments to improve teaching and learning (Danielson, 2013). An elementary and middle school teachers may be more skilled in developing assessments and using the assessments outcomes for instructional decisions. Perhaps, the pressure of standardized testing in middle and elementary schools encourages teachers to assess students more frequently to better prepare them for the rigorous tests.

Based on inferential analysis, the statistically significant differences related to Standard 4 existed between the middle and high school teachers and between elementary and high school teachers. The significance, p values for ANOVA test at α = 0.05 level was 0.001 (p = 0.001, $p < 0.05$). The significance, p value for the post hoc, Tukey's HSD at α = 0.05 level was 0.008 (p = 0.001, $p < 0.05$). The elementary school teachers performed higher compared to high school teachers on Standard 4. The group mean for elementary school teachers was 67% (M = 0.67, SD = 0.23) and the group mean for high school teachers was 45% (M = 0.45, SD = 0.22) for Standard 4. The middle school teachers performed higher on Standard 4 compared to high school teachers. The significance, p value for the Tukey's HSD at α = 0.05 level was 0.008 (p = 0.008, $p < 0.05$). The group mean for middle school teachers was 63% (M = 0.63, SD = 0.24) and the group mean for high school teachers 45% (M = 0.45, SD = 0.22). No evidences occurred that statistically significant differences exist between the middle and elementary school teachers. For results see Table 8 in Appendix G.

The Standard 4 requires teachers to use the assessment information for instructional planning and to manage student educational development (AFT, NCEM, & NEA, 1990). Danielson (2013) defined a distinguished teacher as an expert in using various types of assessment to direct instructional improvements. Importantly to note, that Protheroe (2009) argued that students demonstrated an increase in scores on standardized tests in schools with staff trained to use assessment data effectively.

Based on inferential analysis, statistically significant differences related to Standard 6 existed between the middle and high school teachers. The significance, p values for ANOVA test at α = 0.05 level was 0.028 (p = 0.028, $p < 0.05$). The significance, p value for the post hoc Tukey's HSD test at α = 0.05 was 0.024 (p = 0.024, $p < 0.05$). The middle school teachers performed higher compared to their peers form high schools. The group mean for middle school teachers was 63% (M = 0.63, SD = 0.26) and the group mean for high school teachers was 45% (M = 0.45, SD = 0.23). For results see Table 8 in Appendix G.

The Standard 6 refers to teacher' ability to communicate assessments results. A distinguished teacher, according to Danielson (2013), needs to know how to interpret the tests' results and communicate the information about students' learning to students, parents and other educators. An assessment literate teacher selects or develops and administers different types of assessments, uses the data from the assessments to inform instructional decisions, and communicates the assessments results to students and their parents (Mertler & Campbell, 2005; Popham, 2011; Stiggins, 2002).

5.3 Research Question 3

What is the statistical comparison of assessment literacy level between groups of teachers who taught 0-4 years, 5-10 years, 11-20 years, and more than 21 years?

The years of experience was not a significant factor in teacher performance. On average, teachers with more than 10 years of experience performed slightly higher compared to other two groups. Figure 3 demonstrates New Jersey teachers' composite score on standards based on years of experience.

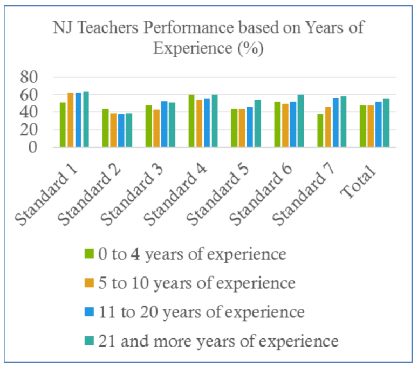

Figure 3. NJ teachers' performance based on years of experience

Based on descriptive statistics, on average, teachers with 21 and more years of experience performed higher compared to other groups. The average score of teachers with 21 and more years was 55%. The second highest score of 52% was for teachers with 11 to 20 years of experience. Teachers with 0 to 4 years and 5 to 10 years of experience had the same score of 48%. Table 4 in Appendix C shows teachers composite scores based on years of experience.

In comparison to previous study, New Jersey teachers demonstrated different results. Davidheiser (2013) reported the highest score for the group of teachers with 4 to 10 years of experience (among the high school teachers). Davidheiser reported the lowest score for the group of teachers with 0 to 3 years of experience. Teachers with 11 to 20 years and 21 and 35 years performed similarly (Davidheiser, 2013).

The pattern in teachers' performance for standards emerged as follows: Teachers with 21 and more years of experience performed higher on Standards 1, 4, 5, 6 and 7. Teachers with 0 to 4 years of experience performed higher on Standards 2 and 4. Teachers with 11 to 20 years of experience demonstrated the highest score on Standard 3 only.

Inferential analysis did not show evidences that the statistically significant differences exist between groups of teachers with different years of experience. These findings call for conclusion that teacher knowledge in assessment theory did not change over the course of teacher career. In addition, these findings may imply the lack of professional development related to educational assessment.

5.4 Research Question 4

What is the statistical comparison of assessment literacy level between teachers of tested and nontested subjects?

In this study, tested (mathematics and English) and nontested (all other subject matter) teachers formed two groups. The rationale for the classification was that in the state of New Jersey students from grade 3 to 8 and once in the high school take state standardized tests in mathematics and English. The researcher of this study hypothesized that tested teachers are more literate in assessment theory due to overwhelming testing and preparation for the standardized tests. In total, 33 tested teachers (17 mathematics and 16 English teachers) and 49 nontested teachers participated in this study. The tested and nontested teachers' scores were almost the same with the exception for Standard 7. Tested teachers demonstrated better results. Figure 4 shows New Jersey teachers' composite score on standards based on subject taught.

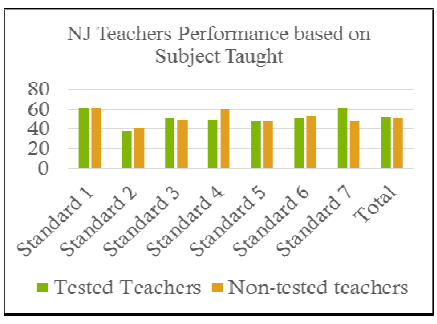

Figure 4. NJ teachers' performance based on subject taught

Based on descriptive statistics, on average, both groups of teachers performed similarly on ALI. The average score of tested teachers was 52% and for nontested was 51%. The pattern in teachers' performance for Standards emerged as follows: Tested teachers performed higher on Standards 1, 3 and 7 and nontested teachers performed higher on Standards 2, 4, and 6. Table 5 demonstrates the teachers composite scores based on tested and nontested classification. For results see Table 5 in Appendix D.

Inferential analysis demonstrated evidences that the statistically significant differences exist between the two groups of teachers for Standard 7. The significance p level for independent sample t-test was 0.003 ($p < 0.05$, two-tailed). The tested teachers performed higher compared to nontested teachers. The mean group for tested teachers was 62% ($M = 0.62$, $SD = 0.26$). The Standard 7 implies to teachers' skills to recognize unethical or illegal practices. Again, this evidence suggested that due to the pressure of the standardized testing tested teachers may receive more training and professional development in the area of assessments. For results see Table 15 in Appendix N.

5.5 Research Question 5

Is there statistically significant difference in the level of assessment literacy between groups of teachers based on level of education attained?

When analyzing data for question 5 only two groups, instead of four, emerged. Twenty-three teachers with bachelor's degree and 57 teachers with master's degree completed the survey forming two groups. The comparison of teachers' assessment literacy occurred only between two groups (teachers with the bachelor and master degrees) because only one teacher with the educational specialist degree and one teacher with the doctoral degree participated in the study. These two teachers were excluded from the inferential analysis.

Teachers with the master's degree outperformed teachers with the bachelor's degree. Figure 5 demonstrates New Jersey teachers' composite score on standards based on educational degree attained.

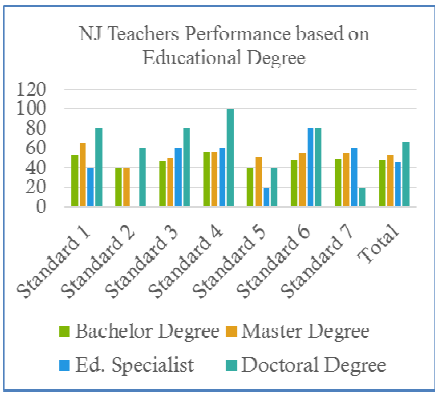

Figure 5. NJ teachers' performance based on educational degree

The composite score of teachers with the master's degree was 53% and the composite score of teachers with the bachelor's degree was 48%. The average score of one teacher with the educational specialist degree was 46% and the average score of one teacher with the doctoral degree was 66%. Based on descriptive statistics the conclusion was that the teacher's educational degree had an impact on teacher assessment literacy. Teachers with master's degree performed higher on every standard. Perhaps, an assessment and measurement course in a master's degree program positively influenced teacher's performance. For result see Table 6 in Appendix E.

Inferential statistical analysis did not demonstrate evidences that the statistically significant differences related to Standard 7 exist in population. In comparison to the previous study New Jersey teachers demonstrated different results. Davidheiser (2013) reported that the group means for teachers with the master degree and bachelor degree were similar. For teachers with bachelor's degree: $M = 24.31$, $SD = 5.536$ (p. 65). For teachers with master's degree: $M = 24.33$, $SD = 4.85$.

6. Limitation

The study had several limitations. The low response rate of 10.3% presented the first limitation. Although, 169 teachers out of 798 responded to the survey, yielding to 21% participation rate, only 82 fully completed surveys

served as a foundation for statistical analysis. A greater number of participants would lead to more precise statistical analysis. However, based on effect size of $d = 0.45$ and confidence level of 95% ($\alpha = 0.05$) the Power Analysis for the *F*-tests was 0.9541217 for the total sample size of 81 participants. The $N = 82$ allowed to sufficiently conduct statistical analysis.

Another limitation occurred during data analysis. When analyzing data for question 5 only two groups, instead of four, emerged. Twenty-three teachers with bachelor's degree and 57 teachers with master's degree completed the survey forming two groups. The comparison of teachers' assessment literacy occurred only between the groups of teachers with bachelor and master degrees.

The conceptual framework for this study was that the assessment literate teachers are effective (Danielson, 2013; Popham, 2011; Stiggins, 2002). This concept presumed the next limitation that the minimal research exists describing the relationship between the level of teacher assessment literacy and students' academic achievements. The school district's vision, mission and professional development plan, preservice teacher preparation program and teacher's individual involvement in professional learning influence teacher's knowledge in assessment theory.

7. Conclusions

On average, teachers from high achieving schools performed better compared to teachers from low achieving schools. The statistically significant differences occurred between the middle, elementary and high school teachers. The middle and elementary school teachers demonstrated better results for Standard 2 (Developing Appropriate Assessment Methods Appropriate for Instructional Decisions), Standard 4 (Using Assessment Results when Making Decisions about Individual Students, Planning Teaching, Developing Curriculum, and School Improvement) and Standard 6 (Communicating Assessments Results to Students, Parents, other Lay Audience, and other Educators). The assessment literacy barely improved as teachers' years of experience increased. The tested and nontested teachers' scores were almost the same, except for Standard 7 (Recognizing Unethical, Illegal, and otherwise Inappropriate Assessment Methods and Uses of Assessment Information). The statistically significant differences occurred between the tested and nontested teachers only for Standard 7. Teachers with the master's degree performed higher on every standard compared to teachers with the bachelor's degree.

New Jersey teachers' average score of 51% in assessment literacy confirmed an alarming statement made by Popham (2011) and Gareis and Grant (2015) that only a few teachers are conversant of how accountability tests work. For instance, only 4% of teachers answered item #28 correctly. This item refers to teachers' skills to implement the standardized tests without breaking standardization. Only 6% of teachers answered item #27 correctly, and 22% of teachers answered item #10 correctly. These items refer to teachers' ability to interpret results from the standardized tests. Eighteen items answered incorrectly by most of the teachers validated Stiggins (2002) statement that the educators were unable to differentiate information obtained from the assessments.

In conclusion, the patterns in teacher assessment literacy, found in this study, indicated that teachers' competence in assessments did not improve during the last two decades. As Gareis and Grant (2015) stated, "the evidence that teachers continue to be ill-prepared in the domain of assessment persists to the present day" (p. 7). The results of this study supported the findings of previous researchers stating that teachers were concerned about the quality of the SGO assessments (Hu, 2015; Pollins, 2014). The data from unreliable SGOs provides misleading inferences about teaching and learning and subsequently do not have impact on teacher practices. The educational policymakers should acknowledge the consequences of utilizing data from the low quality assessments in teacher evaluation. Before holding teachers accountable for the students' achievements on high stake tests the district leaders should re-evaluate the school culture related to assessments. To maximize the intended purposes and potential benefits of the educational assessments and the SGO the school leaders should promote teacher assessment literacy. The SGO test should serve as a tool for instructional improvements and not a measure of teacher effectiveness, and the outcomes from the properly completed SGO process should reflect students' true academic growth.

8. Recomendations

8.1 Recommendation for Future Research

This study was a quick snapshot of New Jersey public school teachers' competence in student assessments. Replicating this study as a statewide research to generalize conclusions to the entire population of school teachers in the state of New Jersey was the initial recommendation. The larger sample size would serve to secure

the sufficient number of participants in different demographic groups to conduct more precise inferential analysis (Bock et al., 2010; Johnson & Christensen, 2014). The results of the statewide research can help to improve the policy related to teacher evaluation in the state of New Jersey.

Another recommendation was developing a rubric or a system to evaluate teacher's proficiency in educational assessments. The developers of the existing instruments did not specify a minimum requirement for teachers in assessment literacy. The next step should be investigating a relationship between the teacher proficiency in assessments and students' academic achievements. The question for the future inquiry is: Do teachers with the higher score on ALI have more proficient students as measured by the standardized tests or school district benchmarks?

The third recommendation involves exploring the relationship between the SGO attainment status and a teacher-assessment-literacy level. The question for the future researcher is: Do teachers with the higher score on ALI have more students achieving their goals as measured by SGO? Conducting a qualitative research to explore how teachers perceive their assessment practices in relation to school culture, district vision and mission; and how teachers use the SGO data for instructional purposes between the pre SGO test and post SGO test was the final recommendation for the future research.

8.2 Recommendations for Future Practice

The first recommendation to school leaders was to provide teachers with support related to all seven standards but with the priorities given to the standards with the lowest scores. Specifically, remediation of how to develop assessments for different purposes because the lowest score was for Standard 2, Developing Appropriate Assessment Methods Appropriate for Instructional Decisions. The next attention should be given to mastering skills in creating rubrics and grading system to evaluate student knowledge because the second lowest score was on Standard 5.

Based on study results, on average, teachers from high achieving schools outperformed their peers from low achieving schools. The school administration of low achieving schools should look at the assessment practices and teacher professional development plan of high achieving schools. On average, the middle and elementary school teachers demonstrated higher results compared to high school teachers. The school administrators can explore the differences in practice related to assessment methods between the middle, elementary, and high schools.

Further, on average, teachers with 21 and more years of experience performed slightly better compared to other groups. The school administration should establish or reinforce the teacher-peer-mentoring program and assign a tenured, highly effective, teacher to a novice teacher. However, when giving an additional responsibility to a teacher, school leaders should think of the future obstacles, and possibly provide to teacher-mentor some kind of rewards or compensations.

On average, teachers with the master's degree performed better on every standard compared to teachers with the bachelor's degree. The school district leaders should encourage teachers to continue their education and professional learning. One of the district options may be providing teachers some incentives and recognition for completing additional courses related to their subject area.

It is important to note that, the researcher of this study did not have an intent to diminish the practical individual knowledge of teachers. The researcher believes that the classroom teachers have fundamental understanding of how the school culture and district's vision influence the assessment practices in the classrooms. The goal of this research was to measure and compare the assessment literacy among different groups of teachers by using quantitative instrument.

References

American Federation of Teachers [AFM], National Council on Measurement in Education [NCME], & National Education Association [NEA]. (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues in Practice*, *9*(4), 30-32. https://doi.org/10.1111/j.1745-3992.1990.tb00391.x

Baker, B., Oluwole, J., Green, P., & Preston, C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archives*, *21*(5). https://doi.org/10.14507/epaa.v21n5.2013

Barone, T. (2012). *Complex assessments, teacher inferences, and instructional decision-making* (Doctoral dissertation). University of California, Berkeley. Retrieved from https://scholar.google.com

Bock, D., Velleman, P., & DeVeaux, R. (2010). *Stats: Modeling the World* (3rd ed.). Boston: Pearson/Addison-Wesley.

Callahan, K., & Sadeghi, L. (2015). Teacher perceptions of the value of teacher evaluations: New Jersey's ACHIEVE NJ. *International Journal of Educational Leadership Preparation*, *10*(1), 46-59. Retrieved from http://www.ncpeapublications.org/

Danielson, C. (2013). *The framework for teaching: Evaluation instrument*. Princeton, NJ: Danielson Group.

Darling-Hammond, L. (1999). Chapter 2: Reshaping teaching policy, preparation, and practice: Influences of the National Board for Professional Teaching Standards [NBPTS]. *Advances in Program Evaluation*, 25-53.

Darling-Hammond, L. (2014). One piece of the whole: Teacher evaluation as part of a comprehensive system for teaching and learning. *American Educator*, *38*(1), 4-13. Retrieved from http://www.aft.org/our-news/periodicals/american-educator

Davidheiser, S. A. (2013). *Identifying areas for high school teacher development: A study of assessment literacy in the central bucks school district* (Ed.D.). Available from ProQuest Dissertations & Theses Full Text: The Humanities and Social Sciences Collection.

Gareis, C., & Grant, L. (2015). Assessment literacy for teacher candidates: A focused approach. *Teacher Educators' Journal*, *2015*(N/A), 4-21. Retrieved from http://www.tandfonline.com/toc/utte20/current

Gill, B., Bruch, J., & Booker, K. (2013). *Using alternative student growth measures for evaluating teacher performance: What the literature says*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

Gill, B., English, B., Furgeson, J., & McCullough, M. (2014). *Alternative student growth measures for teacher evaluation: Profiles of early-adopting districts*. Regional Educational Laboratory Mid-Atlantic. Retrieved from http://www.relmidatlantic.org/

Gitomer, D. (2011). Road maps for learning and teacher evaluation. *Measurement: Interdisciplinary Research and Perspectives*, *9*(2-3), 146-148. https://doi.org/10.1080/15366367.2011.603616

Groen, M. (2012). NCLB--the educational accountability paradigm in historical perspective. *American Educational History Journal*, *39*(1), 1-14. Retrieved from http://www.infoagepub.com/american-educational-history-journal

Hamilton, P. (2014). *Understanding teachers' concurrent knowledge of assessment literacy and curriculum-based measurement* (M.A.). Available from ProQuest Dissertations & Theses Full Text: The Humanities and Social Sciences Collection.

Hanushek, E., & Haycock, K. (2010). Effective teacher in every classroom. *Education Next*, *10*(3), 46-52. Retrieved from http://educationnext.org/an-effective-teacher-in-every-classroom/

Hendrickson, K. (2012). Learning from Finland: Formative assessment. *The Mathematics Teacher*, *105*(7), 488-502. https://doi.org/10.5951/mathteacher.105.7.0488

Herman, J., & Linn, R. (2013). *On the road to assessing deeper learning: The status of smarter balanced and PARCC assessment consortia*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from http://cresst.org/publications/cresst-publication-3192/

Hu, J. (2015). *Teacher evaluation based on an aspect of classroom practice and on student achievement: A relational analysis between student learning objectives and value-added modeling* (Ph.D.). Available from ProQuest Dissertations & Theses Full Text: The Humanities and Social Sciences Collection.

Johnson, B., & Christensen, L. (2014). *Educational research: Quantitative, qualitative, and mixed approaches* (5th ed.). Los Angeles, CA: Sage.

Lacireno-Paquet, N., Morgan, C., & Mello, D. (2014). *How states use student learning objectives in teacher evaluation systems: A review of state websites*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory North-east & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

Marder, M. (2012). Measuring teacher quality with value-added modeling. *Kappa Delta Pi Record*, *48*(4), 156-161. https://doi.org/10.1080/00228958.2012.733929

Mathis, W. (2011). Is education the key to global economic competitiveness? *Teacher Educator*, *46*(2), 89-97. https://doi.org/10.1080/08878730.2011.555637

Mertler, G. A., & Vannata, R. A. (2002). *Advanced and multivariate statistical measures: Applications and interpretations*. Los Angeles, CA: Pyrczak Publishing.

Mertler, G., & Campbell, C. (2005). *Measuring teachers' knowledge & application of classroom assessment concepts: Development of the Assessment Literacy Inventory*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

New Jersey Department of Education [NJDOE]. (2015). *New Jersey public schools fact sheet*. Retrieved July 1, 2015, from http://www.state.nj.us/education/data/fact.htm

Odo, D. (2015). Improving urban teachers' assessment literacy through synergistic individualized tutoring and self-reflection. *Networks: An Online Journal for Teacher Research*, *17*(2), 1-13. Retrieved from http://journals.library.wisc.edu/index.php/networks

Onosko, J. (2011). Race to the top leaves children and future citizens behind: The devastating effects of centralization, standardization, and high stakes accountability. *Democracy and Education*, *19*(2), 1-11. Retrieved from http://democracyeducationjournal.org/home/vol19/iss2/1

Perry, M. L. (2013). *Teacher and principal assessment literacy* (Ed.D.). Available from ProQuest Dissertations & Theses Full Text: The Humanities and Social Sciences Collection.

Plake, B. S., Impara, J. C., & Fager, J. J. (2005). Assessment competencies of teachers: A National survey. *Educational Measurement: Issues and Practice*, *12*(4), 10-39. https://doi.org/10.1111/j.1745-3992.1993.tb00548.x

Pollins, T. A. (2014). *Student learning objectives: A Rhode Island case study* (Ed.D.). Available from ProQuest Dissertations & Theses Full Text: The Humanities and Social Sciences Collection.

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *Teacher Educator*, *46*(4), 265-273. https://doi.org/10.1080/08878730.2011.605048

Program for International Student Assessment [PISA]. (2015). International student assessment. In *OECD Fact Book* (pp. 160-161).

Rentner, D., & Kober, N. (2014). *Common core state standards in 2014: District implementation of consortia-developed assessments*. Center on Education Policy. Retrieved from http://www.cepdc.org/displayDocument.cfm?DocumentID=442

Riordan, J., Lacireno-Paquet, N., Shakman, K., Bocala, C., & Chang, Q. (2015). *Redesigning teacher evaluation: Lessons from a pilot implementation (REL 2015–030)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

Siegel, M., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education*, *22*(4), 371-391. https://doi.org/10.1007/s10972-011-9231-6

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758-765. https://doi.org/10.1177/003172170208301010

Sullivan, G., & Feinn, R. (2012). Using effect size—Or why the p value is not enough. *J Grad Med Educ.*, *4*(3), 279-282. https://doi.org/10.4300/JGME-D-12-00156.1

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *1*(2), 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Wallace, M., & White, T. (2015). Secondary mathematics preservice teachers' assessment perspectives and practices: An evolutionary portrait. *Mathematics Teacher Education and Development*, *16*(2), 25-45. Retrieved from https://www.merga.net.au/ojs/index.php/mted/index

Appendix A

Teachers' Composite Scores Based on School Assignemnt

Table 2. Teachers' composite scores based on school assignment

Standards	High Achieving Schools	Low Achieving Schools
#1 Choosing Appropriate Assessment Methods	65%	56%
#2 Developing Appropriate Assessment Methods	43%	33%
#3 Administering, Scoring, and Interpreting the Results of Assessment	50%	49%
#4 Using Assessment Results to Make Decision	56%	56%
#5 Developing Valid Grading Procedures	48%	44%
#6 Communicating Assessment Results	56%	47%
#7 Recognizing Unethical or Illegal Practices	55%	50%
Total number of Teachers	52	30

Appendix B

Teachers' Composite Scores Based on Grade Level Taught

Table 3. Teachers composite scores based on grade assignment

Standards	Elementary	Middle	High
#1 Choosing Appropriate Assessment Methods	59%	65%	60%
#2 Developing Appropriate Assessment Methods	42%	49%	31%
#3 Administering, Scoring, and Interpreting the Results of Assessment	53%	50%	47%
#4 Using Assessment Results to Make Decision	67%	63%	45%
#5 Developing Valid Grading Procedures	48%	48%	46%
#6 Communicating Assessment Results	55%	63%	45%
#7 Recognizing Unethical or Illegal Practices	55%	50%	53%
Total number of Teachers	24	23	35

Appendix C

Teachers Composite Scores Based on Years of Experience

Table 4. Teachers composite scores based on years of experience

Standards	0 to 4 years	5 to 10 years	11 to 20 years	21 and more years
#1 Choosing Appropriate Assessment Methods	52%	62%	62%	63%
#2 Developing Appropriate Assessment Methods	44%	39%	38%	39%
#3 Administering, Scoring, and Interpreting the Results of Assessment	48%	43%	53%	51%
#4 Using Assessment Results to Make Decision	60%	54%	55%	60%
#5 Developing Valid Grading Procedures	44%	44%	46%	54%
#6 Communicating Assessment Results	52%	49%	52%	60%
#7 Recognizing Unethical or Illegal Practices	38%	46%	56%	58%
Total	48%	48%	52%	55%
Total number of Teachers	10	16	39	17

Appendix D

Teachers' Composite Scores Based on Tested and Nontested Subjects

Table 5. Teachers composite scores based on tested and nontested subjects

Standards	Tested Teachers	Nontested Teachers
# 1 Choosing Appropriate Assessment Methods	62%	61%
#2 Developing Appropriate Assessment Methods	38%	40%
#3 Administering, Scoring, and Interpreting the Results of Assessment	51%	49%
#4 Using Assessment Results to Make Decision	50%	60%
#5 Developing Valid Grading Procedures	47%	47%
# 6 Communicating Assessment Results	51%	54%
#7 Recognizing Unethical or Illegal Practices	62%	46%
Total	52%	51%
Total number of Teachers	33	49

Appendix E

Teachers' Composite Score Based on Educatioanl Degree

Table 6. Teachers' composite scores based on educational degree

Standards	Master Degree	Bachelor Degree	Ed. Specialist Degree	Doctoral Degree
# 1 Choosing Appropriate Assessment Method	65%	53%	40%	80%
#2 Developing Appropriate Assessment Methods	40%	39%	0%	60%
#3 Administering, Scoring, and Interpreting the Results of Assessment	50%	48%	60%	80%
#4 Using Assessment Results to Make Decision	50%	48%	60%	100%
#5 Developing Valid Grading Procedures	50%	40%	20%	40%
# 6 Communicating Assessment Results	54%	48%	80%	80%
#7 Recognizing Unethical or Illegal Practices	55%	50%	60%	20%
Total	52%	47%	46%	66%
Total number of Teachers	57	23	1	1

Appendix F

T Test Results for Standards in Question 1

Table 7. T test results for standards in question 1

N =82	*F*-for Levene test	Sig for Levene test	*t*-statistic	*df*	Sig 2-tailed	95% Confidence Interval Low	Upper	Mean difference
						Upper		
Standard 1	0.055	0.815	1.648	80	0.103	-0179	0.190	0.09
Standard 2	0.487	0.487	1.956	80	0.054	-0.001	0.209	0.104
Standard 3	0.094	0.760	0.123	80	0.903	-0.102	0.11	0.007
Standard 4	0.141	0.710	0.094	80	0.925	-0.108	0.119	0.005
Standard 5	0.000	0.999	0.938	80	0.341	-0.0543	0.151	0.048
Standard 6	1.029	0.313	1.551	80	0.125	-0.025	0.201	0.088
Standard 7	3.242	0.076	1.089	80	0.280	-0.049	0.168	0.059

Appendix G

ANOVA Tests Results for Standards in Question 2

Table 8. ANOVA test results for standards in question 2

Standards (N=82)	Sum of Squares	*df*	Mean Square	*F*	Significance
Standard 1					
Between Groups	0.048	2	0.024	0.444	0.643
Within Groups	4.255	79	0.054		
Total	4.302	81			
Standard 2					
Between Groups	0.433	2	0.217	4.193	0.019
Within the Groups	4.082	79	0.052		
Total	4.516	81			
Standard 3					
Between Groups	0.038	2	0.019	0.337	-0.715
Within the Groups	4.461	79	0.056		
Total	4.50	81			
Standard 4					
Between Groups	0.858	2	0.429	8.240	0.001
Within Groups	4.112	79	0.052		
Total	4.970	81			
Standard 5					
Between Groups	0.012	2	0.006	0.112	0.895
Within Groups	4.098	79	0.052		
Total	4.098	81			
Standard 6					
Between Groups	0.438	2	0.219	8.735	0.028
Within Groups	4.632	79	0.059		
Total	5.070	81			

Appendix H

Robust Test of Equality of Means for Standard 7 in Question 2

Table 9. Robust test of equality of means for standard 7 in question 2

Standard 7 (*N* =82)				
	Statistics	*df1*	*df2*	Sig.
Welch	0.383	2	50.210	0.684
Brown-Forsythe	0.324	2	74.170	0.725

Appendix I

Tukey HSD Test for Standard 2 in Question 2

Table 10. Tukey HSD test for standard 2 in question 2

Dependent Variable: Standard 2 (N =82)						95% Confidence Interval	
	Grade Level Taught	Grade Level Taught	Mean Differences	Std. Error	Sig.	Lower Bound	Upper Bound
Tukey HSD	Middle	High	0.1726	0.061	0.016	0.0269	0.3184
		Elementary	0.0702	0.066	0.542	-0.0881	0.2287
	High	Middle	-0.1726	0.061	0.016	-0.3184	-0.0269
		Elementary	-0.1023	0.060	0.212	-0.2462	0.0415
	Elementary	Middle	0.0702	0.066	0.542	-0.2287	0.0881
		High	0.1023	0.0600	0.212	-0.0415	0.2462

Appendix J

Tukey Test for Standard 4 in Question 2

Table 11. Tukey HSD test for standard 4 in question 2

Dependent Variable: Standard 4 (N =82)						95% Confidence Interval	
	Grade Level Taught	Grade Level Taught	Mean Differences	Std. Error	Sig.	Lower Bound	Upper Bound
Tukey HSD	Middle	High	0.1890	0.0612	0.008	0.0427	0.3353
		Elementary	-0.0318	0.0665	0.881	-0.1909	0.1271
	High	Middle	-0.1890	0.0612	0.008	-0.3353	-0.0427
		Elementary	-0.2209	0.0604	0.001	-0.3653	-0.0765
	Elementary	Middle	0.0318	0.0665	0.881	-0.1271	0.1909
		High	0.2209	0.0604	0.001	0.0765	0.3653

Appendix K

Tukey Test for Standard 6 in Question 2

Table 12. Tukey HSD test for standard 6 in question 2

Standard 6 N =82						95% Confidence Interval	
	Grade Level Taught	Grade Level Taught	Mean Differences	Std. Error	Sig.	Lower Bound	Upper Bound
Tukey HSD	Middle	High	0.1746	0.0649	0.024	0.0194	0.3299
		Elementary	0.0760	0.0706	0.531	-0.0926	0.2448
	High	Middle	-0.1746	0.0649	0.024	-0.3299	-0.0194
		Elementary	-0.0985	0.0641	0.280	-0.2518	0.0547
	Elementary	Middle	-0.0760	0.0706	0.531	-0.2448	0.0926
		High	0.0985	0.0641	0.280	-0.0547	0.2518

Appendix L

ANOVA Test Results for Standards in Question 3

Table 13. ANOVA test results for standards in question 3

Standards (N=82)	Sum of Squares	*df*	Mean Square	*F*	Significance
Standard 1					
Between Groups	0.103		0.034	0.639	0.592
Within Groups	4.199	78			
Total	4.302	81	0.054		
Standard 2					
Between Groups	0.026	3	0.009	0.149	0.930
Within the Groups	4.490	78	0.058		
Total	4.516	81			
Standard 4					
Between Groups	0.055	3	0.018	0.293	0.831
Within Groups	4.915	78	0.063		
Total	4.970	81			
Standard 5					
Between Groups	0.115	3	0.038	0.747	0.527
Within Groups	3.995	78	0.051		
Total	4.110	81			
Standard 6					
Between Groups	0.119	2	0.040	0.624	0.602
Within Groups	4.951	79	0.063		
Total	5.070	81			
Standard 7					
Between Groups	0.377	3	0.126	2.318	0.082
Within Groups	4.234	78	0.054		
Total	4.611	81			

Appendix M

Robust Test of Equality of Means for Standard 3 in Question 3

Table 14. Robust test of equality of means for standard 3 in question 3

Standard 3 (N =82)				
	Statistics	*df1*	*df2*	Sig.
Welch	0.383	2	50.210	0.684
Brown-Forsythe	0.324	2	74.170	0.725

Appendix N

T Test Results for Standards in Question 4

Table 15. T test results for standards in question 4

N = 82	*F*-for Levene test	Sig for Levene test	*t*-statistic	*df*	Sig 2-tailed	95% Confidence Interval Low Upper	Upper	Mean difference
Standard 1	0.055	0.544	0.114	80	0.910	-0.097	0.109	0.059
Standard 2	1.281	0.261	-0.340	80	0.735	-0.125	0.088	-0.0181
Standard 3	0.055	0.815	1.648	80	0.103	-0.766	0.135	0.0861
Standard 4	0.927	0.339	-1.838	80	0.070	-0.210	-0.008	-0.1010
Standard 5	0.047	0.829	0.065	80	0.948	-0.098	0.105	0.0033
Standard 6	0.000	0.984	-0.597	80	0.552	-0.179	0.078	0.0337
Standard 7	1.857	0.177	3.069	80	0.003	0.055	0.258	0.1569

Appendix O

ANOVA Test Results for Standards in Question 5

Table 16. ANOVA test results for standards in question 5

Standards (N=82)	Sum of Squares	*df*	Mean Square	*F*	Significance
Standard 1					
Between Groups	0.311	3	0.104	2.028	0.117
Within Groups	3.991	78	0.051		
Total	4.302	81			
Standard 2					
Between Groups	0.198	3	0.066	1.195	0.317
Within the Groups	4.317	78	0.055		
Total	4.516	81			
Standard 3					
Between Groups	0.121	3	0.040	0.718	0.544
Within the Groups	4.371	78	0.056		
Total	4.5	81			
Standard 4					
Between Groups	0.197	3	0.066	1.071	0.366
Within Groups	4.774	78	0.061		
Total	4.970	81			
Standard 5					
Between Groups	0.261	3	0.087	1.766	0.161
Within Groups	3.848	78	0.049		
Total	4.110	81			
Standard 6					
Between Groups	0.213	3	0.071	1.143	0.337

Within Groups	4.856	78	0.062		
Total	5.070	81			
Standard 7					
Between Groups	0.144	3	0.048	0.841	0.476
Within Groups	4.467	78	0.057		
Total	4.611	81			

Appendix P

Cronbach Alpha Coefficient

Reliability Statistics	
Cronbach's Alpha	N of Items
.772	35

Figure 6. Cronbach Alpha for ALI