Exploring Teacher Assessment Literacy through the Process of Training Teachers to Write Assessment Items

Heather Peltier Wright

University of South Florida, wrighth@mail.usf.edu

Follow this and additional works at: http://scholarcommons.usf.edu/etd

Part of the Educational Assessment, Evaluation, and Research Commons, Instructional Media Design Commons, and the Teacher Education and Professional Development Commons
Exploring Teacher Assessment Literacy through the Process of Training Teachers to Write Assessment Items

by

Heather Peltier Wright

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Curriculum and Instruction with an emphasis in Instructional Technology Department of Educational and Psychological Studies College of Education University of South Florida

Major Professor: Yiping Lou, Ph.D.
Eun Sook Kim, Ph.D.
Yi-Hsin Chen, Ph.D.
James Hatten, Ph.D.

Date of Approval:
March 1, 2017

Keywords: Assessment literacy, professional development, item development, teacher training

Copyright © 2017, Heather Peltier Wright
ACKNOWLEDGMENTS

I would like to thank all of my dissertation committee members, who spent so many hours reading, re-reading, and providing invaluable feedback to me on my work. You made the impossible feel possible! Dr. Yiping Lou, thank you for your mentorship, guidance, and support over the past five years. Dr. Eun Sook Kim and Dr. Yi-Hsin Chen, I have learned so much from your courses and from your guidance on this study. Dr. James Hatten, I appreciate your assistance and support. Also a special thank you to Dr. Robert Dedrick for grounding me in the foundations in measurement, without which this project would not have been possible, and to Dr. Jennifer Wolgemuth, for helping a reluctant learner with post-positivist leanings better appreciate qualitative research methods. All of you helped me set high standards and gave me the confidence I needed to succeed at this work, and I will be forever grateful.
# TABLE OF CONTENTS

List of Tables .................................................................................................................................. iv

List of Figures ................................................................................................................................... vi

Abstract .......................................................................................................................................... vii

Chapter One: Introduction ............................................................................................................... 1
    Study Context & Background ............................................................................................. 1
    Historical Context ......................................................................................................... 2
    Assessment Development Project ........................................................................... 4
    Statement of the Problem ............................................................................................ 5
    Purpose ............................................................................................................................ 7
    Research Questions ............................................................................................................. 8
    Conceptual Framework ....................................................................................................... 8
        Teacher-Level Characteristics................................................................................. 9
        Assessment Professional Development .......................................................... 12
    Significance ....................................................................................................................... 15
    Delimitations ..................................................................................................................... 16
    Limitations ........................................................................................................................ 16
    Definitions of Terms ......................................................................................................... 18

Chapter Two: Review of the Literature ......................................................................................... 20
    Literature Search Method.................................................................................................. 20
    Challenges with Existing Research................................................................................... 20
    Teacher Assessment Literacy and Assessment Quality .................................................... 24
    Impact of Assessment Policies.......................................................................................... 27
    Impact of Assessments on Students .................................................................................. 29
    Types of Professional Development .................................................................................. 31
    Assessment-Related Professional Development Opportunities and Challenges .............. 33
    Feedback and Assessment Item Development .......................................................... 35
    Media Effects .................................................................................................................... 38
    Summary ........................................................................................................................... 40

Chapter Three: Method .................................................................................................................. 42
    Research Paradigm ............................................................................................................ 43
    Research Design ................................................................................................................. 43
        Training Mode .............................................................................................................. 44
    Participants ......................................................................................................................... 45
    Data Sources ....................................................................................................................... 53
Chapter Four: Results ....................................................................................................................62
  Research Question One ..................................................................................................... 62
    Outcome Variable: Item Discrimination/Slope Estimate ........................................... 63
  Model 1: Unconditional baseline model for item discrimination.............................. 64
  Model 1a: Between-writer differences in item discrimination based on years of teaching experience ................................................................. 65
  Model 1b: Between-writer differences in item discrimination based on prior assessment experience beyond classroom level ........................................ 66
  Model 1c: Between-writer differences in item discrimination based on prior assessment experience beyond school level ..................................... 66
  Model 1d: Between-writer differences in item discrimination based on graduate degree ........................................................................................................... 67
  Model 1e: Between-writer differences in item discrimination based on National Board certification ................................................................................. 68
    Outcome Variable: Number of Reviews by Item .................................................... 68
  Model 1f: Unconditional baseline model for number of reviews by item ....................... 70
  Model 1g: Between-writer differences in number of reviews by item based on years of experience ................................................................. 70
  Model 1h: Between-writer differences in number of reviews by item based on prior assessment experience only at classroom level ......................... 71
  Model 1i: Between-writer differences in number of reviews by item based on prior assessment experience beyond the school level ......................... 72
  Model 1j: Between-writer differences in number of reviews by item based on graduate degree ................................................................. 73
  Model 1k: Between-writer differences in number of reviews by item based on National Board certification ................................................................. 73

Research Question Two .................................................................................................... 74
  Outcome Variable: Number of Reviews by Item .................................................... 74
  Model 2a: Between-writer differences in number of reviews by item based on participant role ................................................................. 74
  Model 2b: Between-writer differences in number of reviews by item based on participant role and prior assessment development experience ... 76
    Outcome Variable: Proportion of Accepted Items .................................................... 77

Research Question Three .................................................................................................. 79
  Outcome Variable: Number of Reviews by Item .................................................... 79
  Model 3a: Between-writer differences in number of reviews by item based on training type.
LIST OF TABLES

Table 1: Training Mode by Sample ................................................................. 44
Table 2: Comparison of Training Modes for Assessment Development Project........ 46
Table 3: Participant Demographics for Each Sample ....................................... 48
Table 4: Prior Assessment Experience by Sample ........................................... 51
Table 5: Participant Affiliation by Sample ..................................................... 51
Table 6: Training Evaluation Questionnaire Participant Demographics .............. 54
Table 7: Correlation Matrix between Teacher Variables ................................. 58
Table 8: Descriptive Statistics for Number of Item Reviews and Discrimination .... 63
Table 9: Parameter Estimates for the Models Examining the Relationship Between Item Writer Characteristics and Test Item Discrimination ............. 69
Table 10: Parameter Estimates for the Models Examining the Relationship Between Item Writer Characteristics and the Number of Reviews for Each Item Written by the Item Writer ......................................................... 75
Table 11: Multiple Regression Model Summary of Proportion of Accepted Items as an Outcome Variable ............................................................ 78
Table 12: Coefficients for Multiple Regression Models of Proportion of Accepted Items as an Outcome Variable ......................................................... 78
Table 13: Parameter Estimates for the Model Examining the Relationship Between Mode of Training and the Number of Reviews by Item for Each Item Written by the Item Writer ......................................................... 81
Table 14: Parameter Estimates for the 2 Models Examining the Relationship Between Mode of Training and the Item Discrimination for Each Item Written by the Item Writer ......................................................... 81
Table 15: Linear Regression Model Summary of Proportion of Accepted Items as an Outcome Variable, Using Mode of Training as a Predictor .......... 82
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>Coefficients for Linear Regression Model of Proportion of Accepted Items as an Outcome Variable, Using Mode of Training as a Predictor</td>
<td>82</td>
</tr>
<tr>
<td>17</td>
<td>Percent of Responses Related to Each Theme</td>
<td>84</td>
</tr>
<tr>
<td>18</td>
<td>Chi-square Test for Collaboration with Other Educators</td>
<td>86</td>
</tr>
<tr>
<td>19</td>
<td>Results of Chi-square Test Post-hoc Analysis Using Standardized Residuals, Collaboration</td>
<td>87</td>
</tr>
<tr>
<td>20</td>
<td>Chi-square Test for Cognitive Complexity and Rigor</td>
<td>88</td>
</tr>
<tr>
<td>21</td>
<td>Results of Chi-square Test Post-hoc Analysis Using Standardized Residuals, Cognitive Complexity and Rigor</td>
<td>89</td>
</tr>
<tr>
<td>22</td>
<td>Chi-square Test for Alignment of Assessment and Instruction</td>
<td>90</td>
</tr>
<tr>
<td>23</td>
<td>Chi-square Test for Bias and Sensitivity</td>
<td>91</td>
</tr>
<tr>
<td>24</td>
<td>Results of Chi-square Test Post-hoc Analysis Using Standardized Residuals, Bias and Sensitivity</td>
<td>92</td>
</tr>
<tr>
<td>25</td>
<td>Chi-square Test for Rubric Development</td>
<td>95</td>
</tr>
<tr>
<td>26</td>
<td>Results of Chi-square Test Post-hoc Analysis Using Standardized Residuals, Rubrics</td>
<td>96</td>
</tr>
<tr>
<td>27</td>
<td>Descriptive Statistics for Evaluation Survey Results</td>
<td>98</td>
</tr>
<tr>
<td>28</td>
<td>ANOVA for Total Score on Evaluation Survey, by Mode of Training</td>
<td>98</td>
</tr>
<tr>
<td>29</td>
<td>ANOVA for Response to Each Item, by Mode of Training</td>
<td>99</td>
</tr>
<tr>
<td>30</td>
<td>Results of 2PL IRT Analysis for Items with Field Test Data</td>
<td>144</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1: Logic model of teacher assessment literacy professional development...........10
Figure 2: Relationship between assessment terms used throughout this study..............19
Figure 3: Relationship between each sample..............................................................49
Figure 4: Student population race and ethnicity.........................................................55
Figure 5: Evidence and analysis by research question................................................57
ABSTRACT

The purpose of this study was to examine the process and impact of assessment training content and delivery mode on the quality of assessment items developed by the teachers in a two-year assessment development project. Teacher characteristics were examined as potential moderating factors. Four types of delivery mode were employed in the project: synchronous online, asynchronous online, in-person workshop, and blended (a combination of online and in-person training). The quality of assessment items developed by participating teachers was measured via: 1) item acceptance rate, 2) number of item reviews (as an indicator of how many times accepted items were rejected before being approved), and 3) psychometric properties of the items (item difficulty and item discrimination) in the field test data.

A teacher perception survey with quantitative and qualitative data was used to explore teacher perception of the training across the four modes and the anticipated impact of the project participation the teachers expected on their classroom assessment practices.

Multilevel modeling and multiple regression were used to examine the quality of items developed by participants, while constant comparative analysis, a chi-square test, and ANOVA were employed to analyze participants’ responses to a participation survey.

No pre-existing teacher variables were found to have a significant impact on the item discrimination values, though prior assessment development experience beyond that of the classroom level was found to have a significant relationship with the number of reviews per item. After controlling for prior assessment development experience, participant role was found to
have a significant \( (p < .01) \) impact on the number of reviews per item. Items written by participants who served as both item writers and reviewers had a significantly lower number of reviews per item, meaning their items were rejected less frequently than items written by participants who served as item writers only. No differences in item quality were found based on the mode of training in which item writers participated.

Responses to the training evaluation survey differed significantly by mode of training at \( p < .001 \). The in-person trained group had the lowest total rating, followed by the online asynchronous group, while the online synchronous group had the highest overall rating of the training. Participant responses to open-ended questions also differed significantly by mode of training.
CHAPTER ONE:
INTRODUCTION

Study Context & Background

The quality and appropriateness of assessments used by teachers is a significant part of what happens in the classroom. Although much attention has been given to understanding standardized, summative assessments administered to students, largely because of the high stakes these hold for teachers, schools, and districts, a much smaller body of research has been conducted on understanding teacher-developed classroom assessments (Gotch, 2012). However, these represent the majority of assessments given to students. While students typically only take one standardized assessment per course each year, in the United States, teacher-developed classroom assessments are often given on a monthly, weekly, or even daily basis. In addition to being a substantial determiner of student grades and a valuable communication tool between school and home, classroom assessments are used by teachers for instructional planning (Brookhart, 2004; Gullickson, 1984; Marsh, Pane & Hamilton, 2006). Marso and Pigge (1993) reported that students will have taken between 400 and 1,000 classroom assessments during their K-12 educational careers, while Stiggins (1999) estimated that teachers spend 30 to 50 percent of their professional time on assessment-related activities. There is also evidence that grading and assessment practices can impact students’ self-identity and self-efficacy (Brookhart & DeVoge 1999; Thomas & Oldfather, 1997). Given these facts, the dearth of research into both the nature
of teacher-developed classroom assessments and teachers’ acquisition of assessment competency represents a significant research gap.

Prior explorations of teacher assessment literacy competence reveal concerns about teachers’ ability to develop or select classroom assessments that measure higher order thinking skills and that are aligned with state standards. Wright, Foran, Holmes, and Lou (2016) found that only 1.3 percent of the questions on the studied sample of classroom assessments measured the highest levels of Bloom’s cognitive process dimensions, Evaluate and Create, and only 6.1 percent of questions measured level 3 of Webb’s Depth of Knowledge. While the majority of mathematics items examined (94 percent) were aligned to state content standards, 41.2 percent of English language arts items were not aligned to state content standards. The most common problems found with the misaligned English language arts items were either alignment to a standard that was outside of the scope of the course description or teacher re-creation of a standard within the course description, to lower its cognitive demand. For example, one teacher identified the standard to which a question was aligned as “LACC.910.RL.2.4 Recognize the author’s tone.” The actual wording of standard LACC.910.RL.2.4 asks students to “analyze the cumulative impact of specific word choices on meaning and tone” (Common Core State Standards Initiative [CCSSI], 2014), a more cognitively demanding task than the teacher’s revised version. This prior research highlights the need for additional teacher training on assessment literacy and research into teachers’ acquisition of assessment literacy.

Historical Context

Assessment has become an important part of the political landscape, and any research into this area is situated within a political context. This study took place in Florida, which, over the past decade, has made several efforts to implement teacher evaluation programs based on student assessments. In 2006, the Florida legislature approved the Special Teachers are
Rewarded (STAR) performance pay program for teachers, which provided $147.5 million in funding for teachers whose students showed learning gains on the statewide assessment of reading and mathematics, the Florida Comprehensive Assessment Test (FCAT) or, for courses and grade levels not assessed by the FCAT, on optional district-developed assessments (U.S. Department of Education, 2007). Because of a lack of political support, the STAR program was replaced in March 2007 by the Merit Award Program (MAP), which offered school districts more flexibility in determining both the award size and proportion of teachers awarded, but which still required the award criteria to be based on student performance on assessments. MAP differed from STAR, in that it required districts to use content area assessments for subjects other than those assessed by the FCAT, whereas STAR had given districts the option to use district-developed assessments, state-developed assessments, or a combination thereof (U.S. Department of Education, 2007). In order to alleviate pressure on school districts to rapidly develop assessments for courses not covered by the FCAT, the Florida Department of Education collected a bank of over 550 assessments from school districts (primarily from Hillsborough County); however, because of time constraints, districts had minimal time to review or evaluate the quality of these assessments. One school district developed 313 assessments over the course of three weeks, 200 of which had errors, which led to multiple teacher appeals of compensation decisions (Hobbs, 2007). Due to a lack of funding, by 2011, only three school districts participated in the Merit Award Program, and it was repealed by House Bill 7087 (Florida Department of Education, 2011).

In 2010, the state of Florida was awarded a Race to the Top grant by the federal government, which again became the impetus for teacher evaluation reforms linked to student assessment results. In 2011, Florida Senate Bill 736 was signed into law, codified in two
The first of these two statutes, 1012.34, deals with teacher evaluation systems and was amended to require that at least 50 percent of a classroom teacher’s evaluation be based upon student learning growth, as assessed through statewide assessments, or, for courses or grade levels not assessed through statewide assessments, as assessed through district-implemented assessments (2013). The second statute, 1008.22 required school districts to administer an end of course assessment for every course offered in the district, and to develop or acquire assessments to measure student mastery of the course content for any courses not covered by statewide assessments (2013). Thus, between the years of 2011 and 2014, school districts in Florida were tasked with not only developing a large number of assessments in a very limited time frame, but with developing assessments that could, with some legal defensibility, be used as a significant portion of each teacher’s evaluation.

Assessment Development Project

In July 2013, a collaborative of Florida public school districts was awarded the first of two grants by the Florida Department of Education that enabled the collaborative to develop assessments for a total of over 200 secondary courses without existing state-developed assessments. More than 400 teachers across the state of Florida participated in this project over the course of two years; they received professional development, either online or in person, on principles of assessment development, and used an online software program to author their own items and to review others’ items.

Test items developed by the project went through three reviews: a level one content review, a level two bias and sensitivity review, and a level three proofreading and formatting review. The item writer and reviewer checklist used to evaluate item quality is found in Appendix A.
All item writers were required to be licensed in the content area for which they developed items, or in a related area. No minimum years of experience were required for item writers. In contrast, teachers who served as reviewers at any of the three levels were required to have at least three years of experience in teaching. Level one content reviewers were required to be licensed in the content area for which they developed items, or in a related area, while level two bias and sensitivity reviewers were required to have either ESOL or ESE licensure, or both. Finally, level three proofreaders were selected by the project manager from the pool of level two reviewer participants, as those who demonstrated the most meticulous, scrupulous review practices. The group of level three reviewers was the smallest of all participant groups, since this tier of review was composed of the most detail-oriented reviewers who also had a sufficient grasp of the editing and proofreading process to ensure high-quality items from a grammar and formatting standpoint.

Statement of the Problem

Prior studies of teacher assessments demonstrate that there are concerns with teachers’ assessment literacy in terms of both the use (or misuse) and quality of teacher-created and teacher-selected assessments (Benson, 1997; Brookhart, 2004; Fleming & Chambers, 1983; McMorris & Boothroyd, 1993; Wright et al., 2016). An equally troublesome concern, however, is the dearth of research into these assessments. Gotch (2012) found only 36 articles published in peer-reviewed journals over a 21-year period that studied teacher creation of assessments or use of assessment results. Assessment has become a large industry within education, with an estimated $1.7 billion spent by states on student assessment, and with only six vendors accounting for 89 percent of these assessment costs (Chingos, 2012). Comparatively little funding has been spent on the development of teacher proficiency in student assessment, in contrast. While there has been a significant emphasis on research into these standardized
assessments, teacher-developed classroom assessments, which represent the majority of assessments taken by students in any given year, have received relatively little attention by researchers, particularly given the fact that these assessments are very high-stakes for students.

In Florida, where this study takes place, standardized assessments are high-stakes for students, with certain assessments serving as promotion or graduation requirements (Florida Department of Education, 2014a; O’Connor, 2014). Classroom assessments, as a major determinant of student grades, also hold high stakes for students, though these stakes are given much less attention than that accorded to standardized assessments. According to the Florida Department of Education (2014b), a grade point average of 2.0 on a 4.0 scale is a graduation requirement. Given these high stakes for students, the lack of research into teachers’ assessment practices and the quality of teacher-developed classroom assessments is particularly concerning.

Despite holding high stakes for students, classroom assessments receive very little scrutiny. Most teachers are primarily, if not entirely, responsible for developing their own classroom assessments, which receive very little, if any, oversight or feedback by their teaching peers or administrators (Black & Wiliam, 1998a; Volante & Cherubini, 2011). In Florida, some contracts between the teachers’ union and the school district include clauses of grading autonomy for teachers, allowing administrators to make changes to grades only in the case of mis-recorded, transposed, or poorly documented grades (Polk Education Association, 2013). In districts with such agreements, administrative oversight for teacher grading practices is not allowed, further isolating teachers within their own practice.

Additionally, in a recent survey of new graduates from teacher preparation programs in Florida, teachers reported that, of the six Florida Educator Accomplished Practices, they felt least prepared in the area of assessment (Milton et al., 2013). Principals agreed, rating teachers lowest
in this area, with 11.6 percent of teachers judged by administrators as either unsatisfactory or needs improvement, when it comes to assessments (Milton et al., 2013). These findings highlight the need to provide professional development for teachers on assessment literacy. There is a further need for the examination of the content, format, and process of professional development designed to improve the quality of assessment developed by teachers.

Purpose

The purpose of this study was to examine the process and impact of assessment training content and delivery mode on the quality of assessment items developed by the teachers in a two-year assessment development project. Four types of delivery mode were employed in the project: synchronous online, asynchronous online, in-person workshop, and blended (a combination of online and in-person training). While the basic training goals were the same across the training modes, the training materials and process of training varied somewhat across the delivery mode to make better use of the affordances available for each specific mode of delivery.

The quality of assessment items developed by participating teachers was measured via: 1) item acceptance rate, 2) number of item reviews (as an indicator of how many times accepted items were rejected before being approved), and 3) psychometric properties of the items (item difficulty and item discrimination) in the field test data.

A teacher perception survey with quantitative and qualitative data was used to explore teacher perception of the training across the four modes and the anticipated impact of the project participation the teachers expected on their classroom assessment practices.

In addition, teacher level features such as years of teaching experience, type of degree, National Board certification, and prior assessment experience were examined as potential moderating factors.
Research Questions

This study sought to answer the following questions:

Q1: Is there any significant relationship between the following teacher-level factors:
   - years of experience
   - prior assessment development experience
   - type of degree, and
   - National Board certification
   and the participating teachers’ ability to write high-quality assessment items?

Q2: What impact, if any, does the role selected by teachers (item writers vs. item writers/reviewers) have on teachers’ abilities to write high quality assessment items?

Q3: What impact, if any, does the mode of training have on the quality of items written by teachers?

Q4: What do teachers report as the effects of participation in an assessment development project, and are there differences in these effects by mode of training?

Q5: Is there a relationship between teachers’ perceptions of the assessment literacy training and the mode of training (in-person vs. online) they received?

Conceptual Framework

A logic model for this study is shown in Figure 1. The input prior to the project was teachers’ prior assessment literacy as measured by their prior classroom assessment quality, which is a result of teacher level characteristics including years of teaching experience, National Board Certification, prior assessment development experiences, and type of degree that they have. These classroom assessments, submitted by teachers as part of their application to the project, were studied by Wright, Foran, Holmes and Lou (2016), using an item quality rubric that examined style and format considerations such as spelling and grammar and clarity of wording.
as well as content considerations like alignment to standard and cognitive demand. Findings varied by content area; the majority of items written to assess English language arts content were not aligned to content standards, though the majority of items written to assess mathematics content were aligned to state content standards. The majority of items were found to measure low cognitive complexity levels. Over 12 percent of items were found to have style or formatting issues such as grammar, spelling, or punctuation problems.

During the assessment collaborative project, the participating teachers received assessment training first in four different modes: online asynchronous, online synchronous, face-to-face, and blended via self-selection. After the training, the participants either wrote items and/or wrote and reviewed items, receiving and/or providing feedback on the items written, and make revisions based on the feedback. The output or results of the project in terms of teacher assessment literacy was measured by the number of times an item is reviewed before acceptance, the rate of item acceptance, and the psychometric quality of the items field tested at the teacher level.

**Teacher-Level Characteristics**

Kershaw (1993) hypothesized that teacher characteristics such as age, gender, education level, teaching experience, related work experience, certification route, program area, and type of school impacted their attitude toward assessment, which in turn impacted their use of student assessment data to inform educational decision making. His study, however, found no significant or substantial relationships between these teacher characteristics and teachers’ use of assessment data for educational decision making.
Figure 1. Logic model of teacher assessment literacy professional development
Williams and Rink (2003) also examined teacher-level factors, but in relation to accuracy of scoring students’ responses to physical education performance task assessments. Gender and teacher training on how to score the assessment were hypothesized as factors that could have a relationship with the accuracy of their scores. However, no significant relationships were found between either of these factors and teachers’ accuracy of scoring. Similarly, Veldhuis and van den Heuvel-Panhuizen (2014) in a study of primary school teachers’ assessment profiles in mathematics education found that including teacher age, gender, grade taught, or textbook use did not improve the fit of their model of these teachers’ assessment practices.

Mazzie (2008) examined both student- and teacher-level factors in relation to performance on science assessments. While student-level factors were found to have a significant relationship with their scores on the assessment, the teacher-level variables of teacher scores on the Teacher Quality Research Test of Assessment Literacy Skills and teacher participation in a Teacher Quality Research project on classroom assessment had no significant relationship with student achievement.

Different from the above findings, Sato, Chung, and Darling-Hammond’s (2008) research found significant differences between National Board certified and non-National Board certified teachers on the quality of formative assessment rubrics that teachers developed. However, the study did not examine other teacher-level variables.

Although this was the only study that could be located that explored specifically the variables of National Board certification and assessment literacy, other researchers have explored the relationship between National Board certification and student performance on standardized assessments, and have found that National Board certified teachers receive higher scores on value added models used to evaluate their students’ performance (Cowan & Goldhaber, 2016).
Jarr (2012) found no significant relationships between years of teaching experience and performance on the Assessment Results Interpretation and Use total score, but did discover that teacher participants who reported having previously participated in some form of professional development focused on the use of assessments to guide instruction had higher levels of assessment self-confidence. However, when compared to the outcome of performance in locating, interpreting, and using data, as indicated by the survey developed by the researchers, these teachers did not demonstrate higher levels of competence, indicating that their previous professional development had increased their self-confidence though not necessarily their performance. The reverse was found to be true for participants who had reportedly taken assessment-related college courses. These participants expressed the same degree of self-confidence as those who had not taken similar courses, but their performance on the Assessment Results Interpretation and Use Survey was stronger than that of their peers without assessment-related college coursework.

Considering the inconsistent findings in the prior studies and that the outcome variables examined in this study differ from those in earlier studies, teacher-level factors such as years of experience, prior assessment development experience, National Board certification, and type of degree will be examined to determine whether they have a significant impact on teacher assessment literacy in terms of item quality.

Assessment Professional Development

While teacher professional development workshops have traditionally followed an expert transmission model, this grant project utilized a combination of expert transmission and peer learning.
Teacher participants self-selected into one of four types of training: online asynchronous (self-paced), online synchronous (two 90-minute webinar sessions with individual activities submitted to the instructor), in-person (6 hours), or a combination of two or more of the training options (blended learning). Participants were encouraged to attend multiple training sessions, as they felt necessary, creating the fourth mode of training: blended. Based on prior research, mode of training administration itself is unlikely to lead to significant differences in the performance of teachers (Fishman et al., 2013; Fisher et al., 2010; Powell et al., 2010). However, the time spent in training for this project was longer for participants who participated in the in-person and self-paced online trainings than for the participants who opted for the webinar training, while the time spent in training was longest for participants who chose to attend both the in-person training as well as one of the two online options. Earlier research by Koh (2011) found the variable of length of training to have a significant impact on teachers’ acquisition of assessment literacy, so one might reasonably expect to see differences in the outcomes between participants who participated in the blended mode, when compared to those who selected other modes of training.

In all four modes of training, teachers were given a combination of direct instruction as well as feedback by an expert or a peer (depending on mode of training) on their own classroom assessments. After completing the formal training portion of the project, teachers were placed into peer groups, where, depending on role as writer or reviewer (or both), either received feedback on assessment items that they developed or provided feedback on assessment items developed by others. In prior research by Li (2007), learners reported cognitive and learning gains through the process of giving and receiving peer feedback. There is a substantial body of work that supports the theory that engaging in the process of providing feedback utilizes the cognitive processes of analysis, problem representation, solution development, solution

DeLuca and colleagues (2012) identified three components of successful contemporary professional development related to assessment literacy: active, collaborative learning; ongoing, contextualized learning; and process-based learning and reflective practice. Malone (2013) found that the second element, contextualized learning, was valued by teachers. Research questions one and two both explore the first two elements, though question one deals with the direct instruction component, while question two examines the teacher’s role as recipient of peer feedback or as both provider and recipient of peer feedback. The third element, reflective practice, is measured by question four, which examines the impact of project participation on teachers’ professional practices, as expected by the teachers.

While the project allowed for teachers to self-select into the role of writer or reviewer, teachers were also allowed to serve in both roles, with the only restriction being that they could not review their own items. The majority of the teachers opted to participate as both writers and reviewers, and thus served as both recipients and providers of peer feedback. Prior studies have shown significant correlations between learning outcomes and providing feedback to others (Cho & Cho, 2011; Snowball & Mostert, 2013; Cevik, 2015), while receiving feedback has more inconsistent results (Chen et al., 2009; Cho & MacArthur, 2010; Cho & Cho, 2011; Pelgrim, Kramer, Mokkink, & Van der Vleuten, 2013; Cevik et al., 2015). This study also explored the relationship between the role selected by the item writer (writer only vs. writer and reviewer) and the quality of the items written by the writer as measured by the number of reviews for each of their items and the proportion of accepted items.
Significance

This study examined the process and impact of training teachers to develop high-quality assessment items. Teacher-level variables were examined, to determine whether any of these were associated with the development of higher- or lower-quality assessment items. Earlier studies exploring the relationship between variables such as years of experience, type of degree, and subject or grade level taught on teachers’ assessment literacy used questionnaires to measure assessment literacy (Alkharusi, 2011; Hoover, 2009; Zhang & Burry-Stock, 2003). In contrast, this study focused on empirical outcomes: the psychometric properties of assessment items developed by participating teachers. Additionally, this study built upon the work of Sato, Chung, and Darling-Hammond (2008) to examine the variable of National Board certification, a variable seldom examined in relation to teachers’ assessment literacy. Sato, Chung, and Darling-Hammond (2008) found significant increases in scores on formative assessment rubrics for teachers pursuing National Board certification, compared to their counterparts who were not pursuing the certification. This study explored whether this variable is significantly related to test item performance and item acceptance rate.

Koh (2011) conducted a study to explore the relationship between the type of professional development delivered and teachers’ assessment literacy; however, the variable examined was length, rather than mode, of intervention. This study explores a combination of length, mode, and instructional activities within four different intervention types. This study also examined expected effects of participation in the project on teachers’ classroom assessment practices, as self-reported by teacher participants. The impact of this study includes recommendations for training teachers on best practices in assessment, as well as a greater understanding of the factors involved in teacher assessment literacy development. Because
assessment is, or ought to be, closely linked to pedagogical practices, research into assessment inevitably returns valuable information about instruction as well.

**Delimitations**

This research is limited to studying the quality of assessment items created by teachers who participated in the assessment literacy collaborative project, which took place between July 2013 and June 2015, the Race to the Top project period. This sets its own restrictions on participants:

- All participants were certified teachers in the state of Florida.
- Participants were limited to teachers of secondary (grades 6-12) subject areas.
- Participants were limited to teachers of the following content areas: mathematics, science, English language arts, Junior ROTC, social studies, career/technical education, and world languages.

Additionally, for research questions one and three, since the examination of item quality must include items that were used on assessments to measure student learning, and since the item bank is larger than the number of field-tested items, this further restricted the number of teachers whose work was studied. Of the 419 teachers who participated in the project, item acceptance data was available for 32.7 percent (137 teachers), while item field test data was available for 14.8 percent (62 teachers).

**Limitations**

This study had several limitations. While the grant-funded project followed the principles of educational design research, and modifications to the training were made throughout the lifespan of the project, based on input from earlier teacher participants, the analysis of project effects were done post-hoc. Because of this, modifications that could possibly have been made to improve the effectiveness of the training will be omitted. For example, the results of the
training evaluation survey were not used to modify the training received by the teachers participating in the cohort studied by this project, although they were used to modify the online training delivered to later cohorts of teachers after the conclusion of the grant project.

Another limitation of this research study pertains to the training variable itself. There are confounding factors within this variable, since teachers who participated in the blended mode of training received a longer training than did those who participated in either online mode or in the face-to-face training. Because of this, had there been a relationship between training type and item quality, it would have been difficult to tell whether the effect was due to the mode of training, the length of training, or both mode and length of training.

Generalizability is limited by some shortcomings of the data collected. While data on years of teaching experience, area of licensure, type of degree, and prior assessment experience were collected by the assessment collaborative as part of the demographic data for teacher participants, demographic data such as race or ethnicity was not collected. Thus, the portion of this research that focuses on an analysis of secondary data collected by the grant project was limited to the demographic information that was collected. Teacher participants in the project were volunteers who were willing to spend time outside of school being trained on item writing and reviewing items, so the pool of study participants was composed of those who were likely already interested in learning more about assessment. Because of this, the findings may not generalize to a wider group of teachers with less interest in assessment.

Another limitation of the study also pertains to the type of data collected. The student performance data on items were limited to students from a single district, and the number of student responses available for each item ranges from 190 to 4,952. These limitations impact the generalizability of the results.
Of the 570 items, written by 62 teachers for which student performance data is available, only 20 of the items, written by 8 teachers, were from teachers who participated as writers only. Because of the small sample size and unbalanced sample, a comparison of item performance data for items written by item writers versus those written by item writers/reviewers was impossible. Thus, research question two was only able to be answered by an analysis of the item acceptance rate and number of reviews for each item, not by student performance data by item.

A further limitation of the study pertains to the teacher perception survey data. While teachers can report on ways they believe participation changed their own practices, these conclusions relied upon self-reports, which can be biased.

Definitions of Terms

There are several terms that will be used throughout this study, which I define here as they are used.

Assessment literacy – In this study, I used the definition of assessment literacy presented in the Standards for Teacher Competence in Educational Assessment of Students (American Federation of Teachers et al., 1990). Assessment literacy encompasses teachers’ abilities to select and develop appropriate assessment methods for instructional decisions; to administer, score, and interpret results of teacher-created and externally-produced assessments; to use these assessment results to make decisions about students, curriculum, and pedagogy; to develop valid pupil grading procedures; to communicate assessment results to multiple stakeholders; and to recognize unethical, illegal, or inappropriate assessment methods (American Federation of Teachers, et al., 1990).

Measurement literacy – The definition of measurement literacy, as used here, was that proposed by Gotch (2012): the ability to understand and work with the results of standardized tests.
**Teacher developed classroom assessment** – This referred to any assessment, either formative or summative, selected or created by the teacher for use in his or her classroom. Excluded from this definition were assessments mandated by authorities outside of the classroom teacher, such as district- or state-mandated assessments.

**Formative assessment** – The definition of formative assessment, as used here, was any assessment whose primary purpose is to gather feedback in order to guide changes to teaching and learning practices. Formative assessments were used synonymously with assessments for learning, in this study.

**Summative assessment** – Summative assessment referred to any assessment whose primary goal is to measure the level of student proficiency or success at the end of an instructional unit. While summative assessment data can be used formatively, the differing primary purpose of assessment serves as the differentiator between these two categories. These types of assessments were also referred to as assessments of learning.

Figure 2 illustrates the relationship between the different assessment types discussed in this study. These terms, as will be discussed in the following section, are critical to any discussion of assessment literacy; often, the different terminology used to describe the same phenomena, or its reverse, the use of same terminology to describe different phenomena, present a challenge to aggregating and sharing findings from researchers.

<table>
<thead>
<tr>
<th><strong>Classroom Assessments</strong></th>
<th><strong>Accountability Assessments</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonyms: Teacher-developed assessments, Teacher-created assessments</td>
<td>State assessments, Standardized assessments</td>
</tr>
<tr>
<td><strong>Formative</strong></td>
<td><strong>Summative</strong></td>
</tr>
<tr>
<td>Assessments for learning</td>
<td>Assessments of learning</td>
</tr>
</tbody>
</table>

*Figure 2. Relationship between assessment terms used throughout this study.*
CHAPTER TWO:

REVIEW OF THE LITERATURE

Literature Search Method

The following literature review covers the past three decades of research into teacher assessment literacy, assessment quality, assessment-related professional development, and the impact of providing and receiving feedback on learning. In order to uncover studies that were candidates for inclusion, I searched research databases such as ERIC, Ebsco, Education Full Text, and JSTOR. By using search terms such as “assessment literacy,” “assessment quality,” and “teacher assessment,” I was able to discover a limited number of articles. Challenges with locating studies will be discussed in more detail below. After searching with the above keywords, I began using synonyms such as “test” or “measurement” in place of “assessment.” These search terms uncovered primarily studies published within peer-reviewed journals. To broaden my search to include other sources such as conference presentations, I used a snowball sampling method, searching for articles that cited any that I had uncovered, and searching through the references section of studies to discover additional works that were relevant to this topic.

Challenges with Existing Research

Current studies of classroom assessment present several challenges for researchers attempting to synthesize information across studies. First and foremost of these is difficulty locating research on assessment practices due to inconsistent terminology. While the majority of studies on assessment attempt to distinguish between types and purposes of assessments, the nomenclature used to serve this purpose is inconsistent. Although the most frequent terms used
are formative and summative assessment, there is a significant body of work that distinguishes between assessment for learning and assessment of learning, while still others draw a distinction between classroom assessment and accountability assessment (Black and Wiliam, 1998b, DeLuca, et al., 2012, Stiggins, 2006, Popham, 2009). In addition to making these studies difficult to locate, the distinctions, sometimes subtle and sometimes pronounced, between these types of assessments (classroom vs. formative vs. assessment for learning) also create a challenge in aggregating research.

A second challenge with current research is that many studies do not report information necessary to make a determination regarding the internal and external validity of the study. Failures to describe the demographics, and sometimes even number, of study participants limits the generalizability of results (Koh, 2011; Malone 2013; Mertler, 2005). Study designs that introduce more than one variable (such as both length of time spent in professional development and type of professional development) reduce the internal validity of the study, since it is unclear whether it was one or both of these variables that affected the outcome (Koh, 2011). Additionally, some studies have participant groups that are limited by gender, ethnicity, or subject/grade-level taught, which impacts the generalizability of these results (Benson, 2009; Malone 2013).

Other challenges related to assessment research deal with the complexities inherent in assessment. It is difficult to measure the effects of assessment in isolation since a change in assessment practices often involves a pedagogical shift in the classroom (Black & Wiliam, 1998a; Volante & Cherubini, 2011). Introducing increased formative assessment, changing the focus of assessment from a normative approach to mastery learning, all represent significant shifts in instruction as well as assessment (Black and Wiliam, 1998a; Stiggins, 2006). Any
research conducted on the impact of assessment must also take the multi-faceted relationship between assessment and instruction into consideration.

Besides the complex relationship between assessment and curriculum, there are other underlying issues that must be addressed in order to fully understand assessment and its impact on students. Black and Wiliam (1998a) list eleven such underlying aspects of assessment, including assumptions of learning that underlie curriculum and instruction, roles and responsibilities of teachers and students in learning and assessment, beliefs of teachers about their students’ abilities and future prospects, issues relevant to race, gender, and socioeconomic status, and the extent to which the context in which studies are situated is artificial and the possible effects of this artificiality on the generalizability of results (p. 45). Eyal (2012) adds to this the additional dimension of digital assessment literacy and describes ways in which assessment must adapt in an era of technology-driven instruction.

Yet another challenge in researching assessment-related issues, and particularly in measuring teachers’ assessment literacy, pertains to the measurement instrument being used. As Gotch (2012) aptly points out, instruments based on the Standards for Teacher Competence in Educational Assessment of Students target areas of assessment literacy that are identified by experts as being important for teachers (American Federation of Teachers, National Council on Measurement in Education, and National Education Association, 1990). Because there has been, to date, no concerted effort to develop an assessment literacy-measurement instrument based on a job task analysis of actual assessment tasks that teachers are required to carry out during the course of the school year, there may be a disconnect between the characteristics of assessment literacy as measured by the Standards and those actually needed in a classroom setting.
Still another instrument-related challenge relates to the translation (or lack thereof) of assessment literacy as demonstrated by an instrument to the implementation of best practices in assessment in teachers' classrooms. Benson (1997) studied a group of teachers who participated in a three-year assessment literacy program, and who demonstrated knowledge on a measurement instrument of how to promote higher-order thinking among their students, but whose own classroom assessment tasks were low complexity. Gotch (2012) suggested measuring assessment literacy beyond that demonstrated by the survey instrument, to determine whether teachers can demonstrate this literacy through classroom practice. Because the ultimate goal of assessment literacy instruction is to improve classroom practices that impact students, such research is a necessary piece in the evaluation of any teacher assessment literacy professional development.

One final confounding facet of assessment research is that of multiple purposes for assessment. Teachers have reported using assessment for multiple purposes: to motivate students, to focus student attention, to assign grades, to change and focus their own instruction, to predict students’ performance on standardized assessments, to encourage accountability for students and teachers, and to use for grouping students by ability level (Benson, 1997; McMorris and Boothroyd, 1993; Reynolds, 1992). Given these multiple purposes for assessments that fall under the general heading of “formative,” “summative,” “classroom assessment,” or “accountability assessment,” it is not surprising that researchers have difficulty generalizing results from one study across other instructional situations. Consensus on terminology within the educational research community, as well as precision in terminology to capture some of these nuances, may help ensure that researchers attempting to generalize results across studies can do so with a reasonable degree of confidence.
Teacher Assessment Literacy and Assessment Quality

While correlations have been shown between higher teacher assessment literacy and higher quality classroom assessments, ample documentation exists that teachers have generally low levels of assessment literacy and demonstrate sometimes alarming classroom assessment practices (Benson, 1997; Gotch, 2012; Mertler, 2005). Black and Wiliam (1998a) found a preponderance of low-level, knowledge recall questions in teacher-developed assessments. These findings coincide with those of other researchers, who have found that the majority of classroom assessment items require students to memorize, not to demonstrate the application of, knowledge (Fleming and Chambers, 1983; McMorris and Boothroyd, 1993; Marso & Pigge, 1993). In an analysis of 342 classroom tests from teachers in the Cleveland school district, Fleming and Chambers (1983) discovered that approximately 80 percent of items across all subject areas evaluated knowledge of terms, knowledge of facts, and knowledge of rules and principles. Fewer than 2 percent of items were essay questions, which were generally avoided by even English teachers; 71 percent of items on senior high English tests assessed knowledge of facts.

These findings were consistent with those by Wright, Foran, Holmes, and Lou (2016), who found that 67.7 percent of studied items on English Language Arts assessments were written at low complexity levels (level one of Webb’s Depth of Knowledge), though mathematics items were primarily written at moderate complexity levels, with 61.6 percent of the sampled mathematics items measuring level two of Webb’s Depth of Knowledge. Items measuring high cognitive complexity (level three of Webb’s Depth of Knowledge) were extremely limited (6.8 percent in English Language Arts and 5.2 percent in mathematics), while no items in either sample measured level four of Webb’s Depth of Knowledge.
Black and Wiliam (1998b) hypothesize a correlation between insufficient wait times by teachers (when awaiting student responses to a question) and a preponderance of recall-type questions; questions that place low cognitive demands on students are relatively easy to answer in a classroom setting, whereas questions that require higher-order thinking also require longer wait times for students to compose their answers.

When Fleming and Chambers (1983) evaluated the Cleveland teachers’ classroom assessments for errors in the test items themselves, they found that 15-20 percent of items had errors in grammar, spelling, punctuation, or numbers. Short answer items were often ambiguous, with several possible answers that could be defended as being correct, while essay items often failed to include the approximate amount of time that should be spent on constructing a response (Fleming and Chambers, 1983). McMorris and Boothroyd (1993) found similar results in an evaluation of middle school science and math assessments; 35 percent of short answer/completion items contained errors, and 20 percent of multiple choice items contained errors. Wright et al. (2016) found that the most frequent problems with items were constructed response items that failed to clearly delineate expectations or that had unclear scoring rubrics, though spelling or grammar errors were also found in 16 percent of items. For the English Language Arts sample studied by Wright et al. (2016), alignment of items to standards also presented a serious problem, with only 41 percent measuring a standard that was present in the course description.

The types of errors found by McMorris and Boothroyd (1993) also coincided with the findings of Fleming and Chambers (1983); the most frequent completion-type errors were items that called for an ambiguous or nonspecific response, or structural errors such as blanks in the beginning or middle of a statement. Most common multiple choice errors included
nonhomogeneity of response options and the related error of the longest option being the answer key, though Wright et al. (2016) also found problems with clear wording of question stems and answer choices. Other multiple choice item errors included cues present in the question stem, spelling or grammatical errors, or stems not written as a question.

Although there are concerns with the assessments being developed and used by teachers, there are also more fundamental problems caused by a lack of assessment literacy. Teachers often display confusion regarding the purpose of assessment, with a focus on grading, rather than student learning, as the outcome (Black and Wiliam, 1998a; Black and Wiliam, 1998b). Milton et al. (2013) reported that, when surveyed regarding the six Florida Educator Accomplished Practices, Assessment was one of the two areas in which teachers felt least prepared by their teacher preparation institutions. Additionally, some teachers use classroom assessments to predict student performance on standardized tests, rather than to provide them with the information that they actually need about student learning (Black and Wiliam, 1998b). The multiple purposes of assessment, and confusion about the role played by each type of assessment, can lead to a tension between the formative and summative purposes served by assessment.

While there is ample evidence of poor practices in teacher-developed assessments, it is not so clear how teachers can correct these errors. One possible solution, receiving guidance from peers or from administrators who serve in the role of instructional leaders, appears to be currently unlikely. Volante and Cherubini (2011), in a study of the assessment literacy of school administrators, found that administrators generally scored low on assessment self-efficacy questions. Participants reported a lack of confidence in their own knowledge about assessment, as well as a lack of professional development in assessment and evaluation. This was consistent
with the findings from Marso and Pigge’s (1993) study, in which district testing directors reported providing little or no support to teachers in the area of assessment creation. School administrators also struggled with their own capacity to provide support to teachers in assessment development. Marso and Pigge (1993) found a negative correlation between school administrators’ ratings of teachers’ test creation skills and the observed quality of these teacher-developed assessments.

This situation is compounded at the secondary level, where departmentalization makes discussions among teachers about assessment practices the exception rather than the rule (Volante and Cherubini, 2011; Black and Wiliam, 1998b). These findings concur with those of Black and Wiliam (1998a), who report a mistrust and lack of awareness of other teachers’ work with assessments. This mistrust and awareness is compounded by policies that link assessment to high-stakes decisions regarding teacher evaluations as well as student grades or promotions.

**Impact of Assessment Policies**

Assessment-related policies emphasize summative assessment almost exclusively, with very little, if any, emphasis placed on formative assessment (DeLuca et al., 2012; Stiggins, 2009). Considering the funds spent yearly on the development, scoring, and evaluation of standardized, summative assessments, relatively few funds are spent on training teachers to effectively develop and administer classroom assessments, although the majority of instructional decisions made in the classroom are made on the basis of formative assessments (Stiggins, 2009). Although classroom tests are the basis of instructional and grading decisions, significantly fewer resources are spent evaluating the quality of these assessments, compared to the ample documentation and research available on state-level, summative assessments. Just as many teachers report using assessments as a motivational strategy with their students, policy makers, who were generally successful in school and who responded to this motivation, often
believe that the same holds true for every student and implement policies that call for an increase in standardized testing every year (Florida Department of Education, 2013b; Stiggins, 2009). These policies can have a significant impact on classroom instruction, as accountability tests often impact what and how students are taught (Popham, 2009).

Stiggins (2006) identified three levels of assessment-related information: the classroom level, wherein assessments are used for instructional decisions, the program evaluation and support level, wherein assessment data is used to identify students in need of additional support of measure program effects, and the institutional and policy level, which uses assessment data to evaluate programs, guide policy decisions, and allocate resources. Each of these consumers of assessment data has different assessment needs, and a balanced assessment program that can provide data for each level is more appropriate than an assessment program based exclusively on standardized, summative assessments administered once yearly.

The creation of the Standards for Teacher Competence in Educational Assessment of Students by the American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association in 1990 heralded the recognition of essential teacher skills in assessment development, selection, use, and interpretation of results, given this climate of assessment-driven accountability. Seven standards are outlined in this document:

Standard 1: Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.

Standard 2: Teachers should be skilled in developing assessment methods appropriate for instructional decisions.

Standard 3: The teacher should be skilled in administering, scoring and interpreting the results of both externally produced and teacher-produced assessment methods.
Standard 4: Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.

Standard 5: Teachers should be skilled in developing valid pupil grading procedures that use pupil assessments.

Standard 6: Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.

Standard 7: Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information (American Federation of Teachers, et al., 1990).

Despite this formal recognition and documentation of requisite skills, relatively few teacher and administrator preparation programs currently include instruction in assessment literacy (Stiggins, 2009; Volante and Cherubini, 2011). In a study conducted by Mertler (2005), both pre-service and in-service teachers scored lowest on Standard 5, which pertains to the ability to develop valid pupil grading procedures using assessments. These results are not surprising, given the lack of training provided to teachers in assessment literacy, as well as the documented misalignment of classroom grades. Informal assessment has the greatest impact on instruction and learning, though formal assessment has the greatest effect on student grades (Benson, 1997; Stiggins, 2009).

Impact of Assessments on Students

Because education is always, ultimately, about the students’ learning experiences and outcomes, assessment must be viewed through the lens of its impact on students. Earlier literature reviews of formative assessment experiments resulted in effect sizes of .4 to .7, when norm-referenced, competition-based assessments are replaced by assessments for learning (Black
and William, 1998b; Stiggins, 2006). Low-achieving students appear to be most impacted by the effective implementation of formative assessment, whereas, conversely, there is evidence that tying performance feedback to grades instead of learning is detrimental to these lower-achieving students.

Self-assessment, as a type of formative assessment appears to be an effective learning strategy for students. However, students cannot learn to be effective self-assessors unless they have a clear understanding of the learning targets and an understanding of descriptors of performance at each level (Black and William, 1998b; Harland and Sawdon, 2011; Stiggins 2006). Stiggins (2006) defines seven principles of assessment for learning:

1. Instruction should begin with an explanation of learning targets couched in student-friendly language.
2. Students should be shown samples of work at ranges of quality.
3. Students should be given access to descriptive feedback, by their teacher, by peers, or by both, that will help them understand how to improve.
4. Students should be taught how to generate descriptive feedback.
5. Teachers should introduce one facet of quality at a time.
6. Students should be taught the practice of focused revision, improving work by one facet at a time.
7. Students should be taught how to understand, keep track of, and reflect on changes in their proficiency (p. 17)

Teaching students how to become proficient self-assessors, and providing them with rapid, descriptive feedback on their performance can have significant impacts on student achievement. However, too many students become confused through mixed messages they receive from their
teachers; teachers may tell students to become risk takers, and to focus on the process of learning more than the product, but this can be undermined when summative assessments have the greatest effect on grades, and formative assessments, which have the greatest impact on student learning, have little effect on student grades.

Popham (2009) also claims that low-quality assessments have an impact on student attitudes toward the teacher, and that flawed tests discourage students. Though these assumptions appear to be anecdotal only, there is logic to the claim that poorly designed classroom assessments can have negative impacts on student learning, and certainly student learning when measured by flawed assessments is suspect at best.

Finally, there are indirect effects of assessment on students, such as those described by Hoover & Abrams (2013), who studied teachers’ use of assessment to modify classroom instruction. The majority of teachers (94 percent) reported using assessment results to aid their reteaching of concepts, while 92 percent stated that they used assessment results to help pace future instruction.

Types of Professional Development

In their Teaching and Learning International Survey (TALIS), the Organisation for Economic Cooperation and Development (2009) categorized formal professional development into seven types: courses or workshops, education conferences or seminars, qualification or degree program, observation visits to other school sites, participation in a professional network of teachers, individual or collaborative research, and mentoring, peer observation, or coaching. In their exploration of the use of these seven types of professional development across 23 countries, courses and workshops were the most common, followed by education conferences or seminars, then participation in a teacher network, individual or collaborative research,
mentoring/peer observation/coaching, observation visits to other schools, and, finally, qualification or degree program.

Boyle, White, and Boyle (2004), focusing on long-term professional development, identified nine strategies: study groups, mentoring, research/inquiry, drop-in clinics, observation of colleagues, coaching, networks, sharing practice, and onsite or online courses. Participating teachers expressed the greatest satisfaction with sharing practice, followed by observation of colleagues, onsite courses, workshops, and mentoring. Only 48 percent of participants expressed satisfaction with online courses as a professional development strategy. However, since this study was conducted, there have been significant advances in technology that may have made online professional development a more viable option.

Yurtsever (2013) classified professional development into two general types: traditional and constructivist, and then into four models. Of the four models, one was classified as traditional: a training model. The remaining three models, mentoring, peer coaching, and self-directed, were classified as constructivist. Turkish teachers in this study expressed stronger preferences for all three constructivist models over the traditional training model.

There are commonalities in models of effective professional development across studies, though the terminology used may differ. Bayar (2014) identified the following six components of effective professional development:

- Match to teacher needs
- Match to school needs
- Teacher involvement in the design or planning of professional development
- Opportunities for active participation
- Long-term engagement, and
• High quality instructors.

Casale (2011) found some of the same components in her exploration of teachers’ perceptions of professional development. Teachers expressed a desire for professional development to be relevant (a match to their needs), to have opportunities for choice in the content as well as the mode of professional development, and to have opportunities for collaboration with their peers. Teachers who had experienced online professional development felt that the online delivery mode had the advantage of being convenient, found it more motivating, and felt that there were fewer distractions. They also enjoyed the ability of online professional development to put them into contact with their counterparts around the country.

One of the most common concerns that teachers expressed about the types of professional development that they were receiving was a lack of follow through. Casale (2011) found that 65.2 percent of teachers reported receiving no coaching or mentoring to assist them with refining their practice, and that 53.1 percent felt they did not receive appropriate feedback after professional development had occurred.

Assessment-Related Professional Development Opportunities and Challenges

Though there have been several studies on assessment quality, and on teacher assessment literacy as measured by a variety of instruments, there are very few studies that evaluate professional development on assessment literacy for teachers. Koh (2011) examined two interventions designed to increase teachers’ assessment literacy. The treatment intervention was ongoing, sustained professional development over two school years, with monthly school meetings and two end of year meetings to review student work together with other teachers, while the comparison condition was a one-day workshop at the beginning of each of the two years, without any follow up. The longer intervention resulted in student work which had a significant increase in higher-order thinking skills; given extended professional development, as
well as a training curriculum that included teacher peer groups collaboratively evaluating student work, the intervention translated to a change in student performance (Koh, 2011). Due to limited participant descriptions, as well as a limited geography, the study cannot be said to have high generalizability, but the findings are certainly intriguing, given some of the challenges facing professional development in assessment literacy.

The findings of the Koh (2011) study align with recommendations by DeLuca et al. (2010; 2012) for best practices in professional development. Traditional teacher professional development workshops follow the expert transmission model, with teachers as passive recipients of information in an artificial setting removed from their classrooms. This decontextualized transmission of knowledge relies upon the teacher to determine how practices transmitted at the workshop may fit into the context of his or her own classroom.

DeLuca et al. (2012) identified three essential elements of effective contemporary professional development. Essential element 1: teacher as active learner, fits in with theories of adult learning that suggest that adults are more likely to “engage in deep learning through collaborative and context-based inquiry” (DeLuca et al., 2012, p. 19). This first element acknowledges that teachers can learn from collaboration with each other.

The second essential element, ongoing, contextualized learning, recognizes the importance of giving teachers context for learning that allows them to use knowledge learned at professional development opportunities in their own classrooms. Just as with element one, this element acknowledges the importance of teacher learning communities, as this job-embedded professional development does not take place in isolation. Malone (2013) discovered something similar when evaluating an assessment literacy module with two groups of participants: teachers and language testing experts. The teacher participants gave feedback regarding the usefulness of
instructional modules, indicating that they desired professional development that was delivered with clarity and that could be understood in the context of a teaching situation.

Finally, the third element stressed by DeLuca et al. (2012), process-based learning and reflective practice, represents a departure from traditional product-based professional development. This reflective model emphasizes a learning goal that includes not only knowledge acquisition, but also improvements to professional practices resulting from belief and behavioral changes.

**Feedback and Assessment Item Development**

Earlier research suggests that feedback can serve as reinforcement of learning, connecting responses to prior stimuli, as additional information, validating or providing incentive to change the initial response, and as a scaffold that helps learners to both construct internal schemata and to analyze their own learning processes (Gielen, Peeters, Dochy, Onghena & Struyven, 2010). Lou, Dedic, and Rosenfield (2003) provide a model of effective feedback, in which feedback from teachers and peers help the learner to close the gap between the current state of performance and the goal state of performance. Cho and Cho (2011) describe the provision of feedback as a constructive learning activity, in which the reviewer must internalize criteria and then transfer this knowledge to help repair another’s work. This theory coincides with that of other researchers who view peer feedback as a useful technique to close the gap between actual performance and the desired standard (Duijnhower, Prins, & Stokking, 2012; Gielen & De Wever, 2015; Kamp, van Berken, Popeijus, Leppink, Schmidt, & Dolmans, 2014).

Lou et al. (2003) examine the provision of feedback within both the online learning and face-to-face learning environment. While face-to-face learning environments have the advantage of real-time verbal and non-verbal communication and feedback, online learning environments have their own considerations for feedback. Using discussion boards, teachers can provide
feedback to benefit more students, as their feedback may be targeted to individual students or to common misconceptions held by multiple students. While face-to-face feedback allows teachers to make instant changes to the learning environment, based upon this real-time feedback, teachers of online courses require more preparation time to make adjustments based on student feedback.

Nicol, Thomson and Breslin (2014) describe peer feedback as a higher order skill, in which learners must engage in evaluative judgments about the work of others and, through a reflective process, about their own work. When providing feedback to others, students make use of these higher order problem solving processes: analyzing, problem representation, solution development and justification, and solution evaluation (Cho & Cho, 2011; Cho & MacArthur, 2010; Nicol et al., 2014; Snowball & Mostert, 2013).

While studies that analyze the effect of feedback on the quality of writing produced by assessors as well as assessees are in the minority, the research that does exist in this arena suggests that learners derive more benefit from providing feedback to others than from receiving feedback on their own work from others (Cevik, 2015). Assessors learn from exposure to others’ writing and from viewing these writings through a critical lens. Cho and Cho (2011) found that reviewers’ comments to their peers had a significant impact on the reviewers’ own writings, which Snowball and Mostert (2013) found that providing feedback to peers helped broaden writers’ perspectives by exposing them to others’ writing, which in turn helped them gain a realistic sense of the value of their own work.

Benefits of peer feedback for those being assessed have also been shown to vary, depending on the type of feedback provided by the assessor. Cevik et al. (2015) and Pelgrim et al. (2013) found that students were more likely to revise their work and to provide specific self-
reflections when they received detailed suggestions. Gielen and De Wever (2015) noted that these same findings held true for assessors. Providing structure or requests for specific feedback improved the quality of the feedback that the assessors were able to produce, though these benefits diminished over time as assessors were able to practice providing structured feedback (Gielen & De Wever, 2015).

Students were also more likely to give credence to negative feedback and to make revisions on the basis of this type of feedback. These findings align with those of Cho and Cho (2011), who found that receiving positive comments on the surface features of the writing had a significant, negative impact on the overall quality of the revised writing product.

Guasch, Espasa, Alvarez, and Kirschner (2013) also discovered meaningful differences in the impact of feedback based on the type of feedback received. They classified feedback into the following four types:

- Corrective feedback: This type of feedback is focused on the assignment requirements, and addresses gaps between product and the desired performance standard.
- Epistemic: Epistemic feedback requests explanation or clarification that requires the assessee to view his or her writing in a critical way.
- Suggestive: Suggestive feedback gives specific directions or guidance on how to proceed, in order to address perceived deficits in the product being assessed.
- Epistemic and suggestive: This category encompasses feedback from both the epistemic and the suggestive categories.

While all types of feedback improved students’ writing in this study, Guasch et al. (2013) found that receiving epistemic feedback resulted in significant improvements, when compared with the receipt of suggestive or corrective feedback. There were no significant differences between the
quality of the writing after receiving epistemic feedback compared to receiving epistemic and suggestive feedback, however.

**Media Effects**

The debate on whether the mode of instructional delivery has an impact on student outcomes gained notoriety in the early nineties with the media effects debate between Clark (1983, 1994) and Kozma (1994). Kozma (1994) contended that the potential for media to influence learning existed because of a potential relationship between the characteristics of a learning environment and the cognitive processes of the learner. Clark (1994), in contrast, argued for a replaceability test wherein the influence of media would be judged by whether it could be replaced by an alternate media and still deliver the same results. His contention was that media effects were typically caused by differences in the instructional design or teaching method, not by differences in the media itself.

Later research into effects of the mode of training (online, face-to-face, or blended/hybrid models that combine online and face-to-face instruction) has yielded mixed results. Bernard et al. (2004), in examining student achievement, attitude, and retention outcomes of distance education compared with face-to-face classroom instruction, found effect sizes of near zero, with wide variability. Fishman et al. (2013), in an exploration of online versus face-to-face professional development on curriculum implementation, found no significant differences in outcomes of teacher efficacy, teachers’ evaluations of the course, teachers’ knowledge of the content, or student outcomes as measured by pre- and post-test gains on the science content on which the teachers were being provided professional development. Similar results were found by Fisher et al. (2010), who found no differences in teacher satisfaction, teacher content knowledge, or student content acquisition, based on type of training (online or face-to-face). Russell, Carey, Kleiman, and Venable (2009), who held instructor, activities, and amount of time
spent in training constant across both online and face-to-face interventions, also found comparable outcomes for teachers’ mathematical understanding, pedagogical beliefs, and instructional practices. Similarly, Wladis, Hachey, and Conway (2014) found no significant differences in STEM course outcomes based on online or face-to-face delivery, after propensity score matching was used.

Not all research supports these findings that there are no differences in training outcomes by mode of delivery, but the results of these studies are mixed and sometimes contradictory. Powell, Diamond, Burchinal, and Koehler (2010) found some differences between online and face-to-face training, but these differences were not consistent; in two cases, outcomes favored the online mode, while in two cases, outcomes favored the face-to-face mode. Similarly, Young and Duncan (2014), in a comparison of over 8,000 university student ratings of online and face-to-face courses, found that online courses were rated significantly higher by students for student effort, while face-to-face courses were rated significantly higher for communication, faculty and student interactions, grading, instructional methods, and course outcomes. Despite the statistical significance of these findings, all effect sizes of differences were very small.

Lou, Bernard, and Abrami (2006) followed the earlier meta-analysis by Bernard et al. (2004) with a closer examination of course components for synchronous and asynchronous online instruction for undergraduate students on achievement. They found that effect sizes for instructor-led synchronous and face-to-face instruction were not significant and were near zero, when instruction was delivered by the same instructor using the same course content and materials. However, when asynchronous online courses incorporated media that facilitated discussion among students, the online students significantly outperformed their peers who received face-to-face instruction.
Brocato, Bonanno, and Ulbig (2015), who also examined instructional evaluations of online courses compared to face-to-face courses, also found that overall, instructors tended to receive higher ratings for their face-to-face courses. While course level and instructor’s gender did contribute significantly to the summary ratings of face-to-face courses, however, this did not appear to be the case for online courses.

Cavanaugh and Jacquemin (2015), who explored grade-based student learning outcomes of online and face-to-face courses, found significant differences that were, as with the findings of Young and Duncan (2014), very small (0.07 GPA points on a 4-point scale). They discovered an interaction effect between grade point average (GPA) and course type. Students with higher GPAs tended to do better in online courses compared to face-to-face courses, while struggling students performed worse in online courses than in face-to-face courses.

In contrast to these findings, though, Smith (2013), in a comparison of face-to-face with blended (combination online and face-to-face) courses, found that students in the blended class reported higher scores on a measure of their perceptions of learning, connectedness, enjoyment, and teacher support. Consistent with earlier findings, though, was the result of no difference in students’ performance on assessments of learning.

Summary
Given promising evidence of formative assessment’s positive effects on student achievement, it behooves educational researchers to better understand teacher acquisition of assessment literacy in order to assist teachers in selecting and developing a variety of measurement instruments. Given some of the concerns about assessment quality raised by researchers, teachers, and administrators, there is a strong need for additional research into effective professional development to better prepare teachers to develop or select high-quality classroom assessments. Assessment has been shown to be a powerful tool for student learning.
but only once we fully understand its uses, and our own purposes underlying these uses, will we truly be able to align instruction and assessment with student needs.
CHAPTER THREE:

METHOD

This study examined teachers’ acquisition of assessment literacy and the process and impact of training teachers to develop assessment items. The study sought to answer the following questions:

Q1: Is there any significant relationship between the following teacher-level factors:

- years of experience
- prior assessment development experience
- type of degree, and
- National Board certification

and the participating teachers’ ability to write high-quality assessment items?

Q2: What impact, if any, does the role selected by teachers (item writers vs. item writers/reviewers) have on teachers’ abilities to write high quality assessment items?

Q3: What impact, if any, does the mode of training have on the quality of items written by teachers?

Q4: What do teachers report as the effects of participation in an assessment development project, and are there differences in these effects by mode of training?

Q5: Is there a relationship between teachers’ perceptions of the assessment literacy training and the mode of training (in-person vs. online) they received?
Research Paradigm

This research was conducted within a pragmatic theoretical framework; in this context, the research conducted was applied research. Pragmatism, developed by C.S. Peirce in the 1870s, and further refined by Chicago School members like Dewey, Mead, and Jane Addams, stresses the relationship between practice and theory, with an emphasis on outcomes (Seigfried, 1999). This framework is suited to mixed methods research, particularly given Venkatesh et al.’s (2013) “dictatorship of the research question,” in which the research question itself drives the selection of the worldview (qualitative vs. quantitative) as well as the research methods used. The desired outcome of this study is to arrive at actionable answers to the research questions, in order to contribute in a practical way to the existing knowledge base about teachers’ assessment literacy, specifically the training of teachers to write assessment items. In the pragmatic viewpoint, validity as a consideration is not independent of the consequences of research (Noddings, 2005); because of this, while there was some use of self-reported data in this study design, triangulation was used to ground teachers’ own beliefs about their assessment literacy in empirical data: their assessment items written for the assessment collaborative project.

Research Design

The research design was a mixed methods, ex-post-facto study, since the data that was used in the study was collected during the 2012-2015 administration of the assessment collaborative project. Research methods used include two-level hierarchical linear models, since items are nested within teachers, a Chi square test of independence, and a constant comparative analysis for open-ended survey results. The independent variables include mode of training, pre-existing teacher factors such as National Board certification, years of experience, and graduate degree, and teacher role (item writer or item writer and reviewer). Dependent variables include open-
ended survey results, training evaluation survey results, and item quality data such as number of reviews per item, item discrimination, and the proportion of accepted items.

**Training Mode**

Participants in the project were able to self-select into one of three training modes: online synchronous through a live webinar, online asynchronous through a self-paced training, or in-person trainings. Because project leads encouraged participants to attend multiple training opportunities, a fourth mode of training, blended, was created for those who attended one of the two online modes of training in addition to an in-person training. The composition of the three samples was very similar in terms of the distribution of participants across the four modes of training (see Table 1). Approximately one-third of the participants (32.8% to 36.5%) opted to attend an in-person training. Approximately one-third of participants (32.8% to 34.1%) elected the online synchronous training, and approximately one-fifth elected the online asynchronous training (16.1% to 20.4%). Between 11.2% and 14.5% of participants chose the blended option, attending an in-person training as well as one of the two online trainings (see Table 1).

Table 1

<table>
<thead>
<tr>
<th>Training Mode</th>
<th>Sample 1: All Participants</th>
<th>Sample 2: Participants with Items Written and in Item Bank</th>
<th>Sample 3: Participants with Item Performance Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online synchronous</td>
<td>34.1% (143)</td>
<td>32.8% (45)</td>
<td>33.9% (21)</td>
</tr>
<tr>
<td>Online asynchronous</td>
<td>18.1% (76)</td>
<td>20.4% (28)</td>
<td>16.1% (10)</td>
</tr>
<tr>
<td>In person</td>
<td>36.5% (153)</td>
<td>32.8% (45)</td>
<td>35.5% (22)</td>
</tr>
<tr>
<td>Blended</td>
<td>11.2% (47)</td>
<td>13.9% (19)</td>
<td>14.5% (9)</td>
</tr>
</tbody>
</table>

While all four modes of training were structured to include the same basic content, the delivery method as well as learning activities differed. The online synchronous and online
asynchronous trainings were the most similar in content, though certain activities such as responding to other participants’ work were dissimilar in execution. For the online synchronous training, delivered through Cisco WebEx, participants had the opportunity to use the chat feature to converse in real time with the instructor and other participants. In the online asynchronous training, delivered through Blackboard CourseSites, participants provided feedback to each other, and received feedback from the instructor, in a discussion board format. The in-person training, while it had the same training objectives as the online trainings, differed in the type of interaction that participants had with each other, as well as in the type of learning activities that occurred. Participants in the in-person training were given the most extensive opportunities to provide and receive feedback on items and were given the most hands-on practice in item writing. An outline of each of these different training modes can be found in Table 2 below, with a more detailed outline provided in Appendix B.

Additionally, the three training opportunities differed in length of time. In-person trainings were conducted over a period of four days, two of which were spent in direct instruction, and two of which were spent working independently or with other participants to develop and review items. The WebEx portion of the online synchronous training lasted for three hours (two 90-minute sessions), with pre-test, post-test, and between-session homework accounting for an additional 90-120 minutes. The online asynchronous portion reportedly took participants between six and eight hours to complete.

Participants

This study took place in the state of Florida. The assessment collaborative project leads distributed recruitment information to all sixty-seven public school districts throughout Florida, though districts followed different recruitment methods when inviting their teachers to
Table 2

Comparison of Training Modes for Assessment Development Project

<table>
<thead>
<tr>
<th>Mode</th>
<th>Online Synchronous</th>
<th>Online Asynchronous</th>
<th>In-Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Time Spent Content</td>
<td>4.5-5 hours</td>
<td>6-8 hours</td>
<td>4 days</td>
</tr>
<tr>
<td>Overview</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Introductions</td>
<td>No</td>
<td>Yes – discussion board</td>
<td>Yes – in person</td>
</tr>
<tr>
<td>Franzipanics quiz &amp; review Provide Feedback to Peers</td>
<td>Yes – synchronous discussion</td>
<td>Yes – quiz &amp; video</td>
<td>Yes – instructor feedback</td>
</tr>
<tr>
<td>Item Specifications &amp; Selected Response Items</td>
<td>Yes – presentation</td>
<td>Yes – presentation</td>
<td>Yes – presentation</td>
</tr>
<tr>
<td>Practice with Selected Response Items</td>
<td>No</td>
<td>Yes – interactive presentation</td>
<td>Yes – editing items individually, then whole group discussion</td>
</tr>
<tr>
<td>Self-reflection on learning Depth of Knowledge &amp; Cognitive Complexity</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Depth of Knowledge Practice</td>
<td>Yes – presentation</td>
<td>Yes – presentation</td>
<td>Yes – presentation</td>
</tr>
<tr>
<td>Copyright Issues</td>
<td>Yes – presentation</td>
<td>Yes – presentation</td>
<td>Yes – presentation</td>
</tr>
<tr>
<td>Constructed Response Item Writing</td>
<td>Yes – presentation</td>
<td>Yes – presentation</td>
<td>Yes – presentation</td>
</tr>
<tr>
<td>Constructed Response Item Writing Practice</td>
<td>Yes – presentation</td>
<td>Yes – presentation</td>
<td>Yes – presentation</td>
</tr>
<tr>
<td>Constructed Response Scoring Rubric</td>
<td>No</td>
<td>Yes – practice</td>
<td>Yes - practice with peer and instructor feedback</td>
</tr>
<tr>
<td>Bias and Sensitivity</td>
<td>Yes – presentation</td>
<td>Yes – presentation and wiki activity</td>
<td>Yes – presentation and group discussion</td>
</tr>
</tbody>
</table>
participate. Some districts shared the advertisement with all teachers, district-wide, while others shared the advertisement only with teachers who were recommended by district-level staff or by administrators. Additionally, the recruitment information was distributed through colleges or departments of education at four public universities in Florida. All participants were required to apply to participate in the project, and applicants were selected for participation by the project manager, based on years of experience, school district, and licensure/subject area.

Table 3 presents the demographics of the three samples included in this study. The teacher sample size for Sample 1 included 419 participants who applied to serve as item writers and/or reviewers for the assessment collaborative project from 2013 to 2015 and who completed the required training and received accounts to develop items in either the state’s grant-funded item bank software, the IBTP. Participants in this sample included active and retired teachers from public and charter schools, employed by forty-five school districts and one university in Florida. Of this sample, 74.7% (313 participants) were female, and 25.3% (106) were male. The majority of participants (62.8%) reported having a graduate degree, while 11.7% reported having earned their National Board certification. Table 4 also contains a detailed breakdown by years of experience for Sample 1. The only data available for Sample 1 was the training evaluations that they completed after the conclusion of the assessment development training.

Of the 419 participants, the only item level writing and review data that was available was for those who wrote items in the IBTP for the grant-funded project. 127 participants who wrote items in the IBTP were included in Sample 2, which was a subset of the larger Sample 1 (see Figure 3). Ten participants were excluded from Sample 2, which had originally 137 participants, because they wrote fewer than three items. The remaining participants in Sample 1 either did not write any items or wrote items in the non-grant funded assessment platform and
Table 3

Participant Demographics for Each Sample

<table>
<thead>
<tr>
<th></th>
<th>Sample 1: Participants who Completed the Training and Evaluation</th>
<th>Sample 2: Participants with Items Written and in Item Bank</th>
<th>Sample 3: Participant with Item Performance Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of participants</td>
<td>419</td>
<td>127</td>
<td>62</td>
</tr>
<tr>
<td>Percent Female</td>
<td>74.7% (313)</td>
<td>74.8% (95)</td>
<td>74.2% (46)</td>
</tr>
<tr>
<td>Percent Male</td>
<td>25.3% (106)</td>
<td>25.2% (32)</td>
<td>25.8% (16)</td>
</tr>
<tr>
<td>Percent with National Board certification</td>
<td>11.7% (49)</td>
<td>15.0% (19)</td>
<td>16.1% (10)</td>
</tr>
<tr>
<td>Percent with graduate degrees</td>
<td>62.8% (263)</td>
<td>67.7% (86)</td>
<td>71% (44)</td>
</tr>
<tr>
<td>Percent with 0-3 years teaching experience</td>
<td>7.4% (31)</td>
<td>6.3% (8)</td>
<td>8.1% (5)</td>
</tr>
<tr>
<td>Percent with 4-6 years teaching experience</td>
<td>12.6% (53)</td>
<td>14.2% (18)</td>
<td>14.5% (9)</td>
</tr>
<tr>
<td>Percent with 7-10 years teaching experience</td>
<td>24.3% (102)</td>
<td>20.5% (26)</td>
<td>17.7% (11)</td>
</tr>
<tr>
<td>Percent with 11-15 years teaching experience</td>
<td>22.0% (92)</td>
<td>20.5% (26)</td>
<td>21.0% (13)</td>
</tr>
<tr>
<td>Percent with 16-20 years teaching experience</td>
<td>11.9% (50)</td>
<td>18.9% (24)</td>
<td>14.5% (9)</td>
</tr>
<tr>
<td>Percent with greater than 20 years teaching experience</td>
<td>21.7% (91)</td>
<td>19.7% (25)</td>
<td>24.2% (15)</td>
</tr>
<tr>
<td>Mean number of items written</td>
<td>106.9</td>
<td>123.4</td>
<td></td>
</tr>
<tr>
<td>SD number of items written</td>
<td>151.0</td>
<td>131.7</td>
<td></td>
</tr>
<tr>
<td>Minimum number of items written</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Maximum number of items written</td>
<td>891</td>
<td>406</td>
<td></td>
</tr>
<tr>
<td>Mean percentage of items accepted</td>
<td>.921</td>
<td>.95</td>
<td></td>
</tr>
<tr>
<td>SD percentage of items accepted</td>
<td>.191</td>
<td>.13</td>
<td></td>
</tr>
</tbody>
</table>

Note. Sample 2 is a subset of Sample 1, and Sample 3 is a subset of Sample 2. Participants with fewer than 3 items written were removed from Sample 2 and Sample 3.
thus item data is not available. The demographics were very similar for these two samples. Like Sample 1, the majority of participants in Sample 2 (75.2%) were female, while more than half had a graduate degree (67.2%). Table 4 contains a detailed description of this sample’s demographics, including years of experience. These participants hailed from thirty-five Florida school districts and one university.

Finally, of the 127 teachers who created three or more items in the IBTP, student performance data from field tests was available for items written by 62 teachers from twenty-three districts and one university. Sample 3 was composed of these 62 participants and was a subset of Sample 2, itself a subset of Sample 1. Thus, Sample 3 contained only participants who had written more than 3 items, since this was an inclusion criteria for Sample 2. Figure 3 illustrates the relationships among the three samples.

![Figure 3. Relationship between each sample.](image-url)
The gender balance for Sample 3 was very similar to that of Samples 1 and 2; 74.2% were female and 25.8% were male. As with the other two samples, the majority of these participants (71%) had earned graduate degrees, and 16.1% were National Board certified. The balance by years of experience was also similar for these three samples.

Participants were also categorized by prior assessment experience (see Table 4). As part of the application process, item writers were asked to describe any relevant prior assessment experiences. Applicant responses were categorized by type of experience. All teachers were classified as having classroom-level assessment experiences, given that teachers develop or select assessments for their own classrooms. Applicants who described participating in common assessment initiatives in their school were grouped as having school-level assessment experiences. Those who described participation in district-wide assessment initiatives were classified as having district-level experiences, while those who described participation in statewide assessment initiatives, either for the state’s standardized assessments or through participation in a statewide item development project similar to the assessment collaborative project were classified as having state-level experiences. Some participants also listed serving in the role of item writer or reviewer for a private corporation, so a fifth category was created: experience with a vendor/private company. Table 4 shows the composition of each sample by experience level. In all three samples, at least 48 percent of teachers had some assessment experience beyond the classroom level, while fewer than 50 percent of participants had prior assessment experience beyond the school level (at the district, state, or vendor level).

Participants were representative of Florida’s geographic diversity, hailing from rural, urban, and suburban school districts. Table 5 shows the distribution of teachers across school districts in Florida. Sample 1 had representation from 45 of the 67 public school districts, as
well as one university representative. Sample 2 included participants from 35 school districts, while sample 3 was composed of participants from 23 school districts.

Table 4

Prior Assessment Experience by Sample

<table>
<thead>
<tr>
<th>Assessment Development Experience</th>
<th>Sample 1: All Participants</th>
<th>Sample 2: Participants with Items Written and in Item Bank</th>
<th>Sample 3: Participants with Item Performance Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any assessment experience beyond classroom level</td>
<td>66.6% (279)</td>
<td>48.0% (61)</td>
<td>61.3% (38)</td>
</tr>
<tr>
<td>School level assessment experience</td>
<td>52.0% (218)</td>
<td>19.7% (25)</td>
<td>27.4% (17)</td>
</tr>
<tr>
<td>Any assessment experience beyond school level</td>
<td>31.5% (132)</td>
<td>34.6% (44)</td>
<td>41.9% (26)</td>
</tr>
<tr>
<td>District level assessment experience</td>
<td>21.0% (88)</td>
<td>25.2% (32)</td>
<td>30.6% (19)</td>
</tr>
<tr>
<td>State level assessment experience</td>
<td>12.2% (51)</td>
<td>13.4% (17)</td>
<td>17.7% (11)</td>
</tr>
<tr>
<td>Assessment experience with a vendor/private company</td>
<td>5.5% (23)</td>
<td>3.9% (5)</td>
<td>1.6% (1)</td>
</tr>
</tbody>
</table>

Table 5

Participant Affiliation by Sample

<table>
<thead>
<tr>
<th>District</th>
<th>Sample 1: Number of participants</th>
<th>Sample 2: Participants with Items Written and in Item Bank</th>
<th>Sample 3: Participants with Item Performance Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alachua</td>
<td>9</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Bay</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Bradford</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Brevard</td>
<td>9</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Brevard</td>
<td>32</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Broward</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Charlotte</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Citrus</td>
<td>13</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>
Table 5 (Continued)

<table>
<thead>
<tr>
<th>District</th>
<th>Sample 1: Number of participants</th>
<th>Sample 2: Participants with Items Written and in Item Bank</th>
<th>Sample 3: Participants with Item Performance Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeSoto</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Escambia</td>
<td>18</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Flagler</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Gadsden</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hardee</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Hendry</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hernando</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Highlands</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hillsborough</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Holmes</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Indian River</td>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Jackson</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Jefferson</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lake</td>
<td>32</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>Lee</td>
<td>13</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Leon</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Levy</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Liberty</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Manatee</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Martin</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Miami-Dade</td>
<td>17</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Okaloosa</td>
<td>11</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Okeechobee</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Orange</td>
<td>14</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Osceola</td>
<td>50</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Palm Beach</td>
<td>22</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Pasco</td>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Pinellas</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Polk</td>
<td>7</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Putnam</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sarasota</td>
<td>17</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Seminole</td>
<td>13</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>St. Lucie</td>
<td>13</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Sumter</td>
<td>9</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Volusia</td>
<td>24</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Wakulla</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Walton</td>
<td>10</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>University</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>No affiliation</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Data Sources

Participant Application

The application survey that all 419 assessment collaborative participants from Sample 1 completed was delivered online, and contained a combination of multiple choice, multiple-select, and open-ended questions. Participants were asked for contact information, demographic information such as years of experience, advanced degrees, affiliated institution or school district, and type of certification, in addition to open-ended questions such as, “Please describe your experience writing or designing curriculum at the school, district, and/or state level.” and “Please describe your experience in assessment development at the school, district, and/or state level.” A copy of the application can be found in Appendix C.

Training Evaluation Questionnaire

After completing the online or in-person training, teachers were asked to complete a 22-question training evaluation questionnaire, found in Appendix D. The questionnaire contained two demographic questions, asking teachers for information about themselves and the training that they attended, and ten selected response questions. The remaining ten questions were open response items. 456 responses to the evaluation questionnaire were collected, though some of these were duplicates, as teachers who participated in more than one training completed the evaluation for each training that they attended. The study included item means, item-to-total correlation, and Cronbach’s alpha for the questionnaire. Table 6 contains the demographic information collected from participants:

The majority of the responses collected were for participants in the online training: 189 responses (41.4%) were for the online self-paced training, while 149 responses (32.7%) were for the online synchronous training. The remaining 118 responses (25.9%) were for the in-person training.
Table 6
Training Evaluation Questionnaire Participant Demographics

<table>
<thead>
<tr>
<th>Role</th>
<th>Percent (Number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>College-level instructor</td>
<td>0.4% (2)</td>
</tr>
<tr>
<td>District-level administrator</td>
<td>5.0% (23)</td>
</tr>
<tr>
<td>Graduate student</td>
<td>0.7% (3)</td>
</tr>
<tr>
<td>Retired teacher</td>
<td>0.2% (1)</td>
</tr>
<tr>
<td>Charter school teacher</td>
<td>0.7% (3)</td>
</tr>
<tr>
<td>Public school teacher</td>
<td>92.8% (423)</td>
</tr>
<tr>
<td>Virtual school teacher</td>
<td>0.2% (1)</td>
</tr>
</tbody>
</table>

Test Items: Item Authoring and Review Data

Test items were developed online by the 127 participants from Sample 2, using the Florida Interim Assessment Item Bank and Test Platform (IBTP) for grant-funded courses. This software platform had the ability to record item-level data such as item writer, item reviewers, number of reviews, item type, standard assessed by the item, and cognitive complexity level, which was measured using Webb’s Depth of Knowledge. This research study used this item level data, available for 13,448 test items, to analyze item quality by writer, examining metrics such as number of reviews as an indicator of item quality (higher number of per-item reviews indicates items that were rejected back to the item writer more frequently by reviewers and thus required more editing before acceptance), and proportion of items that were accepted by the reviewers into the item bank, as a percentage of total number of items submitted by writers.

Test Items: Psychometric Data from Field Test

The student performance data for items developed by the 62 participants in Sample 3 came from field tests administered in a single Central Florida district. The district has more than 150 schools serving a student population of over 90,000. Fifty-six percent of students in the district received free or reduced lunch assistance. Eleven percent of students were classified as English Language Learners, while 12 percent of students have been identified as having a
disability. Five percent of students were identified as gifted. See Figure 4 for a detailed description of the student population, by race and ethnicity.

![Pie chart showing race and ethnicity distribution]

Figure 4. Student population race and ethnicity.

Field test forms were developed by the district’s curriculum specialists and district-level instructional coaches, using the item bank developed by the assessment collaborative project, and were piloted in May 2015. Field test forms were created and administered within the IBTP platform, which allows for a combination of online and paper-based administration, using optical scan features.

This research study analyzed item-level data where available (not all items were selected for inclusion on assessments), using Item Response Theory and values of the \( a \) parameter.

The sample of items for which student performance data is available is a total of 352 items, from 21 field test forms. The number of student responses available for each of the field test forms ranges from 190 to 4,952 responses. The mean number of student responses available
for these items is 1,180, with a standard deviation of 1,473. The median number of responses is 507. All of these items were written by teachers in Sample 3.

Data Analysis

This study employed a variety of data sources and data analysis methods. See Figure 5 for a summary of these sources and methods, by research question. For both types of surveys collected from the participants (application survey and professional development evaluation survey), the statistical software SPSS was used to derive descriptive statistics.

In order to estimate differences in item quality, descriptive and inferential statistics were calculated. Due to the nested nature of the study data (test items nested within teachers), a hierarchical linear model was used to analyze item quality differences, examining the statistical significance of the following teacher-level variables:

- Years of experience: 0-3, 4-6, 7-10, 11-15, 16-20, More than 20
- Degree type: Advanced degree, No advanced degree
- National Board certification: National Board certification, no National Board certification
- Mode of training: Online synchronous, online asynchronous, in person, blended online/in-person

The outcome/response variable was the quality of items created by teachers, as judged by item quality statistics as well as number of reviews required for each item (as an indicator of test item quality). A correlation matrix for teacher factors and delivery mode was first examined to explore the strength of the relationship among these variables (Table 7).

Item statistics were computed using 2-PL Item Response Theory (IRT) to calculate item discrimination. Item discrimination is a measure of how well an item differentiates among individuals of varying levels of the latent trait being measured. While this parameter can vary from $-\infty$ to $+\infty$, values that indicate a good quality item range from 0.8 to 2.5 (deAyala, 2009).
### Research Question | Data Sources | Data Analysis
--- | --- | ---
**Correlational-causal comparative/ex post facto**
RQ1: Is there any significant relationship between the following teacher-level factors: 1a. years of experience, 1b. prior assessment development experience, 1c. type of degree, and 1d. National Board certification and the participating teachers’ ability to write high-quality assessment items?  
**Dependent variables:** number of reviews per item, item discrimination (based on field test data)  
**Moderating variables:** 1a. Years of teaching experience (ordinal), 1b. prior assessment development experience (categorical), 1c. degree (categorical), and 1d. National Board certification (categorical)  
**Two-level hierarchical linear model (items nested within teachers) – means as outcomes model**
RQ2: What impact, if any, does the role selected by teachers (item writers vs. item writers/reviewers) have on teachers’ abilities to write high quality assessment items?  
**Dependent variables:** 2a. number of reviews per item 2b. proportion of accepted items  
**Independent variables:** teacher role (item writer or combination item writer/reviewer)  
**Descriptive and inferential statistics, two-level hierarchical linear model (For 2a, items nested within teachers) For 2b, multiple regression**
RQ3: What impact, if any, does the mode of training have on the quality of items written by teachers?  
**Dependent variables:** # 3a. total item reviews 3b. item discrimination 3c. number of accepted items as a proportion of total items written by the teachers  
**Independent variables:** type of training (online synchronous, online asynchronous, in-person, blended)  
**Descriptive and inferential statistics, two-level hierarchical linear model (For 3a and 3b, items nested within teachers) For 3c, multiple regression Will control for any variables found significant in RQ1 or RQ2**
RQ4: What do teachers report as the effects of participation in an assessment development project, and are there differences in these effects by mode of training?  
**Dependent variables:** Open-ended survey results  
**Descriptive and inferential statistics (Chi-square), constant comparative analysis**
RQ5: Is there a relationship between teachers’ perceptions of the assessment literacy training and the mode of training (in-person vs. online) they received?  
**Dependent variables:** Results of a training evaluation survey on a Likert scale, completed by participants after the training  
**Independent variables:** Mode of training  
**Descriptive and inferential statistics (ANOVA)**

*Figure 5. Evidence and analysis by research question.*

Negative discrimination values indicate that item respondents with lower levels of the latent trait have a higher probability of obtaining a correct response on the item than those with higher levels.
Research question four was answered using constant comparative analysis, to review teachers’ open-ended comments on the perceived impact of the item development training on their own classroom practices. After responses to the open-ended comments are coded, Chi Square analysis will be used to determine whether there are differences in these responses based on the mode of training in which the teacher participated.

Table 7

<table>
<thead>
<tr>
<th></th>
<th>Years Experience</th>
<th>Graduate Degree</th>
<th>NBCT</th>
<th>Prior assessment experience</th>
<th>Online Asynchronous</th>
<th>Online Synchronous</th>
<th>In Person</th>
<th>Blended</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pearson</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sig. (2-tailed)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
</tr>
<tr>
<td><strong>Years Experience</strong></td>
<td>.17**</td>
<td>.31</td>
<td>.31</td>
<td>-.18**</td>
<td>-.21***</td>
<td>.29**</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td><strong>Graduate Degree</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pearson</strong></td>
<td>.17**</td>
<td>1</td>
<td>.02</td>
<td>-.11</td>
<td>-.09</td>
<td>0.08</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sig. (2-tailed)</strong></td>
<td>&lt;.01</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
</tr>
<tr>
<td><strong>NBCT</strong></td>
<td>.31***</td>
<td>0.02</td>
<td>1</td>
<td>.11</td>
<td>-0.08</td>
<td>-0.01</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td><strong>Pearson</strong></td>
<td>&lt;.001</td>
<td>0.75</td>
<td>0.05</td>
<td>0.03</td>
<td>0.12</td>
<td>0.80</td>
<td>0.10</td>
<td>0.35</td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sig. (2-tailed)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
</tr>
<tr>
<td><strong>Prior assessment experience</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pearson</strong></td>
<td>.31***</td>
<td>-.11</td>
<td>.11</td>
<td>1</td>
<td>-0.10</td>
<td>-.23***</td>
<td>.45***</td>
<td>.12***</td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td>&lt;.001</td>
<td>0.05</td>
<td>0.03</td>
<td>0.07</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td><strong>Sig. (2-tailed)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
</tr>
<tr>
<td><strong>Online Asynchronous</strong></td>
<td>-.19***</td>
<td>-.09</td>
<td>-.08</td>
<td>-.10</td>
<td>1</td>
<td>-.18**</td>
<td>-.32**</td>
<td>.46***</td>
</tr>
<tr>
<td><strong>Pearson</strong></td>
<td>&lt;.001</td>
<td>0.09</td>
<td>0.12</td>
<td>0.07</td>
<td>&lt;.01</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sig. (2-tailed)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
</tr>
<tr>
<td><strong>Online Synchronous</strong></td>
<td>-.21***</td>
<td>0.08</td>
<td>-.01</td>
<td>-.23**</td>
<td>-.18**</td>
<td>1</td>
<td>-.68**</td>
<td>-.001</td>
</tr>
<tr>
<td><strong>Pearson</strong></td>
<td>&lt;.001</td>
<td>0.14</td>
<td>0.80</td>
<td>&lt;.001</td>
<td>&lt;.01</td>
<td>&lt;.001</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sig. (2-tailed)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
</tr>
<tr>
<td><strong>In Person</strong></td>
<td>.292</td>
<td>0.05</td>
<td>0.09</td>
<td>.45**</td>
<td>-.32**</td>
<td>-.68**</td>
<td>1</td>
<td>.23**</td>
</tr>
<tr>
<td><strong>Pearson</strong></td>
<td>&lt;.001</td>
<td>0.39</td>
<td>0.10</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sig. (2-tailed)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
</tr>
<tr>
<td><strong>Blended</strong></td>
<td>0.05</td>
<td>0.02</td>
<td>0.05</td>
<td>.118</td>
<td>.46**</td>
<td>-.001</td>
<td>.23**</td>
<td>1</td>
</tr>
<tr>
<td><strong>Pearson</strong></td>
<td>0.35</td>
<td>0.65</td>
<td>0.35</td>
<td>0.03</td>
<td>&lt;.001</td>
<td>0.98</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sig. (2-tailed)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
</tr>
</tbody>
</table>

* p < .05    ** p < .01    ***p < .001
Descriptive and inferential statistics (one-way ANOVA) will be used to answer research question five, to determine whether there is a relationship between teachers’ perceptions of the training, as evidenced by their total scores on the evaluation survey, and the mode of training.

**Hierarchical Linear Modeling**

Because of the nested nature of the data, with multiple items nested within a single writer, hierarchical linear modeling, also called multilevel modeling, was included as an analytic technique. This nested structure can lead to biased standard errors associated with the regression coefficients, when using statistical techniques that do not account for the non-independence of the data, such as ordinary least squares regression or analysis of variance (Hayes, 2006; O’Dwyer & Parker, 2014). Maas and Hox (2005) found that a level two sample size of 50 or less can lead to biased estimates of the second level standard errors. In this study, the level two sample size for models that use item field test data as the outcome variable is 62, while the level two sample size for models that use number of reviews by item as the outcome variable is 137, both of which are above the acceptable value.

**Item Response Theory**

Research questions 1 and 3 called for an indicator of item quality for field tested items, to be used as the outcome variable. Item response theory (IRT) was selected for this measure, since, unlike Classical Test Theory (CTT), IRT is sample invariant (Baker, 2001; Hambleton, Swaminathan, & Rogers, 1991). While in CTT, item difficulty is estimated as the percentage of respondents who answered an item correctly, in IRT, the parameters of the item are a property of the item itself, not the group of respondents. IRT was selected for use in this study, given the limitations of the sample of students used to collect field test data, since all students derived from a single school district.
A three-parameter model was originally selected for this project, since the three parameter model includes the \( a \) parameter, discrimination, the \( b \) parameter, difficulty, and the \( c \) parameter, guessing, which is the probability of responding correctly to the item by guessing alone. However, the three-parameter model did not fit the data sets for all field test forms and its use would have eliminated a large number of items from inclusion in the analysis. Because of this, a two-parameter (2PL) model was used for the final analysis. SAS software version 9.4 was used to calculate the IRT analysis.

**Constant Comparative Analysis**

Constant comparative analysis is a qualitative research technique used to reduce data through iterative coding and recoding (Fram, 2013). Data is compared to other collected data, and categories or codes are identified. The process is repeated over several iterations, with data further reduced over each iteration. Wilson Scott (2004) proposed a 6-question framework for constant comparative analysis when used during grounded theory:

- What is [the category]?
- When does [the category] occur?
- Where does [the category] occur?
- Why does [the category] occur?
- How does [the category] occur?
- With what consequence does [the category] occur or is [the category] understood?

Because this study is not a grounded theory study, and because of the limited nature of the data collected (open response items to four questions on a survey), the focus of this study will be on the first criteria, identification of categories within the participant data. This is most in
alignment with the research question, which asks what teachers report as the effects of participation in an assessment development project.
CHAPTER FOUR:

RESULTS

Research Question One

Hierarchical linear modeling was used to answer research question one, exploring the potential relationship between the following teacher-level factors:

1a. years of experience,
1b. prior assessment development experience,
1c. type of degree, and
1d. National Board certification

and the participating teachers’ ability to write high-quality assessment items. Factors 1a through 1d were used as level two predictors and two different metrics were used as outcome variables: item discrimination and number of reviews per item. SPSS software, version 23, was used for this analysis.

The correlation between these variables was examined (see correlation matrix, Table 7). The strongest relationships were found between years of teaching and other variables. Years of teaching experience was significantly correlated to all variables except for the blended mode of training. Teachers with more experience were more likely to participate in in-person training opportunities than in online training opportunities ($p < .001$), and were also more likely to have National Board certification ($p < .001$) and graduate level degrees ($p < .01$). While graduate degree was significantly positively correlated with years of experience ($p < .01$), interestingly, graduate degree had a significant negative correlation with prior assessment experience ($p = .05$),
meaning that those participants with graduate degrees were less likely to have reported having prior assessment development experience beyond that at the classroom level. In contrast, National Board certification was significantly positively correlated with both years of experience ($p < .001$) and prior assessment development experience ($p = .03$). Participation in the online synchronous training mode was significantly negatively correlated with prior assessment experience ($p < .001$), possibly because it was also negatively correlated with years of teaching experience ($p < .001$).

After examining the correlation matrix, the first outcome variable used to answer this research question was item discrimination. Using item response theory (IRT), a 2PL model was fit, and item discrimination/slope was calculated for each of the 352 items for which field test data was available. Appendix E contains these slope estimates for each of the field tested items. The mean item discrimination was 0.416, with a standard deviation of 0.691 (Table 8). Skewness and kurtosis values were within normal ranges. SAS software, version 9.4, was used for the IRT analysis.

The second outcome variable used for this research question was the number of reviews per item. Because items went through three levels of review, a lower number of reviews indicates that the item was accepted by all reviewers, while values above three mean that the item was rejected at least once. Of the 13,448 items for which these statistics were available, the mean number of reviews was 4.22 (see Table 8), with a standard deviation of 2.20.

**Outcome Variable: Item Discrimination/Slope Estimate**

Six means-as-outcomes models (Models 1 through 1e) were tested to address this research question. Between-writer variability in item discrimination values was independently modeled as a function of (a) years of teaching experience, (b) prior assessment experience
Table 8

*Descriptive Statistics for Number of Item Reviews and Discrimination*

<table>
<thead>
<tr>
<th></th>
<th>Number of item reviews</th>
<th>Item Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of items</td>
<td>13448</td>
<td>352</td>
</tr>
<tr>
<td>Mean</td>
<td>4.22</td>
<td>0.42</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.20</td>
<td>0.69</td>
</tr>
<tr>
<td>Skewness</td>
<td>3.18</td>
<td>0.02</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>15.02</td>
<td>2.64</td>
</tr>
</tbody>
</table>

beyond the classroom level, (c) prior assessment experience beyond the school level, (d) graduate degree, and (e) National Board certification.

**Model 1: Unconditional baseline model for item discrimination.**

A fully unconditional, baseline model was fit, using item discrimination as the outcome variable, to determine whether, without any predictors, item writers differed from each other, on average, on the discrimination of their items. This model was a random intercept-only model, with the fixed component \( \gamma_{00} \) and the random component \( u_{0j} \). The formula for this model is as follows:

Level 1: \( Y_{ij} = \beta_{0j} + r_{ij} \)

Level 2: \( \beta_{0j} = \gamma_{00} + u_{0j} \)

where \( Y_{ij} \) stands for item discrimination, \( \beta_{0j} \) is the average discrimination of items for writer \( j \), and \( r_{ij} \) is the difference between writer \( j \)’s item discrimination and the discrimination of item \( i \). \( \gamma_{00} \) is the grand mean in the level 2 model, while \( u_{0j} \) is the difference between item writer \( j \)’s average item discrimination and the grand mean. Table 9 contains the results of this model. The Intraclass Correlation Coefficient (ICC) was computed for this model based on the variance components; an ICC of near zero would indicate that a simpler technique than hierarchical linear modeling might be appropriate to use (Hayes, 2006). The 0.37 ICC indicated that item writers
accounted for 37 percent of the variance in item discrimination, and thus that hierarchical linear modeling was an appropriate technique to use to model this data.

**Model 1a: Between-writer differences in item discrimination based on years of teaching experience.**

Model 1a incorporated one additional variable at level 2, years of experience for item writers. This was an ordinal variable, with years of experience coded as follows:

- 0-3 years: 1
- 4-6 years: 2
- 7-10 years: 3
- 11-15 years: 4
- 16-20 years: 5
- More than 20 years: 6.

Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where an additional level 2 predictor, years of experience, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:

Level 1: \( Y_{ij} = \beta_{0j} + r_{ij} \)

Level 2: \( \beta_{0j} = \gamma_{00} + \gamma_{01} \text{YRSEXP} + u_{0j} \)

where YRSEXP is the item writer’s years of teaching experience and \( \gamma_{01} \) is the regression coefficient of YRSEXP. Based on the results of this model, which are shown in Table 9, the impact of years of teaching experience on item discrimination was not statistically different from zero, and thus was not included in subsequent models.
Model 1b: Between-writer differences in item discrimination based on prior assessment experience beyond classroom level.

Model 1b incorporated one additional variable at level 2, prior assessment experience for item writers. This was coded as prior assessment experience only at classroom level or any experience beyond classroom level, such as prior experience developing assessments at the school, district, or state level (1 = classroom only, 0 = beyond classroom). Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictor, prior assessment experience at the classroom level only, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:

Level 1: \( Y_{ij} = \beta_{0j} + r_{ij} \)

Level 2: \( \beta_{0j} = \gamma_{00} + \gamma_{02} \text{CLASSEXP} + u_{0j} \)

where CLASSEXP is the item writer’s prior experience at the classroom level only and \( \gamma_{02} \) is the regression coefficient of CLASSEXP. Based on the results of this model, shown in Table 9, the impact of having prior assessment development experience on item discrimination was not statistically different from zero, and thus was not included in subsequent models.

Model 1c: Between-writer differences in item discrimination based on prior assessment experience beyond school level.

Because prior assessment experience was measured at a variety of levels, Model 1c incorporated one additional variable at level 2, prior assessment experience beyond the school level for item writers. This was coded as prior assessment experience only at classroom or school level or any experience beyond the school level, such as prior experience developing assessments at the district or state level (1 = beyond school, 0 = school or classroom only).
Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictor, prior assessment experience beyond the school level, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:

Level 1: \( Y_{ij} = \beta_{0j} + r_{ij} \)

Level 2: \( \beta_{0j} = \gamma_{00} + \gamma_{03}\text{SCHOOLEXP} + u_{0j} \)

where SCHOOLEXP is the item writer’s prior experience beyond the school level and \( \gamma_{03} \) is the regression coefficient of SCHOOLEXP. Based on the results of this model, shown in Table 9, the impact of having prior assessment development experience beyond the classroom level on item discrimination was not statistically different from zero, and thus was not included in subsequent models.

Model 1d: Between-writer differences in item discrimination based on graduate degree.

Model 1d incorporated one additional variable at level 2, graduate degree, which included Master’s degrees, educational specialist degrees, or doctorate degrees. This was coded as 1= graduate degree, 0= no graduate degree. Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictor, graduate degree, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:

Level 1: \( Y_{ij} = \beta_{0j} + r_{ij} \)

Level 2: \( \beta_{0j} = \gamma_{00} + \gamma_{04}\text{GRADUATE} + u_{0j} \)

where GRADUATE is the item writer’s graduate degree status and \( \gamma_{04} \) is the regression coefficient of GRADUATE. Based on the results of this model, shown in Table 9, the impact of
having a graduate degree on item discrimination was not statistically different from zero, and thus was not included in subsequent models.

**Model 1e: Between-writer differences in item discrimination based on National Board certification.**

Model 1e incorporated one additional variable at level 2, National Board certification. This was coded as 1 = National Board certification, 0 = no National Board certification. Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictor, National Board, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:

Level 1: \( Y_{ij} = \beta_{0j} + r_{ij} \)

Level 2: \( \beta_{0j} = \gamma_{00} + \gamma_{05}\text{NBCT} + u_{0j} \)

where NBCT is the item writer’s National Board certification status and \( \gamma_{05} \) is the regression coefficient for NBCT. Based on the results of this model, shown in Table 9, the impact of having National Board certification on item discrimination was not statistically different from zero, and thus was not included in subsequent models.

**Outcome Variable: Number of Reviews by Item**

Several means-as-outcomes models were developed to address this research question. Between-writer variability in number of reviews by item was independently modeled as a function of (a) years of teaching experience, (b) prior assessment experience beyond the classroom level, (c) prior assessment experience beyond the school level, (d) graduate degree, and (e) National Board certification. A final model was then created that incorporated all variables found to have a significant impact on the between-writer variability in number of reviews by item.

68
<table>
<thead>
<tr>
<th>Table 9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameter Estimates for the Models Examining the Relationship</strong></td>
</tr>
<tr>
<td><strong>Between Item Writer Characteristics and Test Item Discrimination</strong></td>
</tr>
<tr>
<td>Model 1</td>
</tr>
<tr>
<td>Fixed effects</td>
</tr>
<tr>
<td>Intercept ($\gamma_{00}$)</td>
</tr>
<tr>
<td>Intercept SE</td>
</tr>
<tr>
<td>Years of teaching experience ($\gamma_{01}$)</td>
</tr>
<tr>
<td>Prior assessment experience at classroom level only ($\gamma_{02}$)</td>
</tr>
<tr>
<td>Any prior assessment experience beyond school level ($\gamma_{03}$)</td>
</tr>
<tr>
<td>Graduate degree = 0 ($\gamma_{04}$)</td>
</tr>
<tr>
<td>NBCT ($\gamma_{05}$)</td>
</tr>
<tr>
<td>ICC</td>
</tr>
<tr>
<td>Random Effects</td>
</tr>
<tr>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>$\tau_{00}$</td>
</tr>
<tr>
<td>Explained Variance</td>
</tr>
<tr>
<td>Deviance (-2LL)</td>
</tr>
</tbody>
</table>

* $p < .05$  ** $p < .01$  *** $p < .001$  + $p = .09$

*Note.* Explained variance is calculated by comparing each subsequent model to Model 1.
Model 1f: Unconditional baseline model for number of reviews by item.

A fully unconditional, baseline model was fit, using number of reviews by item as the outcome variable, to determine whether, without any predictors, item writers differed from each other, on average, on the number of reviews that their items received. This model was a random intercept-only model, with the fixed component $\gamma_{00}$ and the random component $u_{0j}$. The formula for this model is as follows:

**Level 1:** $Y_{ij} = \beta_{0j} + r_{ij}$

**Level 2:** $\beta_{0j} = \gamma_{00} + u_{0j}$

where $Y_{ij}$ stands for number of reviews by item, $\beta_{0j}$ is the average number of reviews by items for writer $j$, and $r_{ij}$ is the difference between writer $j$’s number of reviews by item and the number of reviews for item $i$. $\gamma_{00}$ is the grand mean in the level 2 model (average number of reviews by item overall), while $u_{0j}$ is the difference between item writer $j$’s average number of reviews by item and the grand mean. Table 10 contains the results of this model. The Intraclass Correlation Coefficient (ICC) was computed for this model based on the variance components. The 0.25 ICC indicated that item writers accounted for 25 percent of the variance in number of reviews by item, and thus that hierarchical linear modeling was an appropriate technique to use.

Model 1g: Between-writer differences in number of reviews by item based on years of experience.

Model 1g incorporated one additional variable at level 2, years of experience for item writers. This was an ordinal variable, with years of experience coded exactly as they were in Model 1a. Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where an additional level 2 predictor, years of experience, was included to determine...
whether the unexplained variance would be reduced. The formula for this model is shown below:

Level 1: \( Y_{ij} = \beta_{0j} + r_{ij} \)

Level 2: \( \beta_{0j} = \gamma_{00} + \gamma_{01} \text{YRSEXP} + u_{0j} \)

where \( \text{YRSEXP} \) is the item writer’s years of teaching experience and \( \gamma_{01} \) is the regression coefficient for \( \text{YRSEXP} \). Based on the results of this model, which are shown in Table 10, the impact of years of teaching experience on the number of reviews per item was not statistically different from zero, and thus was not included in subsequent models.

**Model 1h: Between-writer differences in number of reviews by item based on prior assessment experience only at classroom level.**

Model 1h incorporated one additional variable at level 2, prior assessment experience for item writers. This was coded as prior assessment experience only at classroom level or any experience beyond classroom level, such as prior experience developing assessments at the school, district, or state level (1 = classroom only, 0 = beyond classroom). Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictor, prior assessment experience at the classroom level only, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:

Level 1: \( Y_{ij} = \beta_{0j} + r_{ij} \)

Level 2: \( \beta_{0j} = \gamma_{00} + \gamma_{02} \text{CLASSEXP} + u_{0j} \)

where \( \text{CLASSEXP} \) is the item writer’s prior experience at the classroom level only and \( \gamma_{02} \) is the regression coefficient for \( \text{CLASSEXP} \). Based on the results of this model, shown in Table 10, the impact of having prior assessment development experience on item discrimination was
statistically different from zero ($p = .03$). Items written by participants who had prior assessment experience at the classroom level only had, on average, 0.47 more reviews per item than those written by participants who had prior assessment experience beyond the classroom level. This indicates that items written by participants with prior assessment experience at the classroom level only were rejected more frequently than items written by participants who had prior assessment development experience at the school, district, state, or vendor level.

**Model 1i: Between-writer differences in number of reviews by item based on prior assessment experience beyond the school level.**

Because prior assessment experience was measured at a variety of levels, Model 1i incorporated one additional variable at level 2, prior assessment experience beyond the school level for item writers. This was coded as prior assessment experience only at classroom or school level vs. any experience beyond the school level, such as prior experience developing assessments at the district or state level (1 = beyond school, 0 = school or classroom only). Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictor, prior assessment experience beyond the school level, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:

Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{03}\text{SCHOOLEXP} + u_{0j}$

where SCHOOLEXP is the item writer’s prior experience beyond the school level and $\gamma_{03}$ is the regression coefficient for SCHOOLEXP. Based on the results of this model, shown in Table 10, the impact of having prior assessment development experience beyond the classroom level on
number of reviews per item was not statistically different from zero, and thus was not included in subsequent models.

**Model 1j: Between-writer differences in number of reviews by item based on graduate degree.**

Model 1j incorporated one additional variable at level 2, graduate degree, which included Master’s degrees, educational specialist degrees, or doctorate degrees. This was coded as 1 = graduate degree, 0 = no graduate degree. Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictor, graduate degree, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:

**Level 1:** \( Y_{ij} = \beta_{0j} + r_{ij} \)

**Level 2:** \( \beta_{0j} = \gamma_{00} + \gamma_{04}\text{GRADUATE} + u_{0j} \)

where \( \text{GRADUATE} \) is the item writer’s graduate degree status and \( \gamma_{04} \) is the regression coefficient for \( \text{GRADUATE}. \) Based on the results of this model, shown in Table 10, the impact of having a graduate degree on number of reviews per item was not statistically different from zero, and thus was not included in subsequent models.

**Model 1k: Between-writer differences in number of reviews by item based on National Board certification.**

Model 1e incorporated one additional variable at level 2, National Board certification. This was coded as 1 = National Board certification, 0 = no National Board certification. Slope estimates were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictor, National Board, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:
Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{05}\text{NBCT} + u_{0j}$

where NBCT is the item writer’s National Board certification status and $\gamma_{05}$ is the regression coefficient for NBCT. Based on the results of this model, shown in Table 10, the impact of having National Board certification on number of reviews per item was not statistically different from zero, and thus was not included in subsequent models.

Research Question Two

Research question two asked “What impact, if any, does the role selected by teachers (item writers vs. item writers/reviewers) have on teachers’ abilities to write high quality assessment items?”

Outcome Variable: Number of Reviews by Item

The first part of this research question was answered by hierarchical linear modeling, since items were nested within teachers, using the outcome variable number of reviews per item as an indicator of number of times the item was rejected, since items with more than three reviews were those that were rejected at least once.

A model was run using item writer as the level two clustering variable. Role (item writer or item writer/reviewer) was used as a level two variable with fixed effects. Table 9 shows the results of models 2a and 2b, described below.

Model 2a: Between-writer differences in number of reviews by item based on participant role.

Model 2a incorporated one additional variable at level 2, participant role (writer/reviewer vs. writer only). This was coded as $1 = \text{writer and reviewer, 0 = writer only}$. Intercept estimates
Table 10  
*Parameter Estimates for the Models Examining the Relationship Between Item Writer Characteristics and the Number of Reviews for Each Item Written by the Item Writer*

<table>
<thead>
<tr>
<th></th>
<th>Model 1f</th>
<th>Model 1g</th>
<th>Model 1h</th>
<th>Model 1i</th>
<th>Model 1j</th>
<th>Model 1k</th>
<th>Model 2a</th>
<th>Model 2b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ($\gamma_{00}$)</td>
<td>4.05***</td>
<td>3.87***</td>
<td>3.80***</td>
<td>4.18***</td>
<td>3.89***</td>
<td>4.02***</td>
<td>4.74***</td>
<td>4.49***</td>
</tr>
<tr>
<td>Intercept SE</td>
<td>0.11</td>
<td>0.31</td>
<td>0.16</td>
<td>0.13</td>
<td>0.19</td>
<td>0.12</td>
<td>0.24</td>
<td>0.28</td>
</tr>
<tr>
<td>Years of teaching experience ($\gamma_{01}$)</td>
<td></td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior assessment experience at classroom level only ($\gamma_{02}$)</td>
<td></td>
<td>0.47*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior assessment experience beyond school level ($\gamma_{03}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graduate degree ($\gamma_{04}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>National Board certification ($\gamma_{05}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Role (writer vs. writer/reviewer) ($\gamma_{06}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>3.89***</td>
<td>3.90***</td>
<td>3.89***</td>
<td>3.89***</td>
<td>3.89***</td>
<td>3.89***</td>
<td>3.89***</td>
<td></td>
</tr>
<tr>
<td>$\tau_{00}$</td>
<td>1.31***</td>
<td>1.33***</td>
<td>1.27***</td>
<td>1.29***</td>
<td>1.31***</td>
<td>1.32***</td>
<td>1.19***</td>
<td>1.17***</td>
</tr>
</tbody>
</table>

| Deviance (-2LL)      | 56796.27 | 56761.14 | 56792.87 | 56794.53 | 56796.45 | 56796.36 | 56787.44 | 56785.87 |
| Explained Variance   | 0.01     | 0.03     | 0.02     | 0.00     | 0.01     | 0.10     | 0.11     |

* $p < .05$  ** $p < .01$  *** $p < .001$  +$p = .09$

*Note. Explained variance is calculated by comparing each subsequent model to Model 1f.*
were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictor, participant role, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:

Level 1: \( Y_{ij} = \beta_{0j} + r_{ij} \)

Level 2: \( \beta_{0j} = \gamma_{00} + \gamma_{06}\text{ROLE} + u_{0j} \)

where \( \text{ROLE} \) is the participant’s role and \( \gamma_{06} \) is the regression coefficient for \( \text{ROLE} \). Based on the results of this model, shown in Table 9, the impact of serving as both an item writer and reviewer, as opposed to an item writer only, on the number of reviews per item was statistically different from zero (\( p < .01 \)). Serving as both a writer and a reviewer was associated with a decrease in the number of reviews per item, indicating that items written by participants who served as both item writers and reviewers had approximately 0.85 fewer reviews than items written by participants who served as item writers only. This finding means that these items were rejected less frequently than items written by participants who served as item writers only.

**Model 2b: Between-writer differences in number of reviews by item based on participant role and prior assessment development experience.**

Model 2b incorporated two variables at level 2, participant role (writer/reviewer vs. writer only) and prior assessment experience at the classroom level only, since prior assessment experience was found to be significant in Model 1h. Role was coded as 1 = writer and reviewer, 0 = writer only, while prior assessment experience was coded as 1 = classroom experience only, 0 = experience beyond the classroom level. Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictors, participant role and prior assessment experience, were included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:
Level 1: \( Y_{ij} = \beta_{0j} + r_{ij} \)

Level 2: \( \beta_{0j} = \gamma_{00} + \gamma_{06}\text{ROLE} + \gamma_{02}\text{CLASSEXP} + u_{0j} \)

where \( \text{ROLE} \) is the participant’s role, \( \gamma_{06} \) is the regression coefficient for \( \text{ROLE} \), and \( \gamma_{02} \) is the regression coefficient for \( \text{CLASSEXP} \). Based on the results of this model, shown in Table 9, the impact of serving as both an item writer and reviewer, as opposed to an item writer only, on the number of reviews per item was statistically different from zero \( (p < .01) \). Serving as both a writer and a reviewer was associated with a 0.77 decrease in the number of reviews per item, indicating that items written by participants who served as both item writers and reviewers had a significantly lower number of reviews per item. These items were rejected less frequently than items written by participants who served as item writers only.

After controlling for participant role, prior assessment experience beyond the classroom level was no longer found to have a significant relationship with the number of reviews per item \( (p = .09) \).

**Outcome Variable: Proportion of Accepted Items**

The second part of research question two called for an analysis of the relationship between participant role (item writer only vs. item writer and reviewer) and the proportion of accepted items by the participant. The mean percentage of accepted items for all 127 Sample 2 participants was .92, with a standard deviation of .19. The outcome variable, proportion of accepted items, was regressed against the categorical variable participant role. Tables 11 and 12 show the results of this regression. This first model, Model 2c, showed that participant role accounted for 3.7 percent of the variance in proportion of accepted items. While participant role
was a significant predictor of the proportion of accepted items, $F(1,125) = 5.79, p = .02$, it only accounted for a small percentage of the variance.

Because prior assessment experience beyond that of the classroom only was shown in research question one to be a significant predictor of the number of reviews that an item would be given, a second model, Model 2d, was run including both role and prior assessment experience as variables, to determine whether this variable had a significant impact on the proportion of accepted items by each writer. Both role and prior assessment experience combined accounted for 4.6 percent of the variance in proportion of accepted items. As with model 2b, only role was found to be a significant predictor ($p = .01$), while prior assessment development experience beyond the classroom level was not found to be significant ($p = .14$).

Table 11
Multiple Regression Model Summary of Proportion of Accepted Items as an Outcome Variable

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Standard Error</th>
<th>F</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2c</td>
<td>.21</td>
<td>.04</td>
<td>.04</td>
<td>.19</td>
<td>5.79</td>
<td>.02</td>
</tr>
<tr>
<td>2d</td>
<td>.25</td>
<td>.06</td>
<td>.05</td>
<td>.19</td>
<td>4.05</td>
<td>.02</td>
</tr>
</tbody>
</table>

a. Predictor: (Constant), Role
b. Predictors: (Constant), Role, Prior Assessment Experience

Table 12
Coefficients for Multiple Regression Models of Proportion of Accepted Items as an Outcome Variable

<table>
<thead>
<tr>
<th>Model</th>
<th>Beta</th>
<th>t</th>
<th>Significance</th>
<th>Standard Error</th>
<th>Lower Bound 95% Confidence Interval</th>
<th>Upper Bound 95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>2c (Constant)</td>
<td>.84</td>
<td>23.43</td>
<td>.00</td>
<td>.04</td>
<td>.77</td>
<td>.92</td>
</tr>
<tr>
<td>2c (Role)</td>
<td>.10</td>
<td>2.41</td>
<td>.02</td>
<td>.04</td>
<td>.02</td>
<td>.18</td>
</tr>
<tr>
<td>2d (Constant)</td>
<td>.81</td>
<td>19.20</td>
<td>.00</td>
<td>.04</td>
<td>.73</td>
<td>.89</td>
</tr>
<tr>
<td>2d (Role)</td>
<td>.11</td>
<td>2.62</td>
<td>.01</td>
<td>.04</td>
<td>.03</td>
<td>.19</td>
</tr>
<tr>
<td>2d (Experience)</td>
<td>.05</td>
<td>1.50</td>
<td>.14</td>
<td>.03</td>
<td>-.02</td>
<td>.12</td>
</tr>
</tbody>
</table>
Research Question Three

Research question 3 asked, “What impact, if any, does the mode of training have on the quality of items written by teachers?” In order to answer this question, the mode of training was examined in relationship to three different outcome variables: number of reviews by item, as an indicator of the number of times the item was rejected by reviewers, item discrimination, and percentage of accepted items.

Outcome Variable: Number of Reviews by Item

Model 3a: Between-writer differences in number of reviews by item based on training type.

Model 3a used hierarchical linear modeling to incorporate mode of training as a level 2 variable. This was coded as OnlineSync = online synchronous training, Online Async = online asynchronous training, Blended = blended training, and F2F = face to face training. The number of reviews by item was the outcome variable for this model. Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictor, training type, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:

Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}\text{BLEND} + \gamma_{02}\text{INPERSON} + \gamma_{03}\text{ONLINEASYNC} + u_{0j}$

where BLEND indicates participation in the blended mode of training, $\gamma_{01}$ is the regression coefficient of BLEND, INPERSON indicates participation in the in-person mode of training, $\gamma_{02}$ is the regression coefficient of INPERSON, ONLINEASYNC indicates participation in the online asynchronous mode of training, and $\gamma_{03}$ is the regression coefficient for ONLINEASYNC.
Based on the results of this model, shown in Table 13, the impact of mode of training on the number of reviews per item was not statistically different from zero.

**Outcome Variable: Item Discrimination**

**Model 3b: Between-writer differences in item discrimination based on training type.**

Model 3b was a hierarchical linear model that incorporated mode of training as a variable at level 2. This was coded as described above for model 3a. The item discrimination/slope estimate was the outcome variable for this model. Intercept estimates were allowed to vary across teachers in this means-as-outcomes model, where the level 2 predictor, mode of training, was included to determine whether the unexplained variance would be reduced. The formula for this model is shown below:

**Level 1:** \[ Y_{ij} = \beta_{0j} + r_{ij} \]

**Level 2:** \[ \beta_{0j} = \gamma_{00} + \gamma_{01}\text{BLEND} + \gamma_{02}\text{INPERSON} + \gamma_{03}\text{ONLINEASYNC} + u_{0j} \]

where BLEND indicates participation in the blended mode of training, \( \gamma_{01} \) is the regression coefficient of BLEND, INPERSON indicates participation in the in-person mode of training, \( \gamma_{02} \) is the regression coefficient of INPERSON, ONLINEASYNC indicates participation in the online asynchronous mode of training, and \( \gamma_{03} \) is the regression coefficient for ONLINEASYNC. Based on the results of this model, shown in Table 14, the impact of mode of training on the item discrimination was not statistically different from zero.

**Outcome Variable: Proportion of Accepted Items**

Research question three also called for an analysis of the relationship between mode of training and the proportion of accepted items by the participant. The outcome variable,
proportion of accepted items, was regressed against the categorical variable mode of training.

Tables 15 and 16 show the results of this regression. This first model, Model 3f, showed that

Table 13
*Parameter Estimates for the Model Examining the Relationship Between Mode of Training and the Number of Reviews by Item for Each Item Written by the Item Writer*

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Model 3a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ($\gamma_{00}$)</td>
<td>4.03***</td>
</tr>
<tr>
<td>Intercept SE</td>
<td>0.20</td>
</tr>
<tr>
<td>Training type: Blended ($\gamma_{01}$)</td>
<td>0.07</td>
</tr>
<tr>
<td>Training type: In person ($\gamma_{02}$)</td>
<td>-0.16</td>
</tr>
<tr>
<td>Training type: Online Asynchronous ($\gamma_{03}$)</td>
<td>0.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>$\tau_{00}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Explained Variance and Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance (-2LL)</td>
</tr>
<tr>
<td>Variance</td>
</tr>
</tbody>
</table>

* $p < .05$ ** $p < .01$ ***$p < .001$ $^p=.09$

Note. Explained variance is calculated by comparing Model 3a to Model 1f (baseline model).

Table 14
*Parameter Estimates for the 2 Models Examining the Relationship Between Mode of Training and the Item Discrimination for Each Item Written by the Item Writer*

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Model 3b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ($\gamma_{00}$)</td>
<td>0.21</td>
</tr>
<tr>
<td>Intercept SE</td>
<td>0.15</td>
</tr>
<tr>
<td>Training type: Blended ($\gamma_{01}$)</td>
<td>-0.16</td>
</tr>
<tr>
<td>Training type: In person ($\gamma_{02}$)</td>
<td>0.29</td>
</tr>
<tr>
<td>Training type: Online Asynchronous ($\gamma_{03}$)</td>
<td>0.01</td>
</tr>
<tr>
<td>Prior Experience</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>$\tau_{00}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Explained Variance and Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance</td>
</tr>
<tr>
<td>Deviance (-2LL)</td>
</tr>
</tbody>
</table>

* $p < .05$ ** $p < .01$ ***$p < .001$ $^p=.09$

Note. Explained variance is calculated by comparing Model 3b to Model 1 (baseline model).
mode of training accounted for less than one percent of the variance in proportion of accepted items. Additionally, mode of training was not a significant predictor of the proportion of accepted items, $F(1,125) = 0.44, p = .51$.

<table>
<thead>
<tr>
<th>Model</th>
<th>$R$</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Standard Error</th>
<th>$F$</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>3f</td>
<td>.06</td>
<td>.00</td>
<td>.00</td>
<td>.19</td>
<td>.44</td>
<td>.51</td>
</tr>
</tbody>
</table>

a. Predictor: (Constant), Role

<table>
<thead>
<tr>
<th>Model</th>
<th>Beta</th>
<th>t</th>
<th>Significance</th>
<th>Standard Error</th>
<th>Lower Bound 95% Confidence Interval</th>
<th>Upper Bound 95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Constant)</td>
<td>.95</td>
<td>20.45</td>
<td>.00</td>
<td>.05</td>
<td>0.86</td>
<td>1.04</td>
</tr>
<tr>
<td>1 (Mode of Training)</td>
<td>-.01</td>
<td>-.66</td>
<td>.51</td>
<td>.02</td>
<td>-.04</td>
<td>.02</td>
</tr>
</tbody>
</table>

**Research Question Four**

Research question four asked what teachers reported as the effects of participation in an assessment development project and whether these reported effects differed based on the mode of training in which teachers participated. In order to answer this, open-ended responses to the training evaluation survey were used. Of the 419 participants that composed Sample 3, there were 456 responses to the training evaluation survey. There is a greater number of responses than participants because participants were asked to complete the evaluation survey after every training that they completed. Participants who completed both an online and an in-person training were thus asked to complete the evaluation survey for both trainings. Responses were anonymous, so there was no way to tell from the data set which responses were submitted by the
same person. Another result of the anonymity was that there was no way to tell which responses were made by participants who had completed a different training previously. Because of this, responses are coded as only three modes: online asynchronous, online synchronous, and in-person.

Of the 456 responses to the training evaluation survey, 136 were omitted from the analysis for research question four because they skipped the relevant questions (numbers 15 through 18, found in Appendix D), which were optional, open-response items. The remaining 320 responses to questions 15 through 18 were analyzed in order to answer this question. Of these 320 responses, 88 were evaluating an in-person training, 132 were evaluating the online asynchronous training, and the remaining 100 were evaluating the online synchronous training.

The constant comparative method was used for analyzing this data, which is an iterative process. Individual responses to each of the four open-response items were examined discretely, then in relation to other responses and to other items. The result was nine themes that emerged from the responses:

- Insight into standardized assessments and preparation for these assessments
- Improved knowledge of course standards or content
- Collaboration with other educators
- Cognitive complexity and rigor
- Alignment of assessment and instruction
- Insight into issues of bias and sensitivity as related to assessment
- Insight into measurement issues of reliability or validity as related to assessment
- Insight into formatting and mechanics of assessments
- Insight into development of rubrics for constructed response items or performance tasks
Table 17 shows the percent of responses that related to each of these nine themes.

**Table 17**

**Percent of Responses Related to Each Theme**

<table>
<thead>
<tr>
<th>Theme</th>
<th>Overall Mean (N)</th>
<th>In-Person Mean (N)</th>
<th>Online Asynchronous Mean (N)</th>
<th>Online Synchronous Mean (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theme 1: Insight into standardized assessments</td>
<td>.11 (34)</td>
<td>.10 (9)</td>
<td>.08 (10)</td>
<td>.15 (15)</td>
</tr>
<tr>
<td>Theme 2: Knowledge of course standards/content</td>
<td>.14 (46)</td>
<td>.21 (18)</td>
<td>.14 (19)</td>
<td>.09 (9)</td>
</tr>
<tr>
<td>Theme 3: Collaboration with others</td>
<td>.14 (46)</td>
<td>.26 (23)</td>
<td>.10 (13)</td>
<td>.10 (10)</td>
</tr>
<tr>
<td>Theme 4: Cognitive complexity and rigor</td>
<td>.30 (96)</td>
<td>.24 (21)</td>
<td>.26 (34)</td>
<td>.41 (41)</td>
</tr>
<tr>
<td>Theme 5: Alignment of assessment &amp; instruction</td>
<td>.07 (22)</td>
<td>.13 (11)</td>
<td>.07 (9)</td>
<td>.02 (2)</td>
</tr>
<tr>
<td>Theme 6: Bias and sensitivity</td>
<td>.08 (27)</td>
<td>.03 (3)</td>
<td>.14 (19)</td>
<td>.05 (5)</td>
</tr>
<tr>
<td>Theme 7: Reliability and validity</td>
<td>.03 (10)</td>
<td>.00 (0)</td>
<td>.05 (7)</td>
<td>.03 (3)</td>
</tr>
<tr>
<td>Theme 8: Formatting and mechanics</td>
<td>.09 (28)</td>
<td>.07 (6)</td>
<td>.10 (13)</td>
<td>.09 (9)</td>
</tr>
<tr>
<td>Theme 9: Rubric development</td>
<td>.07 (23)</td>
<td>.18 (16)</td>
<td>.05 (6)</td>
<td>.01 (1)</td>
</tr>
</tbody>
</table>

**Theme 1: Insight into Standardized Assessments and Preparation for these Assessments**

The first theme found in participants’ responses was that participation in the assessment development project provided them with additional insight into state or other standardized assessments, as well as how to prepare their students for these assessments. Ten percent of participants, 34 respondents, provided feedback to this respect. Mention of this theme varied by training type and ranged from 15 percent (online synchronous participants) to 7.6 percent (online asynchronous participants). There were no significant differences between the group responses as determined by a chi-square test of homogeneity. Participant comments included responses such as the following:

- “It will help me to write assessments which will parallel the types of questions that my students will be seeing on their EOCs next year. I will be looking at the requirements for writing questions for the State-required EOCs as I write questions for my students.”
• “It helps me better prepare my students by knowing how EOCs are created.”
• “Will need to prepare students for testing and this is a great way to understand the process”
• “I definitely will have a better idea of the types of questions that will be on various exams that my future students will be taking.”

**Theme 2: Improved Knowledge of Course Content or Standards**

The second theme from participants’ evaluation of the training was that fourteen percent (46 participants) felt that they gained an improved knowledge of their course content or course standards through participation. This theme varied across modes of training, with 20.5 percent of in-person participants reporting this as an impact, while 14.4 percent of online asynchronous and 9 percent of online synchronous participants gave responses that showed evidence of this theme. There were no significant differences between the group responses as determined by a chi-square test of homogeneity. Participant feedback that was coded under this category included responses such as the following:

• “It allowed me to be better prepared as a teacher in the content areas that I teach.”
• “Yes, better understanding of benchmarks. Clearer understanding of what I need to teach and cover.”
• “I will be more effective in meeting the standards.”
• “It will make me more aware of the content.”
• “I understand the benchmarks better and to teach the material differently.”
• “I am much more familiar with the benchmarks of the class I am teaching.”
This workshop will help me in the classroom as I assess the benchmarks and standards throughout the year. It also gave me a deeper understanding of the standards by which I assess my students."

“Being familiar with benchmarks will enhance my teaching focus and student success.”

**Theme 3: Collaboration with Other Educators**

The third theme found in participants’ responses was that of collaboration with other educators. Fourteen percent of participants overall (46 participants) mentioned this in their feedback about project participation. There were significant differences between the group responses as determined by a chi-square test of homogeneity, $\chi^2(2, N = 320) = 13.64, p = .001$, as shown in Table 18.

<table>
<thead>
<tr>
<th>Mode of Training</th>
<th>Mentioned Collaboration</th>
<th>Did Not Mention Collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Person</td>
<td>23 (26.1%)</td>
<td>65 (73.9%)</td>
</tr>
<tr>
<td>Online Asynchronous</td>
<td>13 (9.9%)</td>
<td>119 (90.2%)</td>
</tr>
<tr>
<td>Online Synchronous</td>
<td>10 (10.0%)</td>
<td>90 (90.0%)</td>
</tr>
</tbody>
</table>

**Note.** $\chi^2 = 13.64^*, df = 2$. Numbers in parentheses indicate column percentages.

*p = .001

Twenty three participants, or twenty-six percent of participants in the in-person training ($M = 0.26, SD = 0.44$) had responses that had the theme of collaboration with other educators, as opposed to ten percent of participants in the online synchronous ($M = 0.10, SD = 0.30$) and asynchronous ($M = 0.10, SD = 0.30$) trainings.

Using the standardized residual method described by Beasley & Schumacker, 1995, to conduct post-hoc analysis on the chi-square test results, the in-person mode of training group was found to differ significantly at $p = .001$ from the two online training groups (see Table 19).
The formula $\alpha_{\text{adj}} = 1 - (1 - \alpha)^{1/s}$, where $s$ was equal to the number of tests, was used to calculate the $\alpha_{\text{adj}} = .02$. 

Table 19

<table>
<thead>
<tr>
<th>Mode of Training</th>
<th>Adjusted Standardized Residual, Did Not Mention Collaboration</th>
<th>Adjusted Standardized Residual, Did Mention Collaboration</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Person</td>
<td>-3.69</td>
<td>3.69</td>
<td>.001</td>
</tr>
<tr>
<td>Online Asynchronous</td>
<td>1.93</td>
<td>-1.93</td>
<td>.16</td>
</tr>
<tr>
<td>Online Synchronous</td>
<td>1.50</td>
<td>-1.50</td>
<td>.33</td>
</tr>
</tbody>
</table>

Examples of responses that were classified as indicating collaboration with other educators include the following:

- “As a leader at professional development I will try to share the strategies I have learned with the faculty and improve test writing across the school.”
- “As teachers seek advice in assessments, this will help me make sure that the assessments they are creating are useful and reliable indicators of student learning.”
- “I plan to teach thru Staff Development for other teachers how to develop better assessments for students.”
- “I will definitely share information in this training during PD workshops at my school level.”
- “I will be assisting others at the district level write test items for the district, my … question writing/reviewing skills will come in handy!”
- “Will present mini sessions on evaluating and writing improved assessment items”
• “I chair the social studies department so I will share this information with my peers at our meetings.”

**Theme 4: Cognitive Complexity and Rigor**

The fourth theme found in participants’ responses was that of cognitive complexity and rigor. Thirty percent of participants (96 people) mentioned this theme as one of the impacts of participation on their practices. There were significant differences between the group responses as determined by a chi-square test of homogeneity, $X^2(2, N = 320) = 8.47, p = .01$, as shown in Table 20.

Table 20  
*Chi-square Test for Cognitive Complexity and Rigor*

<table>
<thead>
<tr>
<th>Mode of Training</th>
<th>Mentioned Cognitive Complexity</th>
<th>Did Not Mention Cognitive Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Person</td>
<td>21 (23.9%)</td>
<td>67 (76.1%)</td>
</tr>
<tr>
<td>Online Asynchronous</td>
<td>34 (25.8%)</td>
<td>98 (74.2%)</td>
</tr>
<tr>
<td>Online Synchronous</td>
<td>41 (41.0%)</td>
<td>59 (59.0%)</td>
</tr>
</tbody>
</table>

*Note. $X^2 = 8.47^*$, df = 2. Numbers in parentheses indicate column percentages.  
*$p = .01$*

In a post-hoc analysis of the chi-square test results, shown in Table 21, the online synchronous mode of training group was found to differ significantly at $p = .02$ from the other two training groups.

Forty-one percent of participants in the online synchronous training (41 participants) had responses that addressed the theme of cognitive complexity and rigor, as opposed to 23.9 percent of participants in the in-person and 25.8 percent of participants in the online asynchronous trainings. Examples of responses that were classified as indicative of the impact of participation on participants’ understanding and use of cognitive complexity and rigor include the following:
Table 21

Results of Chi-square Test Post-hoc Analysis Using Standardized Residuals, Cognitive Complexity and Rigor

<table>
<thead>
<tr>
<th>Mode of Training</th>
<th>Adjusted Standardized Residual, Did Not Mention Cognitive Complexity</th>
<th>Adjusted Standardized Residual, Did Mention Cognitive Complexity</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Person Online</td>
<td>1.48</td>
<td>-1.48</td>
<td>.33</td>
</tr>
<tr>
<td>Online Asynchronous</td>
<td>1.39</td>
<td>-1.39</td>
<td>.38</td>
</tr>
<tr>
<td>Online Synchronous</td>
<td>-2.89</td>
<td>2.89</td>
<td>.02</td>
</tr>
</tbody>
</table>

- “The activity that asked us to rewrite a questions to increase the complexity of it was of greatest value because I feel like I will use this strategy to make sure I am challenging my students.”
- “Just re-wording ques. [sic.] can make them more complex, therefore I hope to implement that in my assessments.”
- “Differentiate the difference between difficulty and complexity, and how to incorporate into my assessment items.”
- “I will write questions that are more relevant to student demonstration of understanding and stay away from marketed banks that don't really assess learning. I will look at complexity to give me more evidence of student thinking process.”
- “The training helps me to write better assessments for my students and requires them to use more higher order thinking skills”
- “I will make sure to incorporate different levels of questioning in my classroom.”
- “Teachers, for the most part, ask only low-level questions on their assessments. I might be able to help them evaluate their questions and tests... I probably will stress Webb's DOK a bit more.”
• “I plan to re evaluate my old tests to make sure of the DOK and the complexity of the questions.”

**Theme 5: Alignment of Assessment and Instruction**

The fifth theme found in participants’ responses was that of an improvement in the alignment between their assessment and instruction. Twenty-two participants (6.9 percent) mentioned this theme as one of the impacts of participation on their practices.

There were significant differences between the group responses as determined by a chi-square test of homogeneity, $\chi^2(2, N = 320) = 8.06, p = .012$, as shown in Table 22.

<table>
<thead>
<tr>
<th>Mode of Training</th>
<th>Mentioned Alignment</th>
<th>Did Not Mention Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Person</td>
<td>11 (12.5%)</td>
<td>77 (87.5%)</td>
</tr>
<tr>
<td>Online Asynchronous</td>
<td>9 (6.8%)</td>
<td>123 (93.2%)</td>
</tr>
<tr>
<td>Online Synchronous</td>
<td>2 (2.0%)</td>
<td>98 (98.0%)</td>
</tr>
</tbody>
</table>

*Note. $\chi^2 = 8.06^*, df = 2$. Numbers in parentheses indicate column percentages.

There were no significant differences between the group responses as determined by a post hoc analysis of the standardized residuals. Eleven participants in the in-person training (12.5 percent) had responses that addressed the theme of alignment of instruction and assessment, as opposed to nine participants (6.8 percent) in the online asynchronous training and only two participants (2 percent) in the online synchronous training. Examples of responses that were classified as indicative of the impact of participation on participants’ alignment of assessment and instruction include the following:

• “My assessment practices will be stronger, thereby, providing more accurate data in which to align my instruction.”
• “I will be more aware of instruction and aligning assessment with that instruction. Performance data will be used more efficiently to help with this process.”
• “Now more than ever is it relevant to teach close to the standards and prepare students for assessments. My local/classroom level assessments will be more aligned with standards.”
• “I believe this workshop will allow me to plan my instruction with my assessments in mind first. It has also shown me the importance of aligning my assessment directly to the benchmarks and standards in order to better facilitate instruction.”

**Theme 6: Insight into Issues of Bias and Sensitivity**

The sixth theme found in participants’ responses was that of increased insight into issues of bias and sensitivity, relative to assessment practices. Twenty-seven participants (8.4 percent) mentioned this theme as one of the impacts of participation on their practices.

There were significant differences between the group responses as determined by a chi-square test of homogeneity, \( \chi^2 (2, N = 320) = 10.47, p = .01 \), as shown in Table 23.

<table>
<thead>
<tr>
<th>Mode of Training</th>
<th>Mentioned Bias/Sensitivity</th>
<th>Did Not Mention Bias/Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Person</td>
<td>3 (3.4%)</td>
<td>85 (96.6%)</td>
</tr>
<tr>
<td>Online Asynchronous</td>
<td>19 (14.4%)</td>
<td>113 (85.6%)</td>
</tr>
<tr>
<td>Online Synchronous</td>
<td>5 (5.0%)</td>
<td>95 (95.0%)</td>
</tr>
</tbody>
</table>

*Note. \( \chi^2 = 10.472^*, df = 2. Numbers in parentheses indicate column percentages. \( ^*p = .005 \)

In a post-hoc analysis of the chi-square test results, shown in Table 24, the online asynchronous mode of training group was found to differ significantly at \( p = .006 \) from the other two training groups.
<table>
<thead>
<tr>
<th>Mode of Training</th>
<th>Adjusted Standardized Residual, Did Not Mention Bias</th>
<th>Adjusted Standardized Residual, Did Mention Bias</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Person</td>
<td>1.99</td>
<td>-1.99</td>
<td>.14</td>
</tr>
<tr>
<td>Online Asynchronous</td>
<td>-3.21</td>
<td>3.21</td>
<td>.01</td>
</tr>
<tr>
<td>Online Synchronous</td>
<td>1.49</td>
<td>-1.49</td>
<td>.33</td>
</tr>
</tbody>
</table>

Nineteen participants (14.4 percent) in the online asynchronous training had responses that addressed the theme of bias and sensitivity, as opposed to three participants in the in-person training (3.4 percent) and five participants (5 percent) in the online synchronous training. Examples of responses that were classified as indicative of the impact of participation on participants’ understanding of bias and sensitivity issues related to assessment practice include the following:

- “How to look at a test question and make it understandable and fair for all.”
- “How to write bias-free questions”
- “It will help me begin to write non-bias, non-sensitive, differing level of complexity test items in my own classroom.”
- “remove test bias and field test my questions”
- “The Sensitivity and Bias module really reminded me that I must constantly be aware of the material my students read and the activities i require them to do.”
- “I feel I can more effectively spot bias within my test questions.”
- “My assessments will not longer [sic.] be the same. I will take into account various factors such as bias and skill level.”
• “It's always interesting to review biases. As educators some times we do little things and are not aware those are biases.”
• “I now look at bias/sensitivity in a new light as it affects my test making skills.”

Theme 7: Insight into Measurement Issues of Reliability and Validity

The seventh theme found in participants’ responses was that of increased insight into the measurement issues of reliability and validity, relative to assessment practices. Ten participants (3.1 percent) mentioned this theme as one of the impacts of participation on their practices. There were no significant differences between the group responses as determined by a chi-square test of homogeneity, $\chi^2 (2, N = 320) = 4.91, p = .09$.

No participants in the in-person training had responses that addressed the theme of alignment of reliability and validity, as opposed to seven participants (5.3 percent) in the online asynchronous training and three participants (3 percent) in the online synchronous training. Examples of responses that were classified as indicative of the impact of participation on participants’ increased awareness of reliability and validity related to assessment practice include the following:

• “It will enhance the validity and reliability of my testing designs”
• “The workshop reinforces expectations for standards based lesson design and valid/reliable assessment design.”
• “Making sure all exams are valid and reliable”
• “Information provided was relevant in reevaluating classroom test preparation items by assuring that they are valid and reliable, relevant and rigorous.”
• “Testing is a high stakes concern for my students. I want to ensure the items are valid and reliable.”
• “Making sure I am watching for reliability and validity in all assessments.”
• “I've learned how to construct test questions and assessments that are valid and not biased.”

**Theme 8: Insight into Formatting and Mechanics of Assessments**

The eighth theme found in participants’ responses was that of increased insight into best practices of formatting and the mechanics of assessments. Twenty-eight participants (8.8 percent) mentioned this theme as one of the impacts of presentation on their practices. There were no significant differences between the group responses as determined by a chi-square test of homogeneity, $\chi^2(2, N = 320) = 0.62, p = .73$.

Six participants (6.8 percent) in the in-person training had responses that addressed the theme of assessment formatting and mechanics, as opposed to 13 participants (9.8 percent) in the online asynchronous training and nine participants (9 percent) in the online synchronous training.

Examples of responses that were classified as indicative of the impact of participation on participants’ increased insight into best practices in formatting and mechanics of assessments include the following:

• “One great idea that sticks in mind that I have not been dong [sic.] is to have the answers in abc order.”
• “Understanding the way the question style and format is taken into consideration”
• “I am more aware of cues that can lead student answers.”
• “More consistent item formatting.”
• “I think that it will help me to come up with a better format for my test questions.”
• “The information that was of greatest value to me was the formatting guidelines”
• “The rules regarding item writing will be useful in creating assessments for my students, department, and district.”

• “I enjoyed reading about how to format and formulate multiple choice answers.”

• “I will use the Item Style and Format Guide as my blueprint for writing items for my classroom assessment.”

• “The guidelines about ordering answers to math problems, as well as what are good and bad distractors will also enable me to make design better questions, and present material, as well as test taking strategies better.”

**Theme 9: Insight into the Development of Rubrics**

The ninth theme found in participants’ responses was that of increased insight into the development of rubrics for constructed response items or performance tasks. Twenty-three participants (7.2 percent) mentioned this theme as one of the impacts of participation on their practices.

There were significant differences between the group responses as determined by a chi-square test of homogeneity, $\chi^2(2, N = 320) = 23.07, p < .001$, as shown in Table 25.

<table>
<thead>
<tr>
<th>Mode of Training</th>
<th>Mentioned Rubrics</th>
<th>Did Not Mention Rubrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Person</td>
<td>16 (18.2%)</td>
<td>72 (81.8%)</td>
</tr>
<tr>
<td>Online Asynchronous</td>
<td>6 (4.5%)</td>
<td>126 (95.5%)</td>
</tr>
<tr>
<td>Online Synchronous</td>
<td>1 (1.0%)</td>
<td>99 (99.0%)</td>
</tr>
</tbody>
</table>

*Note. $\chi^2 = 23.07^*$, df = 2. Numbers in parentheses indicate column percentages. *$p < .001$
In a post-hoc analysis of the chi-square test results, shown in Table 26, significant differences were found in the proportion of participants who mentioned rubrics by the in-person mode of training group \((p < .001)\) and the online synchronous training group \((p = .02)\).

<table>
<thead>
<tr>
<th>Mode of Training</th>
<th>Adjusted Standardized Residual, Did Not Mention Rubrics</th>
<th>Adjusted Standardized Residual, Did Mention Rubrics</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Person</td>
<td>-4.69</td>
<td>4.69</td>
<td>.00</td>
</tr>
<tr>
<td>Online Asynchronous</td>
<td>1.53</td>
<td>-1.53</td>
<td>.31</td>
</tr>
<tr>
<td>Online Synchronous</td>
<td>2.89</td>
<td>-2.89</td>
<td>.02</td>
</tr>
</tbody>
</table>

Sixteen participants in the in-person training (18.2 percent) had responses that addressed the theme of the development of rubrics for constructed response items and performance tasks, as opposed to only six participants (4.5 percent) in the online asynchronous training and one participant (1 percent) in the online synchronous training. Examples of responses that were classified as indicative of the impact of participation on participants’ understanding of the creation of rubrics include the following:

- “The way that rubrics are leveled is new to me, and will help me write better rubrics/short answer questions in the future.”
- “How to write a rubric and extended questions”
- “How to write rubrics for extended response items”
- “I will have my students practice more using a rubric for the constructed response questions.”
- “How to more effectively create rubrics.”
“It helped me to see a little bit more about how to write test questions and to think deeper about my rubrics.”

“Good info on writing rubrics for extended response questions.’

“More project based activities and creating rubrics with the material.”

“The rubric will help make grading easier.”

**Research Question Five**

Research question five asked whether there was a relationship between teachers’ perceptions of the assessment literacy training and the mode of training that they received. All 456 respondents answered all questions. Table 27 shows the descriptive statistics for each item on the training evaluation survey, including the minimum and maximum response for each item. While the majority of items had a scale of 1 to 3, item numbers 3 and 5 had scales of 1 to 4 and 1 to 5, respectively. Higher values on the Likert scale indicated more positive responses, while lower values indicated more negative responses. The mean total response was computed, and was 29.48, with a standard deviation of 3.85. The minimum total response possible was 13, with a maximum possible response of 33.

As with research question 4, there were only three modes of training available for analysis: in-person, online synchronous, and online asynchronous. There were significant differences between the group responses as determined by one-way ANOVA ($F(2,453) = 49.70$, $p < .001$), as shown in Table 28.
Table 27

Descriptive Statistics for Evaluation Survey Results

<table>
<thead>
<tr>
<th>Question</th>
<th>Concept</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Organized</td>
<td>1</td>
<td>4</td>
<td>3.43</td>
<td>0.75</td>
</tr>
<tr>
<td>4</td>
<td>Objectives</td>
<td>1</td>
<td>3</td>
<td>2.81</td>
<td>0.42</td>
</tr>
<tr>
<td>5</td>
<td>Relevant Activities</td>
<td>1</td>
<td>5</td>
<td>4.65</td>
<td>0.70</td>
</tr>
<tr>
<td>6</td>
<td>Knowledgeable Instructor</td>
<td>1</td>
<td>3</td>
<td>2.79</td>
<td>0.45</td>
</tr>
<tr>
<td>7</td>
<td>Resources</td>
<td>1</td>
<td>3</td>
<td>2.77</td>
<td>0.45</td>
</tr>
<tr>
<td>9</td>
<td>Content Knowledge</td>
<td>1</td>
<td>3</td>
<td>2.59</td>
<td>0.63</td>
</tr>
<tr>
<td>10</td>
<td>Assessment Strategies</td>
<td>1</td>
<td>3</td>
<td>2.71</td>
<td>0.51</td>
</tr>
<tr>
<td>11</td>
<td>Using Data</td>
<td>1</td>
<td>3</td>
<td>2.48</td>
<td>0.71</td>
</tr>
<tr>
<td>12</td>
<td>Professional Growth</td>
<td>1</td>
<td>3</td>
<td>2.65</td>
<td>0.52</td>
</tr>
<tr>
<td>13</td>
<td>Preparation</td>
<td>1</td>
<td>3</td>
<td>2.60</td>
<td>0.54</td>
</tr>
<tr>
<td>N/A</td>
<td>Total Score</td>
<td>13</td>
<td>33</td>
<td>29.48</td>
<td>3.85</td>
</tr>
</tbody>
</table>

Table 28

ANOVA for Total Score on Evaluation Survey, by Mode of Training

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>Df</th>
<th>MS</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>1214.55</td>
<td>2</td>
<td>607.28</td>
<td>49.70</td>
<td>.00</td>
</tr>
<tr>
<td>Within</td>
<td>5535.27</td>
<td>453</td>
<td>12.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6749.82</td>
<td>455</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Given the significant ANOVA F test, a Tukey post hoc analysis was conducted. All three groups were found to be significantly different from each other at $p < .001$. The in-person trained group had the lowest total rating ($M = 26.92$, $SD = 4.75$), followed by the online asynchronous group ($M = 29.75$, $SD = 3.31$), while the online synchronous group had the highest overall rating of the training ($M = 31.17$, $SD = 2.39$).

An ANOVA was run for responses by item, to determine whether there were significant differences between the group responses. Table 29 contains the results of this ANOVA. All items showed significant differences in group responses ($p < .001$) with the exception of item number 13, which related to the extent to which participants felt prepared to begin writing or reviewing items after the training.
Table 29
ANOVA for Response to Each Item, by Mode of Training

<table>
<thead>
<tr>
<th>Item</th>
<th>Source</th>
<th>SS</th>
<th>Df</th>
<th>MS</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3: Organization</td>
<td>Between</td>
<td>57.86</td>
<td>2</td>
<td>28.93</td>
<td>66.89</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>195.90</td>
<td>453</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>253.75</td>
<td>455</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4: Objectives</td>
<td>Between</td>
<td>13.31</td>
<td>2</td>
<td>6.66</td>
<td>44.45</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>67.84</td>
<td>453</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>81.16</td>
<td>455</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5: Relevant Activities</td>
<td>Between</td>
<td>20.15</td>
<td>2</td>
<td>10.08</td>
<td>22.83</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>200.00</td>
<td>453</td>
<td>0.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>220.16</td>
<td>455</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6: Knowledgeable Instructor</td>
<td>Between</td>
<td>8.40</td>
<td>2</td>
<td>4.20</td>
<td>22.82</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>83.39</td>
<td>453</td>
<td>0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>91.79</td>
<td>455</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7: Resources</td>
<td>Between</td>
<td>5.79</td>
<td>2</td>
<td>2.90</td>
<td>15.53</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>84.49</td>
<td>453</td>
<td>0.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>90.28</td>
<td>455</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8: Content Knowledge</td>
<td>Between</td>
<td>15.33</td>
<td>2</td>
<td>7.66</td>
<td>21.04</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>164.99</td>
<td>453</td>
<td>0.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>180.31</td>
<td>455</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9: Assessment Strategies</td>
<td>Between</td>
<td>9.35</td>
<td>2</td>
<td>4.68</td>
<td>19.82</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>106.86</td>
<td>453</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>116.21</td>
<td>455</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10: Using Data</td>
<td>Between</td>
<td>19.27</td>
<td>2</td>
<td>9.63</td>
<td>20.72</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>210.59</td>
<td>453</td>
<td>0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>229.86</td>
<td>455</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11: Professional Growth</td>
<td>Between</td>
<td>7.36</td>
<td>2</td>
<td>3.68</td>
<td>14.28</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>116.79</td>
<td>453</td>
<td>0.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>124.16</td>
<td>455</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12: Preparation</td>
<td>Between</td>
<td>0.69</td>
<td>2</td>
<td>0.34</td>
<td>1.19</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td>Within</td>
<td>130.67</td>
<td>453</td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>131.36</td>
<td>455</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Because the ANOVA showed significant group differences in responses for all items except for item 13, post hoc Tukey tests were performed on all items except for item 13. The post hoc Tukey test for item 3, which assessed the organization of the training, showed that all three groups differed significantly from each other at \( p < .001 \). The in-person trained group had the lowest total rating (\( M = 2.89, SD = 0.90 \)), followed by the online asynchronous group (\( M = 3.46, SD = 0.65 \)), while the online synchronous group had the highest rating of the training’s organization (\( M = 3.83, SD = 0.38 \)).

For item 4, which evaluated the training objectives, the in-person group (\( M = 2.53, SD = 0.58 \)) had significant differences at \( p < .001 \) from the online asynchronous group (\( M = 2.90, SD = 0.30 \)) and from the online synchronous group (\( M = 2.93, SD = 0.28 \)), though the two online groups were not significantly different from each other in their responses to this item.

Item 5, which assessed the relevance of the training activities, had significant differences (\( p < .05 \)) between all three groups. The in-person group had the lowest ratings for this item (\( M = 4.32, SD = 0.92 \)), followed by the online asynchronous group (\( M = 4.67, SD = 0.63 \)), with the training rated highest in this area by the online synchronous group (\( M = 4.87, SD = 0.42 \)). In contrast, item 6, which measured the attendees’ perception of the instructors’ knowledge, had no significant differences between the in-person (\( M = 2.65, SD = 0.50 \)) and online asynchronous groups (\( M = 2.72, SD = 0.52 \)), though the differences between the online synchronous group (\( M = 2.98, SD = 0.14 \)) and the other two groups were significant at \( p < .001 \).

Item 7, which assessed the resources provided as part of the training, showed significant (\( p < .001 \)) differences between the in-person group (\( M = 2.59, SD = 0.54 \)) and both the online asynchronous (\( M = 2.79, SD = 0.42 \)) and the online synchronous (\( M = 2.89, SD = 0.34 \)) groups.
There were no significant differences between the two online groups in their responses to this item.

Responses to item 9, which assessed the impact of the training on participants’ content knowledge, showed significant ($p < .01$) differences among all three groups: in-person ($M = 2.33$, $SD = 0.75$), online asynchronous ($M = 2.58$, $SD = 0.64$), and online synchronous ($M = 2.81$, $SD = 0.39$).

In contrast, item 10, which evaluated the assessment strategies presented during the training, showed significant ($p < .001$) differences only between the in-person group ($M = 2.47$, $SD = 0.64$) and the two online groups. There were no significant differences between the online asynchronous group ($M = 2.79$, $SD = 0.42$) and the online synchronous group ($M = 2.80$, $SD = 0.42$). The same was true for item 11, which asked for participants’ perceptions of the impact of the training on their ability to use data. There were significant ($p < .001$) differences only between the in-person group ($M = 2.14$, $SD = 0.83$) and the two online groups. There were no significant differences between the responses of the online asynchronous group ($M = 2.59$, $SD = 0.64$) and those of the online synchronous group ($M = 2.62$, $SD = 0.61$).

Item 12, which asked for the perceived impact of the training on participants’ professional growth, showed significant ($p < .05$) differences among all three groups, with the in-person group rating this item the lowest ($M = 2.46$, $SD = 0.59$), followed by the online asynchronous group ($M = 2.65$, $SD = 0.51$) then by the online synchronous group ($M = 2.79$, $SD = 0.42$). Item 13, which measured participants’ perceptions of their level of preparedness following the training, was the only item to show no significant differences in the responses of participants based on mode of training.
Summary

Research question one asked whether there was any significant relationship between existing teacher-level factors such as years of experience, prior assessment development experience, type of degree, and National Board certification and the participating teachers’ ability to write high-quality assessment items. Two outcome variables were used as indicators of item quality, in order to answer this question: item discrimination and number of reviews per item as an indicator of the number of times each item was rejected. Using hierarchical linear modeling, between-writer variability in item discrimination values and in number of reviews by item was independently modeled as a function of (a) years of teaching experience, (b) prior assessment experience beyond the classroom level, (c) prior assessment experience beyond the school level, (d) graduate degree, and (e) National Board certification.

The ICC indicated that item writers accounted for 36.9 percent of the variance in item discrimination and 25.2 percent of the variance in the number of reviews by item. None of the pre-existing teacher-level variables had a significant impact on writers’ item discrimination, though prior assessment development experience beyond the classroom level was found to have a significant impact on the number of reviews per item. Having no pre-existing experience in developing assessment items beyond the classroom level was associated with an increase in the number of reviews for items developed by the writer (0.47), indicating that having limited prior experience in assessment development was associated with an increase in the number of times items were rejected by peer reviewers.

Research question two asked what impact, if any, the role selected by teachers (item writers vs. item writers/reviewers) had on teachers’ abilities to write high-quality assessment items. This question was answered in two ways: using hierarchical linear modeling, with the number of reviews per item used as an outcome variable, and using regression, with the
proportion of accepted items by each participant used as an outcome variable. Because prior assessment development experience beyond the classroom level was found to be significant in research question one, it was included as a variable in both models for research question two.

The impact of serving as both an item writer and reviewer, as opposed to an item writer only, on the number of reviews per item was statistically different from zero ($p < .01$). Serving as both a writer and a reviewer was associated with a 0.77 decrease in the number of reviews per item, indicating that items written by participants who served as both item writers and reviewers had a significantly lower number of reviews per item. These items were rejected less frequently than items written by participants who served as item writers only. After controlling for participant role, prior assessment development experience was no longer found to have a significant impact on the number of reviews per item.

When the proportion of accepted items by each participant was regressed against participant role, participant role accounted for 3.7 percent of the variance in proportion of accepted items. While participant role was a significant predictor of the proportion of accepted items, $F(1,125) = 5.79, p = .02$, it only accounted for a small percentage of the variance. This may be due to the fact that the proportion of accepted items was consistently high for all participants ($M = .92, SD = .19$). Because of this, number of reviews per item, which showed greater variability, may be a better indicator of differences among participants. Consistent with the results using number of reviews per item as the outcome variable, only role was found to be a significant predictor of proportion of accepted items ($p = .01$), while prior assessment development experience was not found to be significant ($p = .14$).

The third research question asked what impact, if any, the mode of training had on the quality of items written by participants. Two different outcome variables were used for this
Hierarchical linear modeling was used to answer this research question, since the items were nested within teachers. Mode of training was not found to have a significant impact on the number of reviews by item or on item discrimination. When proportion of accepted items by participants was regressed against mode of training, again, mode of training was not found to have a significant impact.

Research question four asked what teachers reported as the effects of participation in an assessment development project. The constant comparative method was used to analyze participants’ responses to open-ended items on the training evaluation surveys. The result of this analysis was the identification of nine themes: insight into standardized assessments and preparation for these assessments, improved knowledge of course standards or content, collaboration with other educators, cognitive complexity and rigor, alignment of assessment and instruction, insight into issues of bias and sensitivity as related to assessment, insight into measurement issues of reliability or validity as related to assessment, insight into formatting and mechanics of assessments, and insight into development of rubrics for constructed response items or performance tasks.

When responses were analyzed by mode of training, there were significant differences between participants for the following themes, based on mode of training: collaboration with other educators, cognitive complexity and rigor, alignment of assessment and instruction, bias and sensitivity, and insight into the development of rubrics.

Participants from the in-person mode of training were significantly more likely to provide responses that fell within the theme of collaboration with other educators than were participants from the two online modes of training. Participants from the online synchronous mode of training were significantly more likely to provide responses that referred to cognitive complexity
and rigor than were participants from the in-person or online asynchronous mode of training. For the theme of alignment of assessment and instruction, responses from participants from the in-person mode of training differed significantly from those in the online synchronous training group, with participants from the in-person training more likely to reference this theme in their responses, though there were no significant differences between the online asynchronous training group and the other two training groups. Regarding the theme of bias and sensitivity, responses from participants in the online asynchronous training group were significantly more likely to reference this theme than were responses from participants in the in-person training group, though there were no significant differences between the online synchronous group and the other two groups. Finally there were significant differences in the responses of participants from the in-person training group, which were more likely to reference the theme of development of rubrics, when compared to the two online training groups.

Research question five asked whether there was a relationship between teachers’ perceptions of the assessment literacy training and the mode of training that they received. Participants’ responses to the training evaluation survey were analyzed, with the outcome that the responses from participants in all three modes of training were significantly different from each other. Participants in the online synchronous group had the highest overall rating of the training \( (M = 31.17, \ SD = 2.39) \), followed by those in the online asynchronous group \( (M = 29.75, \ SD = 3.31) \). Participants in the in-person training group had the lowest overall rating \( (M = 26.92, \ SD = 4.8) \).

Of the items composing the evaluation questionnaire, the in-person training group’s responses were found to be significantly different from one or both of the online training groups’
responses in all items except for item 13, which measured participants’ level of preparedness following the training.

Thus, though there were no significant differences in the outcomes of the training, based on mode of training, there were significant differences in the participants’ perception of the training.
CHAPTER FIVE:
CONCLUSION AND DISCUSSION

This study sought to explore the impact of participation in an assessment development project on the quality of participating teachers’ assessment items as well as their own perceptions of the impact of participation. Specifically, the variables of mode of training (online synchronous, online asynchronous, in-person, or blended) and role (item writer or item writer and reviewer) were explored, as well as the pre-existing teacher-level factors of years of experience, type of degree, National Board certification, and prior assessment development experience.

Findings by Research Question

Pre-Existing Teacher Characteristics

First, the existing teacher-level factors of years of experience, graduate degree, National Board certification, and prior assessment development experience were examined in relation to two different outcome variables, used as indicators of item quality: item discrimination and number of reviews per item. None of these variables were found to have a significant impact on the item discrimination values, though prior assessment development experience beyond that of the classroom level was found to have a significant relationship with the number of reviews per item. These findings were consistent with that of previous researchers, who found no relationship between these variables and assessment literacy-related outcomes (Jarr, 2012; Kershaw, 1993; Mazzie, 2008; Williams & Rink, 2003). Although these earlier studies examined years of teaching experience, they did not specifically examine the variable of types of
assessments, development experience, so the finding of significance for this variable is a new finding of this study. Because the outcome variable for this research question was assessment item quality as determined by the number of times the item was rejected by peer reviewers, it is logical that those participants with previous assessment development experience would be able to apply these experiences to write items that were acceptable to their peers.

Though Sato, Chung, and Darling-Hammond (2008) did find a significant relationship between National Board certification and participants’ assessment practices, the sample size of this study was quite small (16 participants), and the outcome variable was classroom assessment practices, not the quality of test items developed by participants.

**Participant Role/Providing and Receiving Feedback**

After exploring the relationship between teachers’ pre-existing traits and the quality of the items they developed, variables over which the assessment development project had control were explored next, in research question two. Because peer feedback had been found by other studies to have a significant impact on the desired study outcomes (Cevik, 2015; Cho & Cho, 2011; Snowball & Mostert, 2013), the quality of participants’ items was examined in relationship to the role selected by the participant. Those who served as item writers only received feedback on their test items but did not provide feedback to others, while those who served as item writers and reviewers were in the position of both receiving feedback on their own items and providing feedback to others on their items. Based on the review of literature, it was hypothesized that the items of participants who served as both writers and reviewers would be of higher quality than those of participants who served as item writers only.

After controlling for prior assessment development experience, participant role was found to have a significant ($p < .01$) impact on the number of reviews per item. Items written by
participants who served as both item writers and reviewers had a significantly lower number of reviews per item, meaning these items were rejected less frequently than items written by participants who served as item writers only. Additionally, the variable of prior assessment development experience was no longer found to have a significant relationship with the number of reviews per item, once the variable of role was introduced.

Participant role was also found to be a significant predictor of the percentage of items that were accepted (those items that passed through all three levels of peer review and were given final approval at all levels). For both outcome variables that were used as indicators of item quality, participant role was found to have a significant impact, indicating that the ability to view others’ items, compare these items to a rubric, and provide feedback to other item writers had a significant impact on the quality of the items written by the peer reviewer. This finding was consistent with previous studies regarding the value of providing as well as receiving peer feedback on student academic learning (Cevik, 2015; Cho & Cho, 2011; Snowball & Mostert, 2013), which showed significant correlations between student learning outcomes and providing feedback to others. In this study, the teacher participants who served as both writers and reviewers were able to both receive feedback on their own items and provide feedback to others on their items. Serving in a role that allowed for the provision as well as the receipt of feedback had a significant impact on the quality of assessment items that they wrote. This could be explained by earlier studies (Nicol, Thomson & Breslin, 2014) that found that providing critical feedback meant engaging in evaluative judgments about others’ work, which translated to making the same evaluative judgements about one’s own work, and that being exposed to others’ work provided a broader perspective to participants (Snowball & Mostert, 2013).
Mode of Training Delivery and Quality of Assessment Writing

The third research question for this study examined the relationship between item quality and mode of training that the participant received. Four modes of training (online asynchronous, online synchronous, in-person, and blended) were first examined in relationship to the number of reviews by item, as an indicator of the number of times the item was rejected by reviewers, then in relationship to the item discrimination, for those items for which field test data was available, and finally in relationship to the overall percentage of accepted items by writer. The mode of training was not found to have a significant impact on any of these indicators of item quality. These findings are consistent with those of earlier studies discussed in the literature review, e.g. Fisher et al., 2010; Fishman et al., 2013; Russell et al., 2009; Wladis et al., 2014, who found no significant differences in outcome measures based on mode of training. Even earlier studies that did find significant differences based on mode of training had either mixed results, small effect sizes, or both, e.g. Bernard et al., 2004; Cavanaugh & Jacquemin, 2015; Lou et al., 2006; Powell et al., 2010; Smith, 2013; Young & Duncan, 2014).

Teacher Perception of Participation Effects and Delivery Mode

The fourth research question asked what teachers reported as the effects of participation in an assessment development project and whether these effects differed based on mode of training in which the teacher participated. The data source used for this portion of the study was participants’ open-ended responses on an evaluation survey that was administered after their completion of the training. Because participants were asked to complete the survey after their completion of either an online or an in-person training, and because the responses were anonymous, there was no way to determine which of the respondents participated in both online and in-person trainings. For this reason, there were only three modes of training examined in
relationship to survey results: online synchronous, online asynchronous, and in-person. Through the use of constant comparative analysis, nine themes were identified: insight into standardized assessments and preparation for these assessments, improved knowledge of course standards or content, collaboration with other educators, cognitive complexity and rigor, alignment of assessment and instruction, insight into issues of bias and sensitivity as related to assessment, insight into measurement issues of reliability or validity as related to assessment, insight into formatting and mechanics of assessments, and insight into development of rubrics for constructed response items or performance tasks.

When responses were analyzed by mode of training, there were significant differences between participants for the following themes, based on mode of training: collaboration with other educators, cognitive complexity and rigor, alignment of assessment and instruction, bias and sensitivity, and insight into the development of rubrics.

Participants from the in-person mode of training were significantly more likely to provide responses that fell within the theme of collaboration with other educators than were participants from the two online modes of training. This could be because of the ample opportunities for real-time, face-to-face collaboration that occurred with the in-person training.

Participants from the online synchronous mode of training were significantly more likely to provide responses that referred to cognitive complexity and rigor than were participants from the in-person or online asynchronous mode of training. This could possibly be because participants were given a homework assignment between the first and second training sessions, and this assignment asked participants to increase the cognitive complexity of an item on one of the teachers’ own assessments. Though this same activity was part of the online asynchronous and in-person trainings, there were other hands-on activities as well for these training modes.
Questions of cognitive complexity could have been more prevalent in the minds of participants in the online synchronous training since this may have stood out as the only hands-on activity they were asked to do.

For the theme of alignment of assessment and instruction, responses from participants from the in-person mode of training differed significantly from those in the online synchronous training group, with participants from the in-person training more likely to reference this theme in their responses, though there were no significant differences between the online asynchronous training group and the other two training groups. It is possible that these differences between modes of training could have occurred for this theme due to the varying levels of years of teaching experience by mode of training. Participants in the in-person mode of training were generally more experienced teachers with more prior assessment development experiences, and it is possible that this led to discussions regarding the alignment of assessment and instruction at the in-person trainings that did not occur with the online trainings.

Regarding the theme of bias and sensitivity, responses from participants in the online asynchronous training group were significantly more likely to reference this theme than were responses from participants in the in-person training group, though there were no significant differences between the online synchronous group and the other two groups. One possible cause of these differences is that the online asynchronous training used a wiki page, and participants were able to modify this document anonymously, as opposed to the in-person training, where the training on bias and sensitivity was done through a group discussion, which may have inhibited responses from some participants.

Finally there were significant differences in the responses of participants from the in-person training group, which were more likely to reference the theme of development of rubrics,
when compared to the two online training groups. This could have been because participants in the in-person training group were given the opportunity to practice developing and applying rubrics to prompts along with their peers, whereas there was no hands-on practice for rubric development in the online synchronous training and the practice in the online asynchronous training was done through a discussion board medium, and participants received inconsistent levels of feedback on their rubrics.

**Perceptions of Training by Mode**

These findings also held true for research question five, which asked whether there was a difference in participants’ perceptions of the assessment literacy training, based on mode of training. Similar to the findings above, though there were no significant differences in the empirical outcomes of item quality, there were significant differences based on participants’ perceptions of the trainings, as measured by their responses to the Likert scale items included on the training evaluation survey. All three groups were found to be significantly different from each other at $p < .001$. The in-person trained group had the lowest total rating, followed by the online asynchronous group, while the online synchronous group had the highest overall rating of the training. These findings support the results on research question four, that the perceptions of the trainings differed significantly based on mode of training, though there were no discernible significant differences in terms of the item quality, as measured by number of reviews by item and percentage of accepted items by participant. Part of the reason for these differences could be attributed to the fact that the in-person training was the longest, lasting four full days, while the highest-rated training mode, the online synchronous training, was also the shortest, lasting 4.5 to 5 hours in totality. Participants in the in-person training were given the most opportunities for hands-on, guided practice and for collaboration with their peers, but they may have felt the
length of the training to be excessive, particularly because this group had more prior assessment experience, based on the correlation matrix. Though none of the participants addressed this in their responses to the training evaluation survey, several commented to the trainers at the in-person training that they felt the training time was too long. The presence of a facilitator for the online synchronous training, who led the session and provided real-time responses to questions asked during the training, may also be one of the factors that contributed to the positive ratings of this mode of training. Caspi and Blau (2008) and Gunawardena and Zittle (1997), among others, found that aspects of social presence correlated positively with participants’ perceived learning; thus, theories of social presence could help provide an explanation for why participants rated the online synchronous training higher than the online asynchronous training. Another possibility is that the online synchronous training group, with limited prior assessment experiences, were more likely to receive answers to their questions as well as targeted and adaptive training than the in-person group or the asynchronous group.

These findings from research questions three and five were consistent with those of Smith (2013), who compared face-to-face and blended courses and found that students in the blended course reported higher perceptions of learning but showed no significant differences in their performance on assessments of learning. This is similar to the current study, where there are no significant differences in terms of the actual item quality, though there are differences in participants’ perceptions of the different modes of training.

Implications

Though there are several limitations that impact the generalizability of this study, such as sample size and selection, availability of item field test data, and the comparability of the training opportunities based on length of time, instructor, and activities, this study has several implications for others seeking to provide professional development to teachers on the
development of test items, particularly when taken in context as part of the larger body of research in this area.

The findings related to participant role have important implications for others attempting to train teachers to develop high-quality assessment items. Giving teachers the opportunity to provide feedback to others on their items, as well as to receive feedback on their own items, was found to have a beneficial impact on the quality of their own items. Because this finding is consistent with an earlier body of research on the effects of the provision as well as the reception of feedback, incorporating this opportunity into assessment development trainings may provide improved outcomes in terms of the quality of items developed by participants. Those seeking to train teachers to serve as test item writers may consider combining the roles of item writer and reviewer, to help improve the quality of developed items.

Additionally, there was some indication that providing assessment experiences to teachers beyond that of developing their own classroom assessments was beneficial to their ability to develop quality assessment items. Because participating in the development of assessments for a school, district, or state would typically involve additional training and collaboration with others, it appears to be beneficial to offer these types of opportunities to teachers.

Thirdly, although there were differences in terms of the time spent in training, the instructors of the training, and the actual training activities, there were no significant differences in terms of the item quality produced by participants from the different mode of training. There were differences, though, in terms of participants’ perceptions of the training based on mode, which may be explained by theories of social presence. Because the outcome variable was narrowly defined as test item quality, it is possible that there are other differences in outcomes
that may be beyond the scope of this study. Conversely, it is also possible that there were no differences in terms of the outcome of the training but that teachers perceived differences based on their enjoyment of the training experience. There is some precedent for self-reported outcomes differing from objectively measured outcomes; Smith (2013) in a comparison of face-to-face with blended courses, found that students in the blended class reported higher scores on a measure of their perceptions of learning, though there were no differences in the students’ performance on assessments of learning.

The absence of an impact of mode of training on project outcomes has implications for those seeking to provide training for teachers to serve as test item developers. As long as the quality of instruction is controlled, this finding leads to additional flexibility for those seeking to implement this type of training and for trainees in selecting a mode of training. Because funds for professional development are often limited, and since online trainings are generally associated with lower costs than in-person trainings, those desiring to train teachers in assessment development may choose to pursue online training as a viable replacement for traditional in-person professional development.

Finally, as assessment currently is and will likely continue to be an issue with political implications, the political climate as it pertains to assessment should be taken into consideration when conducting assessment-related training for teachers. Training goals need to include transparency regarding desired outcomes, particularly pertaining to the use of the assessment, including whether it will be summative or formative in nature, whether it will impact teachers’ evaluations, and what impact, if any, the assessment will have on students’ grades.

Future Research

This study focused narrowly on test item quality as the empirical outcome of participation in the assessment development project. Because participants reported broader
impacts of participation than just item quality, future research could study these additional impacts to determine whether mode of training and time spent in training have significant differences in terms of other assessment- or instructional-related outcomes. This research could include observation of classroom assessment and instructional practices, both before and after participation in the assessment training, to determine whether there are differences based on mode of training. Additionally, student learning outcomes could be measured to determine whether participation in an assessment development project does translate into improved student performance, as many participants reported that they believed would occur.

Future research could also focus on the relationship between prior assessment development experience and the mode of training selected. Because participants who reported having prior assessment development experience were significantly more likely to participate in either the in-person or the blended modes of training, this relationship may bear further examination.

Another intriguing finding was the differences in teachers’ perceptions of bias and sensitivity based on mode of training. Future research could examine whether participants are more comfortable expressing themselves on this potentially sensitive topic in an anonymous way than in participating in group discussions of this topic.

Other teacher-level variables that could be examined in the future include whether gender has a relationship on the quality of items produced and, since participation as a reviewer as well as a writer was shown to have a positive impact on the quality of the items written by these participants, the relationship between number of items reviewed and the quality of items written could also be explored.
REFERENCES


Florida Department of Education. (2013b, September 27). *Transition to next generation and computer-based tests in Florida.*


# APPENDIX A:
## ITEM WRITER AND REVIEWER CHECKLIST

<table>
<thead>
<tr>
<th>Item Writer's &amp; Reviewer's Checklist</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall Considerations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the item match the benchmark, and is it aligned to the benchmark in the platform? The content should not go beyond the content limits.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the cognitive complexity match the DOK indicated in the ITEM WRITER SIGN UP sheet?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is the reading level at or below the lowest grade level listed in the standard (i.e. 9-12 standards should be written at or below 9th grade reading level), with the exception of vocabulary included in the benchmark or item specification?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are clichés, textbook language, or jargon used?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are words of emphasis like least, most, greatest, main, opposite, best, etc. in <strong>boldface</strong>? Are restrictive words like Not and Except All emphasized? <strong>Boldface</strong> preferred.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are grammar, punctuation, and spelling correct?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are reference/resource materials included if necessary? E.g., formula sheets, calculator, etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If images or passages are included, is the source cited and/or is the copyright status given?</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Selected Response Items:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are cue words used in the question stem?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are there any grammatical hints in the questions to give it away?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are absolutes used? (i.e. all of the above, none of the above, etc.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is one benchmark measured primarily?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are terms repeated in the stem and responses?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are there only four answers? (only three for grades K-2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are the answer choices in the proper order in regards to the Multiple Choice Answer Choice Guidelines?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is the stem in the form of a question?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is there only one true and defensible answer?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Are rationales provided for each of the three wrong answer choices?  
Rationales should be worded in a way that explain why that answer choice is a plausible distractor.

Are misconceptions and/or logical misinterpretations used purposefully as distractors?

**Constructed Response Items: Short Answer Considerations**

- Does it ask the student to do or take two actions?
- Does the Stimulus provide a clear and concise action or task?
- Does the task appear as the final part of the stimulus?
- Is the type of answer expected clearly stated?
- Can the item be answered within 2 to 5 minutes?
- Can the question be answered with a few words or few sentences?
- Is there a three (3) point scoring guideline / rubric? (score points 0 - 1 - 2).
  
  In addition to a three point, rubric, short answer questions should have a SAMPLE FULL CREDIT RESPONSE. Does the item have this (in addition to) the rubric?

- Does the rubric include scoring points that are not explicit in the prompt?

**Constructed Response: Extended Response & Performance Task Considerations**

- Is the prompt clear, concise and focused?
- Does the task appear as the final part of the stimulus?
- Does the prompt tell the students the expectations for the format of their response?
  
  Is there a five (5) point scoring guideline / rubric? (score points 0-4). Keep in mind that with extended response and performance task, a sample full credit response is NOT required. In addition, English/Language Arts courses may opt to use a rubric with scoring range 1-6 instead of 0-4.

- Does the rubric include scoring points that are not explicit in the prompt? (I.e., use of correct conventions, spelling, etc.)

- Does the prompt encourage an extended response or performance? Explain, discuss, illustrate, etc.

- Do the prompt and the rubric match expectations? I.e., use of correct conventions.
APPENDIX B:  
DETAILED AGENDAS FOR THREE MODES OF TRAINING

Training Agenda for In-Person Training

Day 1:
A. Overview of assessment collaborative project
B. Introductions
C. Test of Franzpanics
D. Item Writing Rules – Overall Considerations (instructor-led presentation)
E. Item Writing Rules – Selected Response Specific (instructor-led presentation)
F. Poorly Constructed Multiple Choice Items (individual then group activity)
G. Reflection on learning (individual reflection)
H. Reviewing Webb’s Depth of Knowledge (instructor-led presentation)
I. Cognitive Complexity Levels and Associated Verbs (instructor-led presentation)
J. Depth of Knowledge Activity: Calibrating Levels of Complexity (individual then group activity)
K. Multiple Choice Item Writing – Moving Through Complexity (individual then group activity)
L. Copyright Concerns and Information (instructor-led presentation)
M. Lexile Reading Level Tools (resource)
N. Bias and Sensitivity (instructor-led presentation and group discussion)

Day 2: Selected Response Item Writing (item writing practice with peer feedback)

Day 3:
O. Good Item Writing Rules – Constructed Response Items (instructor-led presentation)
P. Short Response and Extended Response Rubric Template (instructor-led presentation)
Q. Writing Constructed Response Practice – Generic Scenario (individual then group activity)
R. Content-Based Scenarios – Writing CR Practice (individual then group activity)
S. Sample Constructed Response Items and Student Responses, Rubric, and Evaluation Form (individual then group activity)
T. Constructed Response Item Writing Practice (individual activity with peer feedback)
U. Constructed Response Sample Rubric (individual activity with peer feedback)

Day 4: Constructed Response Item Writing (item writing practice with peer feedback)
Training Agenda for Online Asynchronous Training

Module 1: Introduction, Background & Common Vocabulary
- Introduction to assessment collaborative project (video)
- Overview of Assessment Development: Test Design, Item Specifications, Item Writing, Item Review, Module Summary (presentation)
- Welcome Forum: Introduction & Post a Classroom Assessment (discussion board)
- Franzipanics (video & quiz)

Module 2: Item Process, Types & Format
- Item Specification Guide, Selected Response Items, Constructed Response Items, Item Style & Format (presentation)
- Interactive Practice: Item formatting (interactive presentation)
- Depth of Knowledge Video (video)
- Printable Resources: Depth of Knowledge, Item Style & Format Guidelines, Item Writer & Reviewer Checklist, Multiple Choice Answer Guidelines, Level 2 Item Reviewer Checklist, Good Response Writing Tips
- Rewriting Poorly Worded Questions (discussion board)
- Self-Assessment (quiz)

Module 3: Cognitive Complexity & Difficulty
- Cognitive Complexity & Difficulty (presentation)
- Printable Resources: Depth of Knowledge, Florida’s DOK Framework, Bloom’s Taxonomy & Webb’s Depth of Knowledge
- Increasing the Complexity (discussion board)
- Self-Assessment (quiz)

Module 4: Constructed Response Items
- Constructed Response Items (presentation)
- Printable Resources: Rules Regarding Constructed Response Items, Item Writer & Reviewer Checklist, Constructed Response Rubric, Test Item Specifications & Blueprints
- Constructed Response Item Practice (discussion board)

Module 5: Bias, Sensitivity & Copyright Issues
- Bias & Sensitivity, Copyright & Fair Use (presentation)
- Printable Resources: Bias & Sensitivity Guide, Copyright Rules, Copyright Free Resources, Release Forms
- Bias & Sensitivity Wiki Assignment (activity)
- Self-Assessment (quiz)

Module 6: Assessment Platform
- Forms
- Grant-funded Courses (link to platform training resources)
- Non-grant Funded Courses (link to platform training resources)
- Course Post-Test (quiz)

Module 7: Evaluation
- Item Writer/Reviewer Post-Test (quiz)
- End of Course Evaluation (survey)
Training Agenda for Online Synchronous Training

Session 1:
- Overview of assessment collaborative project and website
- Franzipanics (quiz then instructor review)
- Overview of assessment development process
- Steps for assessment development: Item specifications, test blueprint, item writing, item review
- Item types: Selected response vs. constructed response
- Selected response items: Formatting, examples
- Constructed response items: Gridded response, short answer, scoring rubric, example student responses, extended response, performance based, audio response, video response
- Cognitive complexity: Florida’s DOK framework, sample items

Homework: Increasing the cognitive complexity of items (activity)

Session 2:
- Cognitive complexity vs. item difficulty
- Bias & sensitivity
- Copyright issues
- Grant vs. non-grant policies & procedures
- IBTP vs. Eduphoria (assessment platform training)
- Item Writer/Reviewer Post-Test (quiz)
- End of Course Evaluation (survey)
APPENDIX C:
PARTICIPANT APPLICATION

Last Name *

First Name *

Please select your commitment level. *
○ I am ready to get to work right away
○ I need more information before I can commit.

If you have detailed questions that will help you decide if you want to participate, please ask them below:

What is the best phone number (or phone numbers) to contact you? *

What is the best e-mail address to use to contact you? *

What professional teaching certificates do you currently hold? Please list all that apply. *

Please list your current involvement level with *
○ I attended the June 17-20 workshop in
○ I attended the June 24-27 workshop in
○ I attended the July 15-18 workshop in
○ I completed the online item writer & reviewer training through Moodle.
○ I am a new participant, and have completed no trainings as of yet.
Which DISTRICT-FUNDED courses do you feel you have the expertise to contribute to the development of test items from existing item specifications? *
Please mark all that you would feel COMFORTABLE working on.
- Communications Technology
- Theatre, Cinema & Film Production
- M/J Guitar
- M/J Research
- M/J Speech & Debate
- Research (ELA)
- Television Production
- Chinese
- Aerospace Science 2-3 (ROTC)
- Leadership Education and Training 2-4 (ROTC)
- Automotive Maintenance and Light Repair
- Other District-Developed Assessments
- Other District-Developed Assessments: Access/ESE Courses

We have opportunities for involvement in assessment development in two different roles: 1) Item Writers and 2) Item Reviewers/Editors. Which role(s) would you be interested in? *
- Item writing (development of original items, may include researching and using public domain passages, images, or graphs, or development of new charts, graphs, etc.)
- Item review and editing (editing/revision of work developed by item writers - requires close attention to detail and intimate familiarity with item specifications)

What school district are you currently employed in? *

How many years have you been in education? *
- 0-3 years
- 4-6 years
- 7-10 years
- 11-15 years
- 16-20 years
- 20+ years

If you have any of the following additional qualifications, please indicate that here (check all that apply):
- Advanced degree (Masters, Specialist, Doctorate)
- National Board Certification
- ESOL Endorsement or Certification
- Reading Endorsement or Certification
Please list any professional associations related to your subject that you belong to. Indicate "None" if you are not a member of any professional associations.

Please describe your experience with writing or designing curriculum in your selected area at the school, district, and/or state level.

Please describe your experience in assessment development in your subject area at the school, district, and/or state level.

Please describe any specific strength which you possess that you feel would be advantageous to the writing or reviewing teams (i.e. leadership, technology, proofreading, etc.).

Prior to participating in this project, what, if any, formal training have you had in assessment? Please include any relevant university experience, professional development workshops, etc.

Responses to this question will NOT impact your ability to participate in the project; we are just collecting information about what types of assessment learning opportunities are provided to teachers in general.
If you have any other information about yourself that you would like us to be aware of, please indicate that here. Please include any past experiences with assessment development at the classroom, school, district, or state level that you feel would be relevant.

Optional

Submit
APPENDIX D:
PARTICIPANT TRAINING EVALUATION SURVEY

* 1. Which workshop did you attend?
   - Online Self-Paced Training (through CourseSites)
   - Online Synchronous Training (Webinar with [redacted])
   - Online Synchronous Training (Webinar with [redacted])
   - [Redacted]
   - [Redacted]
   - [Redacted]
   - [Redacted]
   - [Redacted]
   - [Redacted]
   - [Redacted]

* 2. Which statement best describes you?
   - I am a teacher, currently employed by a public school district.
   - I am a teacher, currently employed by a charter school district.
   - I am a teacher, currently employed by a virtual school.
   - I am a district-level administrator or specialist.
   - I am a retired teacher.
   - I am a college-level instructor.
   - I am a graduate student.
   - Other (please specify)

* 3. Was the training well organized?
   - It was very well organized.
* 4. Were the training objectives clearly stated?
   - They were clearly stated.
   - They were stated, but were somewhat unclear.
   - They were not stated.

* 5. Were the training activities relevant to the training objectives?
   - All of the training activities were relevant to the training objectives.
   - More than half of the training activities were relevant to the training objectives.
   - About half of the training activities were relevant to the training objectives.
   - A few of the training activities were relevant to the training objectives.
   - None of the training activities were relevant to the training objectives.

* 6. Were the instructors knowledgeable and engaging?
   - All of the instructors were knowledgeable and engaging.
   - Some of the instructors were knowledgeable and engaging.
   - None of the instructors were knowledgeable and engaging.
* 7. Were you provided with all of the necessary resources that you will need to continue item writing or review on your own?
   - I have all of the resources that I need to continue writing or reviewing items on my own.
   - I have some of the resources that I need to continue writing or reviewing items on my own.
   - I don’t have any of the resources that I need to continue writing or reviewing items on my own.
   Other (please specify)

8. If you answered that you have some or none of the resources you will need to continue writing or reviewing items on your own, what additional resources could we provide you?

9. Did you feel that this workshop enhanced your content knowledge?
   - This workshop enhanced my content knowledge a great deal.
   - This workshop enhanced my content knowledge somewhat.
   - This workshop did not enhance my content knowledge.
   Other (please specify)

10. Did this activity provide information on a variety of assessment strategies?
    - It provided information on many different assessment strategies.
    - It provided information on a few different assessment strategies.
    - It did not provide information on any assessment strategies.
    Other (please specify)
11. Did this workshop provide you with additional skills for using data in instructional decision making?
   - This workshop provided me with many skills for using data in instructional decision making.
   - This workshop provided me with a few skills for using data in instructional decision making.
   - This workshop did not provide me with any skills for using data in instructional decision making.
   Other (please specify)

12. Did this workshop enhance your professional growth?
   - This workshop greatly enhanced my professional growth.
   - This workshop somewhat enhanced my professional growth.
   - This workshop did not enhance my professional growth.
   Other (please specify)

13. Did this workshop effectively prepare you to serve as an item writer or reviewer?
   - I feel prepared to write high quality test items, or to review items and provide high quality feedback to others.
   - I feel somewhat prepared to write high quality test items, or to review items and provide high quality feedback to others, but I feel that I will still need additional support from the instructors.
   - I feel uncertain about my ability to write high quality test items, or to review items and provide high quality feedback to others.
   Other (please specify)

14. What additional support could we provide you to help you feel more confident in your abilities as an item writer or reviewer?
15. How was this workshop relevant to your job?

16. In what ways do you believe this workshop will affect your current instructional or assessment practices?

17. What new ideas have you gained and how do you plan to implement these new ideas in your job or training capacity?

18. What information or activity was of greatest value to you?

19. What specific suggestions do you have to improve this workshop?

20. Did you feel that you gained additional knowledge from participating in the breakout sessions (beyond what you learned in the whole group session)? If so, what do you feel was the benefit of these sessions?
21. If you have participated in the project previously, through the online/virtual collaboration option, which mode of participation do you prefer (online or face to face)? Why? (Please mark N/A if this does not apply to you)

22. Do you have any other comments or suggestions for the workshop organizers?
## APPENDIX E:

**RESULTS OF 2PL IRT ANALYSIS FOR ITEMS**

Table 30

*Results of 2PL IRT Analysis for Items with Field Test Data*

<table>
<thead>
<tr>
<th>UIN</th>
<th>Number of Reviews for Item</th>
<th>Difficulty Estimate</th>
<th>Difficulty Standard Error</th>
<th>PrGTt</th>
<th>Slope Estimate</th>
<th>Slope Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>02-2002500-00209</td>
<td>6</td>
<td>-0.83</td>
<td>0.21</td>
<td>&lt;.001</td>
<td>-1.11</td>
<td>0.24</td>
</tr>
<tr>
<td>02-2002510-00135</td>
<td>5</td>
<td>-1.02</td>
<td>0.30</td>
<td>0.00</td>
<td>-0.81</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2002510-00162</td>
<td>4</td>
<td>-1.18</td>
<td>0.50</td>
<td>0.01</td>
<td>-0.51</td>
<td>0.19</td>
</tr>
<tr>
<td>02-1201310-00172</td>
<td>4</td>
<td>4.05</td>
<td>0.00</td>
<td>&lt;.001</td>
<td>0.32</td>
<td>0.00</td>
</tr>
<tr>
<td>02-2003310-00175</td>
<td>3</td>
<td>-8.73</td>
<td>2.91</td>
<td>0.00</td>
<td>-0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>02-1202340-00193</td>
<td>4</td>
<td>0.05</td>
<td>0.10</td>
<td>0.30</td>
<td>0.63</td>
<td>0.09</td>
</tr>
<tr>
<td>02-2002520-00231</td>
<td>6</td>
<td>13.03</td>
<td>25.32</td>
<td>0.30</td>
<td>0.11</td>
<td>0.22</td>
</tr>
<tr>
<td>02-2106350-00098</td>
<td>12</td>
<td>3.58</td>
<td>2.30</td>
<td>0.06</td>
<td>0.29</td>
<td>0.19</td>
</tr>
<tr>
<td>02-2002500-00213</td>
<td>5</td>
<td>4.77</td>
<td>3.75</td>
<td>0.10</td>
<td>0.27</td>
<td>0.22</td>
</tr>
<tr>
<td>02-2002510-00166</td>
<td>4</td>
<td>7.20</td>
<td>7.08</td>
<td>0.15</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>02-2002500-00241</td>
<td>5</td>
<td>4.48</td>
<td>3.11</td>
<td>0.08</td>
<td>0.33</td>
<td>0.23</td>
</tr>
<tr>
<td>02-1006300-00101</td>
<td>3</td>
<td>0.13</td>
<td>0.24</td>
<td>0.29</td>
<td>0.58</td>
<td>0.15</td>
</tr>
<tr>
<td>02-2003320-00162</td>
<td>3</td>
<td>2.50</td>
<td>0.44</td>
<td>&lt;.001</td>
<td>0.40</td>
<td>0.07</td>
</tr>
<tr>
<td>02-2003310-00175</td>
<td>3</td>
<td>0.62</td>
<td>0.13</td>
<td>&lt;.001</td>
<td>0.54</td>
<td>0.07</td>
</tr>
<tr>
<td>02-1006300-00197</td>
<td>14</td>
<td>-1.94</td>
<td>0.35</td>
<td>&lt;.001</td>
<td>1.04</td>
<td>0.23</td>
</tr>
<tr>
<td>02-1202340-00182</td>
<td>3</td>
<td>3.58</td>
<td>0.53</td>
<td>&lt;.001</td>
<td>1.23</td>
<td>0.25</td>
</tr>
<tr>
<td>02-1211300-00218</td>
<td>3</td>
<td>1.01</td>
<td>0.25</td>
<td>&lt;.001</td>
<td>1.30</td>
<td>0.38</td>
</tr>
<tr>
<td>02-2003310-00180</td>
<td>4</td>
<td>-0.20</td>
<td>0.06</td>
<td>0.00</td>
<td>1.21</td>
<td>0.10</td>
</tr>
<tr>
<td>02-2003310-00187</td>
<td>3</td>
<td>-1.53</td>
<td>0.23</td>
<td>&lt;.001</td>
<td>1.35</td>
<td>0.25</td>
</tr>
<tr>
<td>02-2003320-00107</td>
<td>4</td>
<td>0.26</td>
<td>0.03</td>
<td>&lt;.001</td>
<td>-1.63</td>
<td>0.07</td>
</tr>
<tr>
<td>02-2002510-00059</td>
<td>4</td>
<td>-0.18</td>
<td>0.16</td>
<td>0.13</td>
<td>-1.06</td>
<td>0.23</td>
</tr>
<tr>
<td>02-2002500-00091</td>
<td>3</td>
<td>-0.63</td>
<td>0.16</td>
<td>&lt;.001</td>
<td>-1.47</td>
<td>0.28</td>
</tr>
<tr>
<td>02-2002500-00075</td>
<td>3</td>
<td>-1.38</td>
<td>0.34</td>
<td>&lt;.001</td>
<td>-0.90</td>
<td>0.22</td>
</tr>
<tr>
<td>02-2002510-00107</td>
<td>4</td>
<td>-0.76</td>
<td>0.17</td>
<td>&lt;.001</td>
<td>-1.33</td>
<td>0.26</td>
</tr>
<tr>
<td>02-2002510-00128</td>
<td>4</td>
<td>-0.74</td>
<td>0.06</td>
<td>&lt;.001</td>
<td>-0.80</td>
<td>0.05</td>
</tr>
<tr>
<td>02-1006300-00164</td>
<td>3</td>
<td>14.85</td>
<td>55.69</td>
<td>0.39</td>
<td>0.04</td>
<td>0.14</td>
</tr>
<tr>
<td>02-2002520-00033</td>
<td>3</td>
<td>-0.36</td>
<td>0.24</td>
<td>0.07</td>
<td>-0.66</td>
<td>0.20</td>
</tr>
<tr>
<td>02-1006000-00208</td>
<td>3</td>
<td>1.15</td>
<td>0.71</td>
<td>0.05</td>
<td>0.33</td>
<td>0.17</td>
</tr>
<tr>
<td>02-2002500-00233</td>
<td>5</td>
<td>-0.21</td>
<td>0.15</td>
<td>0.08</td>
<td>-1.10</td>
<td>0.24</td>
</tr>
<tr>
<td>02-2002520-00235</td>
<td>3</td>
<td>-0.26</td>
<td>0.16</td>
<td>0.05</td>
<td>-1.08</td>
<td>0.23</td>
</tr>
<tr>
<td>UIN</td>
<td>Number of Reviews for Item</td>
<td>Difficulty Estimate</td>
<td>Difficulty Standard Error</td>
<td>Slope Estimate</td>
<td>Slope Standard Error</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------</td>
<td>---------------------</td>
<td>---------------------------</td>
<td>----------------</td>
<td>---------------------</td>
<td></td>
</tr>
<tr>
<td>02-2000360-00024</td>
<td>2</td>
<td>-3.39</td>
<td>0.91</td>
<td>-0.43</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>02-2002510-00110</td>
<td>4</td>
<td>-0.96</td>
<td>0.35</td>
<td>0.00</td>
<td>-0.65</td>
<td></td>
</tr>
<tr>
<td>02-2002510-00081</td>
<td>6</td>
<td>-11.31</td>
<td>19.64</td>
<td>0.28</td>
<td>-0.12</td>
<td></td>
</tr>
<tr>
<td>02-2002500-00268</td>
<td>4</td>
<td>-1.17</td>
<td>0.50</td>
<td>0.01</td>
<td>-0.51</td>
<td></td>
</tr>
<tr>
<td>02-2002500-00220</td>
<td>4</td>
<td>-0.48</td>
<td>0.30</td>
<td>0.05</td>
<td>-0.57</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00156</td>
<td>3</td>
<td>0.17</td>
<td>0.21</td>
<td>0.21</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>02-2002510-00114</td>
<td>3</td>
<td>-1.88</td>
<td>0.67</td>
<td>0.00</td>
<td>-0.56</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00243</td>
<td>6</td>
<td>784.28</td>
<td>10409</td>
<td>0.50</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>02-2000360-00115</td>
<td>4</td>
<td>-1.84</td>
<td>0.95</td>
<td>0.03</td>
<td>-0.38</td>
<td></td>
</tr>
<tr>
<td>02-2002500-00253</td>
<td>6</td>
<td>14.36</td>
<td>36.69</td>
<td>0.35</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>02-2000440-00067</td>
<td>3</td>
<td>0.13</td>
<td>0.19</td>
<td>0.25</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00160</td>
<td>5</td>
<td>19.76</td>
<td>28.00</td>
<td>0.24</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>02-2003320-00091</td>
<td>4</td>
<td>13.42</td>
<td>11.68</td>
<td>0.13</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>02-2000350-00158</td>
<td>3</td>
<td>10.94</td>
<td>9.42</td>
<td>0.12</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>02-2000350-00062</td>
<td>3</td>
<td>5.88</td>
<td>3.57</td>
<td>0.05</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>02-2000350-00081</td>
<td>4</td>
<td>5.62</td>
<td>3.18</td>
<td>0.04</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>02-2000350-00025</td>
<td>2</td>
<td>0.35</td>
<td>0.42</td>
<td>0.20</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>02-2002520-00149</td>
<td>4</td>
<td>-2.39</td>
<td>0.33</td>
<td>&lt;.001</td>
<td>-0.29</td>
<td></td>
</tr>
<tr>
<td>02-2003320-00100</td>
<td>3</td>
<td>5.64</td>
<td>2.05</td>
<td>0.00</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00241</td>
<td>4</td>
<td>1.44</td>
<td>0.86</td>
<td>0.05</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>02-2000350-00074</td>
<td>5</td>
<td>2.95</td>
<td>0.97</td>
<td>0.00</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>02-2000350-00026</td>
<td>2</td>
<td>-0.36</td>
<td>0.28</td>
<td>0.10</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>02-2000350-00099</td>
<td>3</td>
<td>3.09</td>
<td>1.00</td>
<td>0.00</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00229</td>
<td>4</td>
<td>-1.53</td>
<td>0.75</td>
<td>0.02</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00196</td>
<td>4</td>
<td>-0.40</td>
<td>0.16</td>
<td>0.01</td>
<td>1.05</td>
<td></td>
</tr>
<tr>
<td>02-2000350-00085</td>
<td>4</td>
<td>0.21</td>
<td>0.24</td>
<td>0.19</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>02-1210300-00150</td>
<td>3</td>
<td>-0.10</td>
<td>0.33</td>
<td>0.38</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>02-2002500-00222</td>
<td>3</td>
<td>-12.03</td>
<td>21.83</td>
<td>0.29</td>
<td>-0.11</td>
<td></td>
</tr>
<tr>
<td>02-2002500-00281</td>
<td>3</td>
<td>-10.07</td>
<td>18.30</td>
<td>0.29</td>
<td>-0.11</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00155</td>
<td>3</td>
<td>-1.19</td>
<td>0.20</td>
<td>&lt;.001</td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td>02-2000350-00120</td>
<td>3</td>
<td>0.49</td>
<td>0.18</td>
<td>0.00</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>02-2000360-00113</td>
<td>3</td>
<td>0.96</td>
<td>0.26</td>
<td>0.00</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>02-2003310-00172</td>
<td>3</td>
<td>0.41</td>
<td>0.14</td>
<td>0.00</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>02-2003310-00111</td>
<td>4</td>
<td>0.66</td>
<td>0.16</td>
<td>&lt;.001</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>02-2002500-00300</td>
<td>4</td>
<td>-49.96</td>
<td>94.84</td>
<td>0.30</td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td>02-2002520-00126</td>
<td>4</td>
<td>423.56</td>
<td>5553.00</td>
<td>0.47</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>02-2000350-00157</td>
<td>3</td>
<td>-0.23</td>
<td>0.16</td>
<td>0.08</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>02-2003310-00092</td>
<td>4</td>
<td>-0.23</td>
<td>0.11</td>
<td>0.02</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>UIN</td>
<td>Number of Reviews for Item</td>
<td>Difficulty Estimate</td>
<td>Difficulty Standard Error</td>
<td>PrGTt Slope Estimate</td>
<td>Slope Standard Error</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------</td>
<td>---------------------</td>
<td>---------------------------</td>
<td>----------------------</td>
<td>----------------------</td>
<td></td>
</tr>
<tr>
<td>02-2003320-00121</td>
<td>3</td>
<td>2.07</td>
<td>0.28</td>
<td>&lt;.001</td>
<td>0.54</td>
<td>0.07</td>
</tr>
<tr>
<td>02-1006300-00152</td>
<td>3</td>
<td>-0.37</td>
<td>0.13</td>
<td>0.00</td>
<td>1.34</td>
<td>0.21</td>
</tr>
<tr>
<td>02-1007300-00096</td>
<td>5</td>
<td>0.26</td>
<td>0.22</td>
<td>0.12</td>
<td>0.64</td>
<td>0.16</td>
</tr>
<tr>
<td>02-2003310-00110</td>
<td>3</td>
<td>-0.43</td>
<td>0.22</td>
<td>0.02</td>
<td>0.61</td>
<td>0.14</td>
</tr>
<tr>
<td>02-1006000-00205</td>
<td>3</td>
<td>-0.10</td>
<td>0.13</td>
<td>0.22</td>
<td>1.39</td>
<td>0.25</td>
</tr>
<tr>
<td>02-2000350-00011</td>
<td>3</td>
<td>-0.57</td>
<td>0.15</td>
<td>&lt;.001</td>
<td>0.67</td>
<td>0.11</td>
</tr>
<tr>
<td>02-2000360-00093</td>
<td>4</td>
<td>-1.42</td>
<td>0.25</td>
<td>&lt;.001</td>
<td>0.68</td>
<td>0.12</td>
</tr>
<tr>
<td>02-2002510-00093</td>
<td>3</td>
<td>-0.66</td>
<td>0.22</td>
<td>0.00</td>
<td>0.66</td>
<td>0.15</td>
</tr>
<tr>
<td>02-2002510-00094</td>
<td>3</td>
<td>1.25</td>
<td>0.14</td>
<td>&lt;.001</td>
<td>0.74</td>
<td>0.08</td>
</tr>
<tr>
<td>02-2000350-00108</td>
<td>3</td>
<td>-0.66</td>
<td>0.15</td>
<td>&lt;.001</td>
<td>0.77</td>
<td>0.12</td>
</tr>
<tr>
<td>02-2000350-00109</td>
<td>3</td>
<td>0.75</td>
<td>0.14</td>
<td>&lt;.001</td>
<td>0.80</td>
<td>0.11</td>
</tr>
<tr>
<td>02-2000350-00024</td>
<td>3</td>
<td>-0.29</td>
<td>0.10</td>
<td>0.00</td>
<td>0.90</td>
<td>0.12</td>
</tr>
<tr>
<td>02-2000350-00072</td>
<td>5</td>
<td>-0.86</td>
<td>0.13</td>
<td>&lt;.001</td>
<td>0.92</td>
<td>0.13</td>
</tr>
<tr>
<td>02-1006300-00093</td>
<td>4</td>
<td>-0.14</td>
<td>0.11</td>
<td>0.11</td>
<td>1.67</td>
<td>0.25</td>
</tr>
<tr>
<td>02-2000440-00108</td>
<td>4</td>
<td>-0.63</td>
<td>0.16</td>
<td>&lt;.001</td>
<td>0.96</td>
<td>0.18</td>
</tr>
<tr>
<td>02-1006300-00099</td>
<td>6</td>
<td>-1.10</td>
<td>0.15</td>
<td>&lt;.001</td>
<td>1.74</td>
<td>0.29</td>
</tr>
<tr>
<td>02-2003320-00092</td>
<td>3</td>
<td>-0.59</td>
<td>0.07</td>
<td>&lt;.001</td>
<td>1.02</td>
<td>0.09</td>
</tr>
<tr>
<td>02-2002500-00083</td>
<td>5</td>
<td>2.71</td>
<td>1.13</td>
<td>0.01</td>
<td>0.58</td>
<td>0.26</td>
</tr>
<tr>
<td>02-2000350-00080</td>
<td>5</td>
<td>-1.35</td>
<td>0.16</td>
<td>&lt;.001</td>
<td>1.15</td>
<td>0.16</td>
</tr>
<tr>
<td>02-2002510-00124</td>
<td>5</td>
<td>-0.36</td>
<td>0.17</td>
<td>0.01</td>
<td>0.82</td>
<td>0.16</td>
</tr>
<tr>
<td>02-2000350-00164</td>
<td>3</td>
<td>-1.56</td>
<td>0.17</td>
<td>&lt;.001</td>
<td>1.35</td>
<td>0.19</td>
</tr>
<tr>
<td>02-1007300-00104</td>
<td>3</td>
<td>-0.98</td>
<td>0.12</td>
<td>&lt;.001</td>
<td>2.40</td>
<td>0.39</td>
</tr>
<tr>
<td>02-2003310 Physical Science-00003</td>
<td>3</td>
<td>1.49</td>
<td>0.06</td>
<td>&lt;.001</td>
<td>-1.65</td>
<td>0.09</td>
</tr>
<tr>
<td>02-Sociology-00017</td>
<td>6</td>
<td>2.51</td>
<td>1.27</td>
<td>0.02</td>
<td>-0.35</td>
<td>0.18</td>
</tr>
<tr>
<td>02-2002510-00185</td>
<td>3</td>
<td>-1.52</td>
<td>0.48</td>
<td>0.00</td>
<td>-0.65</td>
<td>0.20</td>
</tr>
<tr>
<td>02-2000350-00185</td>
<td>3</td>
<td>-0.52</td>
<td>0.21</td>
<td>0.01</td>
<td>-0.90</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2000440-00187</td>
<td>4</td>
<td>-0.36</td>
<td>0.05</td>
<td>&lt;.001</td>
<td>-0.73</td>
<td>0.05</td>
</tr>
<tr>
<td>02-2002510-00134</td>
<td>4</td>
<td>-2.03</td>
<td>0.83</td>
<td>0.01</td>
<td>-0.47</td>
<td>0.19</td>
</tr>
<tr>
<td>02-2002500-00138</td>
<td>4</td>
<td>-3.47</td>
<td>1.64</td>
<td>0.02</td>
<td>-0.44</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2106350-00264</td>
<td>4</td>
<td>13.66</td>
<td>48.87</td>
<td>0.39</td>
<td>0.05</td>
<td>0.17</td>
</tr>
<tr>
<td>02-2002510-00133</td>
<td>4</td>
<td>-4.81</td>
<td>3.63</td>
<td>0.09</td>
<td>-0.26</td>
<td>0.20</td>
</tr>
<tr>
<td>02-2108300-00217</td>
<td>4</td>
<td>-0.02</td>
<td>0.31</td>
<td>0.48</td>
<td>0.44</td>
<td>0.16</td>
</tr>
<tr>
<td>02-2107310-00176</td>
<td>3</td>
<td>1.50</td>
<td>1.22</td>
<td>0.11</td>
<td>0.25</td>
<td>0.18</td>
</tr>
<tr>
<td>02-2108300-00089</td>
<td>6</td>
<td>1.15</td>
<td>0.45</td>
<td>0.01</td>
<td>0.52</td>
<td>0.17</td>
</tr>
<tr>
<td>02-2003310-00045</td>
<td>4</td>
<td>-10.33</td>
<td>8.33</td>
<td>0.11</td>
<td>-0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>02-Sociology-00009</td>
<td>4</td>
<td>-1.03</td>
<td>0.33</td>
<td>0.00</td>
<td>0.64</td>
<td>0.18</td>
</tr>
<tr>
<td>02-2000350-00181</td>
<td>3</td>
<td>4.40</td>
<td>1.63</td>
<td>0.00</td>
<td>0.28</td>
<td>0.10</td>
</tr>
<tr>
<td>02-1211300-00202</td>
<td>3</td>
<td>2.16</td>
<td>0.51</td>
<td>&lt;.001</td>
<td>1.35</td>
<td>0.46</td>
</tr>
<tr>
<td>UIN</td>
<td>Number of Reviews for Item</td>
<td>Difficulty Estimate</td>
<td>Difficulty Standard Error</td>
<td>Slope Estimate</td>
<td>Slope Standard Error</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>---------------------------</td>
<td>---------------------</td>
<td>--------------------------</td>
<td>----------------</td>
<td>---------------------</td>
<td></td>
</tr>
<tr>
<td>02-1202340-00163</td>
<td>3</td>
<td>0.21</td>
<td>0.15</td>
<td>0.08</td>
<td>1.39</td>
<td>0.40</td>
</tr>
<tr>
<td>02-2000350-00166</td>
<td>3</td>
<td>0.08</td>
<td>0.12</td>
<td>0.24</td>
<td>0.72</td>
<td>0.11</td>
</tr>
<tr>
<td>02-2000410-00101</td>
<td>3</td>
<td>-1.24</td>
<td>0.42</td>
<td>0.00</td>
<td>0.47</td>
<td>0.14</td>
</tr>
<tr>
<td>02-2108300-00219</td>
<td>7</td>
<td>-2.01</td>
<td>0.42</td>
<td>&lt;.001</td>
<td>1.24</td>
<td>0.35</td>
</tr>
<tr>
<td>02-Sociology-00011</td>
<td>3</td>
<td>-0.16</td>
<td>0.14</td>
<td>0.12</td>
<td>1.25</td>
<td>0.24</td>
</tr>
<tr>
<td>02-1202340-00156</td>
<td>3</td>
<td>0.20</td>
<td>0.12</td>
<td>0.05</td>
<td>2.21</td>
<td>0.82</td>
</tr>
<tr>
<td>02-1006300-00073</td>
<td>3</td>
<td>-3.09</td>
<td>0.84</td>
<td>0.00</td>
<td>-0.80</td>
<td>0.25</td>
</tr>
<tr>
<td>02-2002500-00198</td>
<td>4</td>
<td>-0.09</td>
<td>0.11</td>
<td>0.21</td>
<td>-1.85</td>
<td>0.35</td>
</tr>
<tr>
<td>02-1006000-00234</td>
<td>3</td>
<td>-3.48</td>
<td>1.82</td>
<td>0.03</td>
<td>-0.36</td>
<td>0.20</td>
</tr>
<tr>
<td>02-2106350-00156</td>
<td>4</td>
<td>-3.50</td>
<td>1.56</td>
<td>0.01</td>
<td>-0.49</td>
<td>0.23</td>
</tr>
<tr>
<td>02-2000410-00012</td>
<td>4</td>
<td>-1.70</td>
<td>0.50</td>
<td>0.00</td>
<td>-0.71</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2108300-00145</td>
<td>3</td>
<td>-3.75</td>
<td>2.81</td>
<td>0.09</td>
<td>-0.22</td>
<td>0.17</td>
</tr>
<tr>
<td>02-2002520-00129</td>
<td>4</td>
<td>-1.37</td>
<td>0.10</td>
<td>&lt;.001</td>
<td>-0.62</td>
<td>0.04</td>
</tr>
<tr>
<td>02-2108300-00162</td>
<td>4</td>
<td>-20.57</td>
<td>48.09</td>
<td>0.33</td>
<td>-0.13</td>
<td>0.31</td>
</tr>
<tr>
<td>02-2003320-00141</td>
<td>4</td>
<td>-1.32</td>
<td>0.13</td>
<td>&lt;.001</td>
<td>-0.45</td>
<td>0.04</td>
</tr>
<tr>
<td>02-2103300-00196</td>
<td>3</td>
<td>-38.18</td>
<td>69.16</td>
<td>0.29</td>
<td>-0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>02-2106350-00166</td>
<td>3</td>
<td>50.38</td>
<td>457.51</td>
<td>0.46</td>
<td>0.02</td>
<td>0.18</td>
</tr>
<tr>
<td>02-2107310-00040</td>
<td>5</td>
<td>42.91</td>
<td>339.43</td>
<td>0.45</td>
<td>0.03</td>
<td>0.20</td>
</tr>
<tr>
<td>02-1211300-00087</td>
<td>6</td>
<td>-3.41</td>
<td>2.66</td>
<td>0.10</td>
<td>-0.29</td>
<td>0.23</td>
</tr>
<tr>
<td>02-2000350-00126</td>
<td>3</td>
<td>-4.11</td>
<td>1.83</td>
<td>0.01</td>
<td>-0.23</td>
<td>0.10</td>
</tr>
<tr>
<td>02-2108300-00065</td>
<td>4</td>
<td>2.52</td>
<td>2.61</td>
<td>0.17</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>02-2002500-00032</td>
<td>3</td>
<td>-3.48</td>
<td>2.45</td>
<td>0.08</td>
<td>-0.27</td>
<td>0.19</td>
</tr>
<tr>
<td>02-2106350-00152</td>
<td>4</td>
<td>28.83</td>
<td>176.46</td>
<td>0.44</td>
<td>0.03</td>
<td>0.17</td>
</tr>
<tr>
<td>02-2002520-00206</td>
<td>3</td>
<td>0.06</td>
<td>0.16</td>
<td>0.34</td>
<td>-1.01</td>
<td>0.23</td>
</tr>
<tr>
<td>02-1201310-00016</td>
<td>4</td>
<td>79.30</td>
<td>959.83</td>
<td>0.47</td>
<td>0.02</td>
<td>0.30</td>
</tr>
<tr>
<td>02-2001340-00163</td>
<td>4</td>
<td>-29.47</td>
<td>39.66</td>
<td>0.23</td>
<td>-0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>02-2107310-00047</td>
<td>3</td>
<td>2.17</td>
<td>1.37</td>
<td>0.06</td>
<td>0.30</td>
<td>0.18</td>
</tr>
<tr>
<td>02-1210300-00072</td>
<td>4</td>
<td>-11.83</td>
<td>9.47</td>
<td>0.11</td>
<td>-0.11</td>
<td>0.09</td>
</tr>
<tr>
<td>02-2108300-00138</td>
<td>4</td>
<td>1.05</td>
<td>0.69</td>
<td>0.06</td>
<td>0.31</td>
<td>0.16</td>
</tr>
<tr>
<td>02-1006000-000227</td>
<td>3</td>
<td>-1.03</td>
<td>0.62</td>
<td>0.05</td>
<td>0.34</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1006000-00232</td>
<td>3</td>
<td>0.70</td>
<td>0.51</td>
<td>0.09</td>
<td>0.35</td>
<td>0.17</td>
</tr>
<tr>
<td>02-2107310-00131</td>
<td>4</td>
<td>2.05</td>
<td>1.11</td>
<td>0.03</td>
<td>0.36</td>
<td>0.19</td>
</tr>
<tr>
<td>02-2002510-00148</td>
<td>4</td>
<td>-6.87</td>
<td>5.67</td>
<td>0.11</td>
<td>-0.30</td>
<td>0.25</td>
</tr>
<tr>
<td>02-2000360-00038</td>
<td>4</td>
<td>3.45</td>
<td>1.61</td>
<td>0.02</td>
<td>0.20</td>
<td>0.09</td>
</tr>
<tr>
<td>02-1006000-00243</td>
<td>3</td>
<td>-1.46</td>
<td>0.72</td>
<td>0.02</td>
<td>0.38</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1210300-00551</td>
<td>3</td>
<td>-21.40</td>
<td>120.80</td>
<td>0.43</td>
<td>-0.03</td>
<td>0.15</td>
</tr>
<tr>
<td>02-2106350-00120</td>
<td>3</td>
<td>-2.59</td>
<td>0.69</td>
<td>&lt;.001</td>
<td>0.23</td>
<td>0.06</td>
</tr>
<tr>
<td>02-2002510-00156</td>
<td>3</td>
<td>-5.32</td>
<td>0.91</td>
<td>&lt;.001</td>
<td>-0.26</td>
<td>0.04</td>
</tr>
<tr>
<td>02-2000360-00002</td>
<td>3</td>
<td>0.09</td>
<td>0.31</td>
<td>0.39</td>
<td>0.25</td>
<td>0.09</td>
</tr>
<tr>
<td>UIN</td>
<td>Number of Reviews for Item</td>
<td>Difficulty Estimate</td>
<td>Difficulty Standard Error</td>
<td>Slope Estimate</td>
<td>Slope Standard Error</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------</td>
<td>---------------------</td>
<td>---------------------------</td>
<td>----------------</td>
<td>---------------------</td>
<td></td>
</tr>
<tr>
<td>02-1211300-001101</td>
<td>4</td>
<td>10.58</td>
<td>25.97</td>
<td>0.34</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>02-1100000-00212</td>
<td>3</td>
<td>0.74</td>
<td>0.19</td>
<td>&lt;.001</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>02-2000360-00053</td>
<td>3</td>
<td>1.77</td>
<td>0.85</td>
<td>0.02</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>02-2000410-00146</td>
<td>3</td>
<td>84.45</td>
<td>1051.00</td>
<td>0.47</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>02-2107310-00036</td>
<td>4</td>
<td>0.74</td>
<td>0.44</td>
<td>0.05</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>02-2106350-00172</td>
<td>5</td>
<td>1.91</td>
<td>0.41</td>
<td>&lt;.001</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>02-1210300-00071</td>
<td>3</td>
<td>28.00</td>
<td>96.57</td>
<td>0.39</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00085</td>
<td>5</td>
<td>-0.16</td>
<td>0.29</td>
<td>0.29</td>
<td>0.47</td>
<td></td>
</tr>
<tr>
<td>02-1006000-00239</td>
<td>3</td>
<td>-0.70</td>
<td>0.36</td>
<td>0.03</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00187</td>
<td>4</td>
<td>1.58</td>
<td>0.53</td>
<td>0.00</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>02-1201310-00001</td>
<td>3</td>
<td>-1.23</td>
<td>1.09</td>
<td>0.13</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>02-2106350-00143</td>
<td>4</td>
<td>1.48</td>
<td>0.28</td>
<td>&lt;.001</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00141</td>
<td>4</td>
<td>-0.05</td>
<td>0.25</td>
<td>0.42</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>02-1211300-00073</td>
<td>3</td>
<td>7.73</td>
<td>8.17</td>
<td>0.17</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>02-1208290-00049</td>
<td>5</td>
<td>7.28</td>
<td>2.58</td>
<td>0.00</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>02-2106350-00153</td>
<td>7</td>
<td>1.36</td>
<td>0.67</td>
<td>0.02</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>02-1210300-00195</td>
<td>3</td>
<td>4.59</td>
<td>4.04</td>
<td>0.13</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00076</td>
<td>3</td>
<td>-0.07</td>
<td>0.23</td>
<td>0.38</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>02-1006300-00123</td>
<td>3</td>
<td>-1.90</td>
<td>0.53</td>
<td>0.00</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>02-2107310-00046</td>
<td>3</td>
<td>1.31</td>
<td>0.47</td>
<td>0.00</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>02-2000440-00032</td>
<td>3</td>
<td>1.39</td>
<td>0.45</td>
<td>0.00</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>02-2000360-00054</td>
<td>4</td>
<td>2.04</td>
<td>0.50</td>
<td>&lt;.001</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>02-2107310-00038</td>
<td>3</td>
<td>-0.45</td>
<td>0.26</td>
<td>0.04</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>02-1100000-00213</td>
<td>4</td>
<td>0.20</td>
<td>0.10</td>
<td>0.02</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>02-2107310-00043</td>
<td>5</td>
<td>-2.32</td>
<td>0.88</td>
<td>0.00</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>02-2000360-00002</td>
<td>6</td>
<td>0.10</td>
<td>0.15</td>
<td>0.26</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>02-2003310-00141</td>
<td>6</td>
<td>4.11</td>
<td>1.03</td>
<td>&lt;.001</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>02-2106350-00171</td>
<td>4</td>
<td>1.18</td>
<td>0.16</td>
<td>&lt;.001</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>02-2000440-00057</td>
<td>3</td>
<td>1.02</td>
<td>0.43</td>
<td>0.01</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>02-2002520-00215</td>
<td>4</td>
<td>-2.19</td>
<td>0.84</td>
<td>0.00</td>
<td>-0.51</td>
<td></td>
</tr>
<tr>
<td>02-2001340-00098</td>
<td>3</td>
<td>0.61</td>
<td>0.23</td>
<td>0.00</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>02-1211300-00118</td>
<td>3</td>
<td>10.34</td>
<td>21.68</td>
<td>0.32</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>02-2106350-00252</td>
<td>3</td>
<td>0.73</td>
<td>0.11</td>
<td>&lt;.001</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>02-2108300-00122</td>
<td>3</td>
<td>1.20</td>
<td>0.31</td>
<td>&lt;.001</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>02-2106350-00268</td>
<td>3</td>
<td>0.79</td>
<td>0.25</td>
<td>0.00</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>02-2000350-00057</td>
<td>3</td>
<td>2.17</td>
<td>0.38</td>
<td>&lt;.001</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>02-2106350-00170</td>
<td>3</td>
<td>-0.63</td>
<td>0.10</td>
<td>&lt;.001</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>02-2001340-00014</td>
<td>3</td>
<td>-1.58</td>
<td>0.55</td>
<td>0.00</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>02-2107310-00039</td>
<td>6</td>
<td>-0.53</td>
<td>0.22</td>
<td>0.01</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>UIN</td>
<td>Number of Reviews for Item</td>
<td>Difficulty Estimate</td>
<td>Difficulty Standard Error</td>
<td>Difficulty PrGT Estimate</td>
<td>Difficulty PrGT Standard Error</td>
<td>Slope Estimate</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------</td>
<td>---------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>-------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>02-1210300-00109</td>
<td>3</td>
<td>0.50</td>
<td>0.32</td>
<td>0.06</td>
<td>0.44</td>
<td>0.16</td>
</tr>
<tr>
<td>02-2000360-00063</td>
<td>3</td>
<td>-1.78</td>
<td>0.31</td>
<td>&lt;.001</td>
<td>0.67</td>
<td>0.12</td>
</tr>
<tr>
<td>02-1210300-00018</td>
<td>3</td>
<td>1.15</td>
<td>0.43</td>
<td>0.00</td>
<td>0.49</td>
<td>0.16</td>
</tr>
<tr>
<td>02-1006300-00205</td>
<td>4</td>
<td>-0.59</td>
<td>0.19</td>
<td>0.00</td>
<td>0.91</td>
<td>0.18</td>
</tr>
<tr>
<td>02-1210300-00039</td>
<td>3</td>
<td>1.26</td>
<td>0.45</td>
<td>0.00</td>
<td>0.50</td>
<td>0.16</td>
</tr>
<tr>
<td>02-1006300-00077</td>
<td>3</td>
<td>-0.19</td>
<td>0.16</td>
<td>0.12</td>
<td>0.94</td>
<td>0.18</td>
</tr>
<tr>
<td>02-1201310-00046</td>
<td>3</td>
<td>1.24</td>
<td>0.45</td>
<td>0.00</td>
<td>0.69</td>
<td>0.25</td>
</tr>
<tr>
<td>02-2003320-00044</td>
<td>3</td>
<td>-1.33</td>
<td>0.25</td>
<td>&lt;.001</td>
<td>0.65</td>
<td>0.12</td>
</tr>
<tr>
<td>02-1006300-00167</td>
<td>4</td>
<td>-0.85</td>
<td>0.19</td>
<td>&lt;.001</td>
<td>1.01</td>
<td>0.19</td>
</tr>
<tr>
<td>02-2108300-00055</td>
<td>4</td>
<td>-1.61</td>
<td>0.35</td>
<td>&lt;.001</td>
<td>1.02</td>
<td>0.26</td>
</tr>
<tr>
<td>02-1006000-00244</td>
<td>4</td>
<td>-0.27</td>
<td>0.16</td>
<td>0.05</td>
<td>1.03</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2003310-00102</td>
<td>4</td>
<td>0.93</td>
<td>0.12</td>
<td>&lt;.001</td>
<td>0.69</td>
<td>0.07</td>
</tr>
<tr>
<td>02-1202340-00141</td>
<td>3</td>
<td>6.99</td>
<td>3.79</td>
<td>0.03</td>
<td>0.83</td>
<td>0.51</td>
</tr>
<tr>
<td>02-2001340-00013</td>
<td>4</td>
<td>0.09</td>
<td>0.09</td>
<td>0.15</td>
<td>0.67</td>
<td>0.07</td>
</tr>
<tr>
<td>02-1006300-00144</td>
<td>4</td>
<td>-1.01</td>
<td>0.20</td>
<td>&lt;.001</td>
<td>1.10</td>
<td>0.20</td>
</tr>
<tr>
<td>02-2103300-00194</td>
<td>3</td>
<td>0.08</td>
<td>0.14</td>
<td>0.29</td>
<td>1.12</td>
<td>0.22</td>
</tr>
<tr>
<td>02-2106350-00169</td>
<td>3</td>
<td>-1.07</td>
<td>0.23</td>
<td>&lt;.001</td>
<td>1.03</td>
<td>0.23</td>
</tr>
<tr>
<td>02-1006300-00176</td>
<td>3</td>
<td>-0.16</td>
<td>0.14</td>
<td>0.12</td>
<td>1.20</td>
<td>0.20</td>
</tr>
<tr>
<td>02-2106350-00184</td>
<td>3</td>
<td>-0.69</td>
<td>0.16</td>
<td>&lt;.001</td>
<td>1.23</td>
<td>0.24</td>
</tr>
<tr>
<td>02-2106350-00089</td>
<td>3</td>
<td>-0.47</td>
<td>0.15</td>
<td>0.00</td>
<td>1.18</td>
<td>0.23</td>
</tr>
<tr>
<td>02-2000410-00034</td>
<td>4</td>
<td>0.44</td>
<td>0.11</td>
<td>&lt;.001</td>
<td>0.96</td>
<td>0.12</td>
</tr>
<tr>
<td>02-2000362-00001</td>
<td>3</td>
<td>-0.74</td>
<td>0.10</td>
<td>&lt;.001</td>
<td>1.27</td>
<td>0.16</td>
</tr>
<tr>
<td>02-2000410-00033</td>
<td>3</td>
<td>0.20</td>
<td>0.09</td>
<td>0.02</td>
<td>1.02</td>
<td>0.13</td>
</tr>
<tr>
<td>02-2106350-00187</td>
<td>4</td>
<td>-0.43</td>
<td>0.14</td>
<td>0.00</td>
<td>1.29</td>
<td>0.23</td>
</tr>
<tr>
<td>02-2002520-00013</td>
<td>6</td>
<td>-1.29</td>
<td>0.17</td>
<td>&lt;.001</td>
<td>1.03</td>
<td>0.14</td>
</tr>
<tr>
<td>02-1202340-00073</td>
<td>3</td>
<td>-0.57</td>
<td>0.08</td>
<td>&lt;.001</td>
<td>1.18</td>
<td>0.13</td>
</tr>
<tr>
<td>02-2108300-00053</td>
<td>5</td>
<td>-0.61</td>
<td>0.14</td>
<td>&lt;.001</td>
<td>1.53</td>
<td>0.30</td>
</tr>
<tr>
<td>02-1006000-00240</td>
<td>3</td>
<td>-1.35</td>
<td>0.20</td>
<td>&lt;.001</td>
<td>1.62</td>
<td>0.34</td>
</tr>
<tr>
<td>02-2106350-00270</td>
<td>3</td>
<td>-0.52</td>
<td>0.11</td>
<td>&lt;.001</td>
<td>1.85</td>
<td>0.31</td>
</tr>
<tr>
<td>02-2108300-00148</td>
<td>4</td>
<td>0.26</td>
<td>0.11</td>
<td>0.01</td>
<td>1.95</td>
<td>0.34</td>
</tr>
<tr>
<td>02-1006000-00199</td>
<td>4</td>
<td>-0.19</td>
<td>0.10</td>
<td>0.04</td>
<td>2.16</td>
<td>0.37</td>
</tr>
<tr>
<td>02-2002500-00150</td>
<td>4</td>
<td>-0.18</td>
<td>0.15</td>
<td>0.11</td>
<td>-1.18</td>
<td>0.25</td>
</tr>
<tr>
<td>02-2002500-00147</td>
<td>4</td>
<td>0.55</td>
<td>0.18</td>
<td>0.00</td>
<td>-1.02</td>
<td>0.24</td>
</tr>
<tr>
<td>02-2000410-00169</td>
<td>3</td>
<td>-0.30</td>
<td>0.18</td>
<td>0.05</td>
<td>-0.92</td>
<td>0.22</td>
</tr>
<tr>
<td>02-1006300-00069</td>
<td>4</td>
<td>-138.95</td>
<td>2451.00</td>
<td>0.48</td>
<td>-0.01</td>
<td>0.16</td>
</tr>
<tr>
<td>02-1006000-00167</td>
<td>3</td>
<td>3.58</td>
<td>5.64</td>
<td>0.26</td>
<td>0.11</td>
<td>0.16</td>
</tr>
<tr>
<td>02-1006000-00095</td>
<td>3</td>
<td>2.81</td>
<td>2.05</td>
<td>0.09</td>
<td>0.24</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1202340-00273</td>
<td>3</td>
<td>13.34</td>
<td>59.37</td>
<td>0.41</td>
<td>0.05</td>
<td>0.20</td>
</tr>
<tr>
<td>02-2106350-00083</td>
<td>4</td>
<td>3.08</td>
<td>2.16</td>
<td>0.08</td>
<td>0.26</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Table 30 (Continued)

<table>
<thead>
<tr>
<th>UIN</th>
<th>Number of Reviews</th>
<th>Difficulty Estimate</th>
<th>Difficulty Error</th>
<th>Difficulty Standard Error</th>
<th>Slope Estimate</th>
<th>Slope Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>02-1006000-00029</td>
<td>6</td>
<td>2.33</td>
<td>1.47</td>
<td>0.06</td>
<td>0.28</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1006000-00044</td>
<td>4</td>
<td>2.27</td>
<td>1.22</td>
<td>0.03</td>
<td>0.29</td>
<td>0.15</td>
</tr>
<tr>
<td>02-1006000-00172</td>
<td>3</td>
<td>1.46</td>
<td>0.92</td>
<td>0.06</td>
<td>0.30</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1006000-00066</td>
<td>4</td>
<td>0.07</td>
<td>0.46</td>
<td>0.44</td>
<td>0.30</td>
<td>0.16</td>
</tr>
<tr>
<td>02-1006000-00101</td>
<td>4</td>
<td>1.56</td>
<td>0.92</td>
<td>0.04</td>
<td>0.32</td>
<td>0.17</td>
</tr>
<tr>
<td>02-2106350-00115</td>
<td>3</td>
<td>-0.26</td>
<td>0.45</td>
<td>0.28</td>
<td>0.31</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1006000-00157</td>
<td>3</td>
<td>2.00</td>
<td>1.10</td>
<td>0.03</td>
<td>0.33</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1006000-00020</td>
<td>3</td>
<td>-2.18</td>
<td>1.09</td>
<td>0.02</td>
<td>0.36</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1006300-00042</td>
<td>3</td>
<td>0.43</td>
<td>0.40</td>
<td>0.14</td>
<td>0.36</td>
<td>0.14</td>
</tr>
<tr>
<td>02-1006000-00031</td>
<td>4</td>
<td>-0.10</td>
<td>0.38</td>
<td>0.40</td>
<td>0.37</td>
<td>0.17</td>
</tr>
<tr>
<td>02-2002500-00063</td>
<td>3</td>
<td>-1.20</td>
<td>0.60</td>
<td>0.02</td>
<td>-0.43</td>
<td>0.18</td>
</tr>
<tr>
<td>02-1006000-00050</td>
<td>3</td>
<td>0.60</td>
<td>0.42</td>
<td>0.07</td>
<td>0.41</td>
<td>0.17</td>
</tr>
<tr>
<td>02-2002500-00046</td>
<td>4</td>
<td>0.72</td>
<td>0.39</td>
<td>0.03</td>
<td>0.42</td>
<td>0.15</td>
</tr>
<tr>
<td>02-2002510-00139</td>
<td>3</td>
<td>-5.17</td>
<td>7.64</td>
<td>0.25</td>
<td>-0.12</td>
<td>0.18</td>
</tr>
<tr>
<td>02-1006300-00036</td>
<td>4</td>
<td>2.90</td>
<td>1.14</td>
<td>0.01</td>
<td>0.42</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1006000-00127</td>
<td>3</td>
<td>-0.97</td>
<td>0.48</td>
<td>0.02</td>
<td>0.43</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1210300-00134</td>
<td>3</td>
<td>22.70</td>
<td>87.74</td>
<td>0.40</td>
<td>0.22</td>
<td>0.87</td>
</tr>
<tr>
<td>02-1006000-00099</td>
<td>5</td>
<td>1.25</td>
<td>0.47</td>
<td>0.00</td>
<td>0.54</td>
<td>0.18</td>
</tr>
<tr>
<td>02-1006000-00011</td>
<td>4</td>
<td>0.03</td>
<td>0.26</td>
<td>0.45</td>
<td>0.54</td>
<td>0.17</td>
</tr>
<tr>
<td>02-2002510-00042</td>
<td>3</td>
<td>-2.48</td>
<td>1.77</td>
<td>0.08</td>
<td>-0.26</td>
<td>0.18</td>
</tr>
<tr>
<td>02-1006000-00119</td>
<td>4</td>
<td>0.98</td>
<td>0.37</td>
<td>0.00</td>
<td>0.58</td>
<td>0.18</td>
</tr>
<tr>
<td>02-2106350-00020</td>
<td>3</td>
<td>0.19</td>
<td>0.10</td>
<td>0.02</td>
<td>0.58</td>
<td>0.06</td>
</tr>
<tr>
<td>02-1006000-00036</td>
<td>6</td>
<td>0.35</td>
<td>0.26</td>
<td>0.09</td>
<td>0.61</td>
<td>0.18</td>
</tr>
<tr>
<td>02-2002510-00044</td>
<td>3</td>
<td>-5.15</td>
<td>5.26</td>
<td>0.16</td>
<td>-0.18</td>
<td>0.19</td>
</tr>
<tr>
<td>02-1006300-00043</td>
<td>3</td>
<td>-1.65</td>
<td>0.43</td>
<td>&lt;.001</td>
<td>0.64</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1210300-00105</td>
<td>3</td>
<td>3.57</td>
<td>1.50</td>
<td>0.01</td>
<td>0.44</td>
<td>0.19</td>
</tr>
<tr>
<td>02-1210300-00107</td>
<td>3</td>
<td>4.69</td>
<td>2.25</td>
<td>0.02</td>
<td>0.44</td>
<td>0.22</td>
</tr>
<tr>
<td>02-1006000-00070</td>
<td>3</td>
<td>-0.04</td>
<td>0.22</td>
<td>0.44</td>
<td>0.65</td>
<td>0.18</td>
</tr>
<tr>
<td>02-1006300-00060</td>
<td>4</td>
<td>-0.75</td>
<td>0.25</td>
<td>0.00</td>
<td>0.70</td>
<td>0.16</td>
</tr>
<tr>
<td>02-1006300-00047</td>
<td>3</td>
<td>-0.53</td>
<td>0.22</td>
<td>0.01</td>
<td>0.71</td>
<td>0.16</td>
</tr>
<tr>
<td>02-1006000-00014</td>
<td>3</td>
<td>-1.53</td>
<td>0.43</td>
<td>0.00</td>
<td>0.72</td>
<td>0.20</td>
</tr>
<tr>
<td>02-2106350-00012</td>
<td>4</td>
<td>-1.00</td>
<td>0.29</td>
<td>0.00</td>
<td>0.73</td>
<td>0.19</td>
</tr>
<tr>
<td>02-1210310-00025</td>
<td>3</td>
<td>1.35</td>
<td>0.24</td>
<td>&lt;.001</td>
<td>0.50</td>
<td>0.08</td>
</tr>
<tr>
<td>02-1006000-00189</td>
<td>7</td>
<td>2.13</td>
<td>0.56</td>
<td>&lt;.001</td>
<td>0.82</td>
<td>0.24</td>
</tr>
<tr>
<td>02-1210300-00179</td>
<td>3</td>
<td>0.68</td>
<td>0.28</td>
<td>0.01</td>
<td>0.56</td>
<td>0.17</td>
</tr>
<tr>
<td>02-1006000-00040</td>
<td>4</td>
<td>-0.90</td>
<td>0.25</td>
<td>0.00</td>
<td>0.84</td>
<td>0.20</td>
</tr>
<tr>
<td>02-1006000-00154</td>
<td>3</td>
<td>-0.37</td>
<td>0.19</td>
<td>0.02</td>
<td>0.87</td>
<td>0.20</td>
</tr>
<tr>
<td>02-1006000-00090</td>
<td>3</td>
<td>-0.86</td>
<td>0.23</td>
<td>0.00</td>
<td>0.89</td>
<td>0.21</td>
</tr>
<tr>
<td>02-1006300-00016</td>
<td>3</td>
<td>-1.17</td>
<td>0.25</td>
<td>&lt;.001</td>
<td>0.89</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Table 30 (Continued)

<table>
<thead>
<tr>
<th>UIN</th>
<th>Number of Reviews for Item</th>
<th>Difficulty Estimate</th>
<th>Difficulty Standard Error</th>
<th>PrGTt</th>
<th>Slope Estimate</th>
<th>Slope Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>02-1006000-00174</td>
<td>7</td>
<td>-0.73</td>
<td>0.22</td>
<td>0.00</td>
<td>0.89</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2002500-00149</td>
<td>4</td>
<td>-2.24</td>
<td>1.83</td>
<td>0.11</td>
<td>0.17</td>
<td>0.13</td>
</tr>
<tr>
<td>02-1006300-00020</td>
<td>4</td>
<td>-0.35</td>
<td>0.16</td>
<td>0.01</td>
<td>1.02</td>
<td>0.19</td>
</tr>
<tr>
<td>02-1006000-00108</td>
<td>3</td>
<td>-0.72</td>
<td>0.19</td>
<td>&lt;.001</td>
<td>1.06</td>
<td>0.22</td>
</tr>
<tr>
<td>02-1210300-00108</td>
<td>3</td>
<td>-0.71</td>
<td>0.20</td>
<td>0.00</td>
<td>0.86</td>
<td>0.20</td>
</tr>
<tr>
<td>02-1006300-00052</td>
<td>3</td>
<td>-0.30</td>
<td>0.15</td>
<td>0.02</td>
<td>1.08</td>
<td>0.19</td>
</tr>
<tr>
<td>02-2000360-00065</td>
<td>3</td>
<td>4.01</td>
<td>1.38</td>
<td>0.00</td>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>02-2106350-00014</td>
<td>3</td>
<td>0.51</td>
<td>0.06</td>
<td>&lt;.001</td>
<td>1.11</td>
<td>0.08</td>
</tr>
<tr>
<td>02-2106350-00015</td>
<td>4</td>
<td>-0.20</td>
<td>0.06</td>
<td>0.00</td>
<td>1.12</td>
<td>0.08</td>
</tr>
<tr>
<td>02-1006000-00080</td>
<td>3</td>
<td>-1.19</td>
<td>0.24</td>
<td>&lt;.001</td>
<td>1.14</td>
<td>0.25</td>
</tr>
<tr>
<td>02-1210300-00141</td>
<td>3</td>
<td>-0.31</td>
<td>0.15</td>
<td>0.02</td>
<td>1.15</td>
<td>0.23</td>
</tr>
<tr>
<td>02-1210300-00208</td>
<td>3</td>
<td>4.34</td>
<td>1.64</td>
<td>0.00</td>
<td>1.23</td>
<td>0.61</td>
</tr>
<tr>
<td>02-2000350-00113</td>
<td>3</td>
<td>2.01</td>
<td>0.41</td>
<td>&lt;.001</td>
<td>0.52</td>
<td>0.11</td>
</tr>
<tr>
<td>02-2106350-00093</td>
<td>4</td>
<td>-1.21</td>
<td>0.08</td>
<td>&lt;.001</td>
<td>1.34</td>
<td>0.11</td>
</tr>
<tr>
<td>02-1211300-00019</td>
<td>3</td>
<td>2.92</td>
<td>0.33</td>
<td>&lt;.001</td>
<td>1.16</td>
<td>0.18</td>
</tr>
<tr>
<td>02-1006000-00062</td>
<td>3</td>
<td>-1.29</td>
<td>0.22</td>
<td>&lt;.001</td>
<td>1.43</td>
<td>0.30</td>
</tr>
<tr>
<td>02-1006000-00120</td>
<td>4</td>
<td>0.10</td>
<td>0.13</td>
<td>0.20</td>
<td>1.48</td>
<td>0.27</td>
</tr>
<tr>
<td>02-1006300-00045</td>
<td>4</td>
<td>-0.73</td>
<td>0.13</td>
<td>&lt;.001</td>
<td>1.61</td>
<td>0.25</td>
</tr>
<tr>
<td>02-1006000-00122</td>
<td>3</td>
<td>-1.11</td>
<td>0.16</td>
<td>&lt;.001</td>
<td>1.84</td>
<td>0.36</td>
</tr>
<tr>
<td>02-2000360-00110</td>
<td>3</td>
<td>0.56</td>
<td>0.10</td>
<td>&lt;.001</td>
<td>1.05</td>
<td>0.13</td>
</tr>
<tr>
<td>02-2008300-00200</td>
<td>6</td>
<td>-1.64</td>
<td>0.43</td>
<td>&lt;.001</td>
<td>-0.77</td>
<td>0.22</td>
</tr>
<tr>
<td>02-2003310-00153</td>
<td>3</td>
<td>-0.71</td>
<td>0.05</td>
<td>&lt;.0001</td>
<td>-0.96</td>
<td>0.05</td>
</tr>
<tr>
<td>02-2001340-00219</td>
<td>4</td>
<td>-0.20</td>
<td>0.14</td>
<td>0.08</td>
<td>-1.29</td>
<td>0.26</td>
</tr>
<tr>
<td>02-2108300-00166</td>
<td>3</td>
<td>-1.12</td>
<td>0.92</td>
<td>0.11</td>
<td>-0.23</td>
<td>0.16</td>
</tr>
<tr>
<td>02-2003310-00159</td>
<td>4</td>
<td>-0.11</td>
<td>0.04</td>
<td>0.00</td>
<td>-1.11</td>
<td>0.05</td>
</tr>
<tr>
<td>02-2107300-00178</td>
<td>3</td>
<td>-15.72</td>
<td>34.73</td>
<td>0.33</td>
<td>-0.12</td>
<td>0.27</td>
</tr>
<tr>
<td>02-2003320-00021</td>
<td>5</td>
<td>-1.57</td>
<td>0.13</td>
<td>&lt;.001</td>
<td>-0.54</td>
<td>0.04</td>
</tr>
<tr>
<td>02-2107310-00204</td>
<td>3</td>
<td>0.81</td>
<td>1.01</td>
<td>0.21</td>
<td>-0.20</td>
<td>0.18</td>
</tr>
<tr>
<td>02-2008300-00128</td>
<td>4</td>
<td>-12.84</td>
<td>37.35</td>
<td>0.37</td>
<td>-0.06</td>
<td>0.16</td>
</tr>
<tr>
<td>02-2107310-00160</td>
<td>3</td>
<td>-6.10</td>
<td>7.24</td>
<td>0.20</td>
<td>-0.17</td>
<td>0.21</td>
</tr>
<tr>
<td>02-1211300-00079</td>
<td>3</td>
<td>0.94</td>
<td>0.19</td>
<td>&lt;.001</td>
<td>0.49</td>
<td>0.08</td>
</tr>
<tr>
<td>02-1202340-00158</td>
<td>3</td>
<td>3.25</td>
<td>1.61</td>
<td>0.02</td>
<td>0.51</td>
<td>0.27</td>
</tr>
<tr>
<td>02-1211300-00184</td>
<td>4</td>
<td>0.43</td>
<td>0.13</td>
<td>0.00</td>
<td>0.55</td>
<td>0.08</td>
</tr>
<tr>
<td>02-2107310-00161</td>
<td>3</td>
<td>-30.27</td>
<td>148.74</td>
<td>0.42</td>
<td>-0.04</td>
<td>0.22</td>
</tr>
<tr>
<td>02-2108300-00193</td>
<td>3</td>
<td>8.07</td>
<td>19.12</td>
<td>0.34</td>
<td>0.07</td>
<td>0.16</td>
</tr>
<tr>
<td>02-2107310-00031</td>
<td>5</td>
<td>2.73</td>
<td>5.68</td>
<td>0.32</td>
<td>0.09</td>
<td>0.18</td>
</tr>
<tr>
<td>02-2108300-00127</td>
<td>3</td>
<td>4.93</td>
<td>6.94</td>
<td>0.24</td>
<td>0.11</td>
<td>0.16</td>
</tr>
<tr>
<td>02-1202340-00107</td>
<td>4</td>
<td>-0.04</td>
<td>0.11</td>
<td>0.34</td>
<td>0.61</td>
<td>0.09</td>
</tr>
<tr>
<td>UIN</td>
<td>Number of Reviews for Item</td>
<td>Difficulty Estimate</td>
<td>Difficulty Standard Error</td>
<td>PrGTt Estimate</td>
<td>Slope Estimate</td>
<td>Slope Standard Error</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------------------</td>
<td>---------------------</td>
<td>---------------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>02-2001340-00086</td>
<td>3</td>
<td>-5.07</td>
<td>3.65</td>
<td>0.08</td>
<td>-0.29</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2107300-00177</td>
<td>3</td>
<td>1.37</td>
<td>1.25</td>
<td>0.14</td>
<td>0.22</td>
<td>0.18</td>
</tr>
<tr>
<td>02-2107310-00188</td>
<td>3</td>
<td>11.89</td>
<td>19.31</td>
<td>0.27</td>
<td>0.14</td>
<td>0.23</td>
</tr>
<tr>
<td>02-2002510-00238</td>
<td>5</td>
<td>2.92</td>
<td>1.57</td>
<td>0.03</td>
<td>0.38</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2000440-00267</td>
<td>4</td>
<td>-12.48</td>
<td>23.03</td>
<td>0.29</td>
<td>-0.11</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2107310-00069</td>
<td>3</td>
<td>2.27</td>
<td>1.96</td>
<td>0.12</td>
<td>0.22</td>
<td>0.18</td>
</tr>
<tr>
<td>02-2106350-00233</td>
<td>3</td>
<td>0.31</td>
<td>0.09</td>
<td>0.00</td>
<td>0.34</td>
<td>0.04</td>
</tr>
<tr>
<td>02-2107300-00189</td>
<td>3</td>
<td>2.52</td>
<td>1.40</td>
<td>0.04</td>
<td>0.34</td>
<td>0.19</td>
</tr>
<tr>
<td>02-2107300-00186</td>
<td>4</td>
<td>-0.95</td>
<td>0.65</td>
<td>0.07</td>
<td>0.35</td>
<td>0.19</td>
</tr>
<tr>
<td>02-2106350-00232</td>
<td>3</td>
<td>3.76</td>
<td>0.65</td>
<td>&lt;.001</td>
<td>0.41</td>
<td>0.07</td>
</tr>
<tr>
<td>02-2108300-00169</td>
<td>3</td>
<td>0.99</td>
<td>0.49</td>
<td>0.02</td>
<td>0.42</td>
<td>0.17</td>
</tr>
<tr>
<td>02-2107300-00205</td>
<td>3</td>
<td>1.33</td>
<td>0.60</td>
<td>0.01</td>
<td>0.47</td>
<td>0.19</td>
</tr>
<tr>
<td>02-1208920-00001</td>
<td>4</td>
<td>0.53</td>
<td>0.11</td>
<td>&lt;.001</td>
<td>0.62</td>
<td>0.09</td>
</tr>
<tr>
<td>02-2107300-00182</td>
<td>3</td>
<td>-0.05</td>
<td>0.30</td>
<td>0.43</td>
<td>0.50</td>
<td>0.20</td>
</tr>
<tr>
<td>02-2107310-00144</td>
<td>3</td>
<td>2.25</td>
<td>1.11</td>
<td>0.02</td>
<td>0.40</td>
<td>0.19</td>
</tr>
<tr>
<td>02-1201310-00139</td>
<td>3</td>
<td>1.35</td>
<td>0.13</td>
<td>&lt;.001</td>
<td>1.05</td>
<td>0.11</td>
</tr>
<tr>
<td>02-2107310-00002</td>
<td>4</td>
<td>-0.28</td>
<td>0.28</td>
<td>0.16</td>
<td>0.56</td>
<td>0.20</td>
</tr>
<tr>
<td>02-2003310 Physical Science-00009</td>
<td>3</td>
<td>-5.23</td>
<td>4.41</td>
<td>0.12</td>
<td>0.17</td>
<td>0.14</td>
</tr>
<tr>
<td>02-2107310-00162</td>
<td>4</td>
<td>1.31</td>
<td>0.57</td>
<td>0.01</td>
<td>0.50</td>
<td>0.19</td>
</tr>
<tr>
<td>02-2106350-00106</td>
<td>3</td>
<td>1.30</td>
<td>0.15</td>
<td>&lt;.001</td>
<td>0.63</td>
<td>0.07</td>
</tr>
<tr>
<td>02-2107310-00009</td>
<td>3</td>
<td>1.50</td>
<td>0.52</td>
<td>0.00</td>
<td>0.62</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2107300-00190</td>
<td>3</td>
<td>0.73</td>
<td>0.32</td>
<td>0.01</td>
<td>0.65</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2106350-00201</td>
<td>4</td>
<td>0.78</td>
<td>0.11</td>
<td>&lt;.001</td>
<td>0.67</td>
<td>0.07</td>
</tr>
<tr>
<td>02-2108300-00185</td>
<td>3</td>
<td>-0.70</td>
<td>0.25</td>
<td>0.00</td>
<td>0.71</td>
<td>0.19</td>
</tr>
<tr>
<td>02-2107300-00206</td>
<td>3</td>
<td>0.48</td>
<td>0.25</td>
<td>0.03</td>
<td>0.73</td>
<td>0.22</td>
</tr>
<tr>
<td>02-2107300-00193</td>
<td>3</td>
<td>1.27</td>
<td>0.39</td>
<td>0.00</td>
<td>0.73</td>
<td>0.22</td>
</tr>
<tr>
<td>02-2106350-00231</td>
<td>3</td>
<td>-0.47</td>
<td>0.08</td>
<td>&lt;.001</td>
<td>0.75</td>
<td>0.07</td>
</tr>
<tr>
<td>02-2107310-00074</td>
<td>3</td>
<td>0.81</td>
<td>0.34</td>
<td>0.01</td>
<td>0.65</td>
<td>0.21</td>
</tr>
<tr>
<td>02-2001340-00141</td>
<td>4</td>
<td>4.66</td>
<td>0.68</td>
<td>&lt;.001</td>
<td>0.33</td>
<td>0.05</td>
</tr>
<tr>
<td>02-2107310-00004</td>
<td>3</td>
<td>0.51</td>
<td>0.28</td>
<td>0.03</td>
<td>0.67</td>
<td>0.21</td>
</tr>
<tr>
<td>02-200520-00001</td>
<td>3</td>
<td>-6.58</td>
<td>2.14</td>
<td>0.00</td>
<td>-0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>02-2108300-00186</td>
<td>3</td>
<td>-0.83</td>
<td>0.23</td>
<td>0.00</td>
<td>0.85</td>
<td>0.20</td>
</tr>
<tr>
<td>02-2106350-00215</td>
<td>5</td>
<td>-0.46</td>
<td>0.18</td>
<td>0.00</td>
<td>0.93</td>
<td>0.20</td>
</tr>
<tr>
<td>02-2107310-00013</td>
<td>3</td>
<td>0.71</td>
<td>0.24</td>
<td>0.00</td>
<td>0.92</td>
<td>0.24</td>
</tr>
<tr>
<td>02-2106350-00056</td>
<td>3</td>
<td>-0.73</td>
<td>0.08</td>
<td>&lt;.001</td>
<td>0.99</td>
<td>0.08</td>
</tr>
<tr>
<td>02-2106350-00236</td>
<td>3</td>
<td>0.21</td>
<td>0.06</td>
<td>0.00</td>
<td>1.00</td>
<td>0.08</td>
</tr>
<tr>
<td>02-2001340-00002</td>
<td>4</td>
<td>-4.93</td>
<td>2.13</td>
<td>0.01</td>
<td>0.55</td>
<td>0.26</td>
</tr>
<tr>
<td>02-2107310-00028</td>
<td>3</td>
<td>0.48</td>
<td>0.20</td>
<td>0.01</td>
<td>0.99</td>
<td>0.26</td>
</tr>
<tr>
<td>02-2107310-00075</td>
<td>3</td>
<td>4.12</td>
<td>1.51</td>
<td>0.00</td>
<td>0.90</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Table 30 (Continued)

<table>
<thead>
<tr>
<th>UIN</th>
<th>Number of Reviews</th>
<th>Difficulty Estimate</th>
<th>Difficulty Standard Error</th>
<th>PrGT1</th>
<th>Slope Estimate</th>
<th>Slope Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>02-2107310-00002</td>
<td>3</td>
<td>2.18</td>
<td>0.53</td>
<td>&lt;.001</td>
<td>0.91</td>
<td>0.26</td>
</tr>
<tr>
<td>02-2003310-00149</td>
<td>4</td>
<td>0.16</td>
<td>0.11</td>
<td>0.07</td>
<td>0.54</td>
<td>0.07</td>
</tr>
<tr>
<td>02-2108300-00006</td>
<td>3</td>
<td>-0.66</td>
<td>0.17</td>
<td>&lt;.001</td>
<td>1.09</td>
<td>0.23</td>
</tr>
<tr>
<td>02-2107310-00029</td>
<td>3</td>
<td>0.24</td>
<td>0.17</td>
<td>0.08</td>
<td>1.11</td>
<td>0.28</td>
</tr>
<tr>
<td>02-1210300-00185</td>
<td>3</td>
<td>0.83</td>
<td>0.17</td>
<td>&lt;.001</td>
<td>1.24</td>
<td>0.27</td>
</tr>
<tr>
<td>02-2107310-00146</td>
<td>4</td>
<td>1.18</td>
<td>0.27</td>
<td>&lt;.001</td>
<td>1.12</td>
<td>0.27</td>
</tr>
<tr>
<td>02-2002510-00173</td>
<td>4</td>
<td>-1.13</td>
<td>0.08</td>
<td>&lt;.001</td>
<td>1.35</td>
<td>0.12</td>
</tr>
<tr>
<td>02-2106350-00057</td>
<td>3</td>
<td>-1.09</td>
<td>0.07</td>
<td>&lt;.001</td>
<td>1.39</td>
<td>0.11</td>
</tr>
<tr>
<td>02-2107300-00175</td>
<td>3</td>
<td>2.11</td>
<td>0.38</td>
<td>&lt;.001</td>
<td>1.44</td>
<td>0.36</td>
</tr>
<tr>
<td>02-1202340-00082</td>
<td>5</td>
<td>3.06</td>
<td>0.82</td>
<td>&lt;.001</td>
<td>4.10</td>
<td>3.45</td>
</tr>
</tbody>
</table>
### APPENDIX G:

**SAMPLE TEST ITEM SPECIFICATIONS/SAMPLE ITEMS USED BY PROJECT**

<table>
<thead>
<tr>
<th>Course</th>
<th>Liberal Arts Mathematics 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reporting Category</td>
<td>Algebra</td>
</tr>
<tr>
<td>Standard</td>
<td>Creating Equations</td>
</tr>
<tr>
<td>Benchmark Number</td>
<td>MAFS.912.A-CED.1.3</td>
</tr>
<tr>
<td>Benchmark</td>
<td>Represent constraints by equations or inequalities, and by systems of equations and/or inequalities, and interpret solutions as viable or nonviable options in a modeling context. For example, represent inequalities describing nutritional and cost constraints on combinations of different foods.</td>
</tr>
<tr>
<td>Item Types</td>
<td>Selected Response (Multiple Choice), Gridded Response, Short Answer</td>
</tr>
<tr>
<td>Cognitive Complexity</td>
<td>High</td>
</tr>
<tr>
<td>Benchmark Clarification</td>
<td>Students will solve systems of equations and inequalities and be able to interpret solutions in terms of the real world context. Additionally, students will be able to interpret the domain and ranges of viable solutions both in terms of the real world context as well as any potential mathematical constraints.</td>
</tr>
<tr>
<td>Content Limits</td>
<td>Items will not assess rational functions. Systems will be limited to linear and exponential equations. Items will not assess systems of linear equations in three variables.</td>
</tr>
<tr>
<td>Stimulus Attributes</td>
<td>While items may be set in a mathematical context, they should focus on real world modeling situations.</td>
</tr>
<tr>
<td>Response Attributes</td>
<td>Not Applicable.</td>
</tr>
</tbody>
</table>
| **Sample Item**         | 1. The cost of 3 large candles and 5 small candles is $16.40. The cost of 4 large candles and 6 small candles is $17.50. Which pair of equations can be used to determine, \( t \), the cost of a large candle, and \( s \), the cost of a small candle?  
   A. \( 3t + 5s = 16.40 \)  
   \( 4t + 6s = 17.50 \)  
   B. \( t + s = 10 \)  
   \( 4t + 6s = 17.50 \)  
   C. \( 3t + 5s = 16.40 \)  
   \( t + s = 8 \) |
<table>
<thead>
<tr>
<th>Course</th>
<th>Forensic Science 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reporting Category</td>
<td>Life Science</td>
</tr>
<tr>
<td>Standard</td>
<td>Organization &amp; Development of Living Organisms</td>
</tr>
<tr>
<td>Benchmark Number</td>
<td>SC.912.L.14.34</td>
</tr>
<tr>
<td>Benchmark</td>
<td>Describe the composition and physiology of blood, including that of the plasma and the formed elements.</td>
</tr>
<tr>
<td>Also Assesses</td>
<td>N/A</td>
</tr>
<tr>
<td>Item Types</td>
<td>Multiple Choice</td>
</tr>
</tbody>
</table>
| Benchmark Clarification | Students will describe the components of blood and the role they play in the human body.  
Students will recognize these components include red blood cells, white blood cells, plasma, and platelets.  
Students will explain how blood carries oxygen, carbon dioxide and other elements essential to the proper functioning of the human organs and tissues. |
| Content Limits       | Items will not assess the specific proteins found in plasma.  
Items will not assess the percentage of plasma or elements of blood.  
Items will not assess blood as a connective tissue or the making of blood in bones. |
| Stimulus Attributes  | May contain graphics, charts, data tables. |
| Response Attributes  | None Specified     |
| Sample Item          | Because carbon monoxide binds easily to hemoglobin, replacing oxygen in the blood. An autopsy of a victim found in a burned building reveals normal levels of dissolved oxygen in the blood. What conclusion can be made based on this evidence?  
A) The victim died in the fire.  
B) The fire was intentionally started.  
C) The victim died before the fire started.  
D) The victim was murdered and the fire set to conceal the crime. |
<p>| Correct Answer: A    |                    |</p>
<table>
<thead>
<tr>
<th>Course</th>
<th>1301070/1301080 MJ Guitar 2,3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>Every art form uses its own unique language, verbal and non-verbal, to document and communicate with the world.</td>
</tr>
<tr>
<td>Benchmark Number</td>
<td>MU.68.O.3.1</td>
</tr>
<tr>
<td>Benchmark</td>
<td>Describe how the combination of instrumentation and expressive elements in a musical work can convey a specific thought, idea, mood, and/or image.</td>
</tr>
<tr>
<td>Also Assesses</td>
<td>MU.68.H.2.3</td>
</tr>
<tr>
<td>Item Type</td>
<td>Multiple Choice, Short Answer, Extended Response</td>
</tr>
<tr>
<td>Benchmark Clarification</td>
<td>Students will describe the potential artistic impact that a provided piece of music conveys. Student will also identify the musical elements that contribute to the overall image, idea or mood conveyed.</td>
</tr>
<tr>
<td>Content Limits</td>
<td>Guitar only. Excerpt needs to be 60-80 seconds long.</td>
</tr>
<tr>
<td>Stimulus Attributes</td>
<td>Audio Recording, Video</td>
</tr>
<tr>
<td>Response Attributes</td>
<td>Factors (not limited to): chord voicing, tempo, dynamics, articulation, chord progression, tonality, instrument selection.</td>
</tr>
<tr>
<td>Sample Item</td>
<td>Identify the mood of the following excerpt. Describe the style and techniques used to create the mood of this piece. List and discuss as many elements as possible.</td>
</tr>
</tbody>
</table>

0: student lists no elements or techniques
1: student lists 1-2 elements
2: student lists 3-4 elements
3: student lists 5-6 elements
4: student lists 7+ elements
Course Name: Creative Writing 1
Benchmark(s): LACC.910.RL.3.7

Test Administration Expectations:
1. The skills assessed by this item should be appropriately incorporated into the classroom instruction.
2. This assessment item should be administered in a safe and supervised classroom environment that conforms to all district policies, standards, and procedures.
3. IEP and 504 student plan accommodations may require additional assessment administration modifications (See Teacher Preparation Guidelines).

Description of the Performance Assessment
Students taking this performance assessment will respond to a performance prompt and a series of short-answer questions.

- Performance prompts asks the student to individually write an extended response based on the criteria outlined in the prompt.
- All student responses must be collected to facilitate scoring and to document each student’s performance.
- Short-answer questions will ask the students to supply a response which may be in the form of words, pictures, and/or diagrams in order to facilitate scoring and to document each student’s response (See Teacher Preparation Guidelines for verbal responses and other accommodations).
- Response sheets are provided for student work. All written work must be completed in the student answer spaces provided.

Materials and Resources
Teachers will need the following materials and resources for students in order to complete this performance assessment:
1. classroom set of reproduced tasks, including the glossary of terms, and the cover page – page 4 must be printed in color
2. one copy of administration guidelines (pages 1-5)
3. classroom set of reproduced student response sheets
4. black/blue ink pens and pencils
5. recording equipment for student accommodations

Teacher Preparation Guidelines
1. This assessment requires an individual performance.
2. Reproduce a classroom set of student task directions, glossary of terms, and student response sheets found in the Student Task Booklet
3. Students may select to use either blue/black pen or pencil for their finished response.
4. Students should be reminded to take the assessment time constraints into consideration in completing their performance prompt response and answering the short-answer questions.
5. Students must include their name/number on all the response sheets.
6. It is recommended and encouraged that the teacher reviews the glossary and scoring rubrics with the students.

Suggestions for Time Management

Students may have as much time as they need to complete the task. Time suggestions are a guide and may be altered to meet individual school, class, and/or student circumstances. It is recommended and encouraged that the teacher reviews the glossary and scoring rubrics with the students.

The following two-day model is a suggested sample timeframe:

**Day One Suggested Time Frame:**
- 15 Minutes: The teacher provides the class with the task and reads it aloud. The students may ask questions. The teacher answers any questions asked and distributes all materials.
- 30 Minutes: The students have 30 minutes to start and answer short-answer questions.
- 10 Minutes: The teacher collects all materials.

**Day Two Suggested Time Frame:**
- 10 Minutes: The teacher distributes all materials to the students and reviews short-answer questions instructions.
- 45 Minutes: The students have 45 minutes to start and complete their extended response based on the criteria outlined in the prompt.
- 5 Minutes: The teacher collects all materials.

**Test Administration Directions**

Students may have as much time as they need to complete the task. All students who remain productively engaged in the task should be allowed to finish their work. In some instances, a few students may require considerably more time to complete the task than other students. In these cases, it is appropriate to relocate these students to a new location to finish. In other cases, the teacher’s knowledge of some students’ work habits or special needs may suggest that students who work very slowly should be tested separately or grouped with similar students for the performance assessment.

Provide the class with the reproduced student pages, which may include the cover page, student prompt, response sheets, rubrics, templates, glossary, and any other required materials prior to beginning the task. Students may highlight and write on these materials during the assessment. Instruct the students to look at the following student pages. Have the students read the directions to themselves as you read them aloud. Answer any clarifying questions the students may have before you instruct them to begin. If this assessment is used for reporting purposes, circle the scoring points on the cover page and/or on the individual student pages.

Say: **Today you will take the Grade 11 Florida Classroom-Based Performance Assessment (FCBPA) Creative Writing 1 Performance Assessment**
Directions:
Analyze Langston Hughes’ poem “Minstrel Man” and Jacob Lawrence’s painting “Tombstone” and complete the assignment that follows.

Assignment:
Art is a medium through which people express their emotions, culture, and ideas. There are many types of art. Among the different types of art are painting, sculpture, music, dance, architecture, and poetry. Below are Langston Hughes’ poem *Minstrel Man* and Jacob Lawrence’s painting *Tombstone*.

Both works of art depict scenes typical of the African-American condition during the Harlem Renaissance. Plan and write an essay in which you compare and contrast the tone of the works of art. In order to write your essay you SHOULD consider imagery, characterization, and mood.

In order to begin, you must first write a thesis statement. Then take notes that you can utilize to write a clear, well-supported, and persuasive essay. You must support your positions with facts and examples from your reading, studies, experiences, or observations.

---

*Minstrel Man* by Langston Hughes

Because my mouth
Is wide with laughter
And my throat
Is deep with song,
You do not think
I suffer after
I have held my pain
So long?

Because my mouth
Is wide with laughter,
You do not hear
My inner cry?
Because my feet
Are gay with dancing,
You do not know
I die?
Florida Classroom-Based Performance Assessment (FCBPA)
Creative Writing 1 Assessment

Glossary

Anaphora – The deliberate repetition of a word or phrase at the beginning of several successive verses, clauses, or paragraphs for rhetorical or poetic effect.

Characterization – the method a writer or artist uses to reveal the personality of a character in a work of art. Personality may be revealed (1) by what the character says about himself or herself; (2) by what others reveal about the character; (3) by the character's own actions; and (4) how the artist portrays the character.

Diction – is the author's choice of words. If an author chooses one word over another, it is probably because that word implies some social or connotative meaning. It is frequently divided into four levels: formal, informal, colloquial, and slang.

Enjambment – when one line of poetry ends without a pause and continues into the next line for its meaning. It is also called a run-on line.

Imagery – Imagery is language that appeals to the senses. It is description that makes the reader feel he or she is "in the setting." There are six basic kinds of imagery: visual (sight), auditory (sound), olfactory (smell), gustatory (taste), tactile (touch), and kinesthetic (movement).

Mood – is the feelings an author or artist creates in the reader or observer.

Personification – A form of metaphor in which human characteristics are attributed to nonhuman things. Personification allows the author a way to give the world life and motion by giving human characteristics to animals, inanimate objects, and abstract ideas.

Repetition – the repetition of a word, phrase, sound, or syllable in poetry for rhetorical effect.

Rhetorical question – a question asked solely to produce an effect or to make a statement, but not expected to receive an answer. The purpose to a rhetorical question is to make a deeper impression upon or provoke deeper though in the hearer or reader than a direct statement would.

Stanza – a related group of lines in a poem, equivalent to a paragraph in prose.

Syntax – the arrangement of words in a sentence, the grammar of a sentence.

Tone – The author’s attitude toward the people, places, and events in a work as revealed by the elements of the author’s style. Tone may be characterized as sad or happy, angry or affectionate, bitter or nostalgic, serious or lighthearted, sympathetic or cruel, or any other feelings and emotions that a person can experience.

Verse – a line of poetry, similar to a sentence in prose.
1. Langston Hughes’ *Minstrel Man* is structured into three rhetorical questions. What effect does Langston Hughes’ use of this rhetorical device have on the reader?

2. What is the point-of-view of Langston Hughes’ *Minstrel Man*? How does Hughes’ use of this point-of-view add to the poem?

3. Identify one literary or poetic device, other than rhetorical questions, used by Langston Hughes in *Minstrel Man* and detail how it adds to the tone and meaning of the poem?

4. What, if any, relationship exists between the subjects in Jacob Lawrence’s *Tombstone*? What could Jacob Lawrence be saying about life during the Harlem Renaissance?

5. What is the mood of Jacob Lawrence’s *Tombstone*? Provide three specific details from the painting that inspire that mood. How did these three details convey that mood to you?

Response to Performance Prompt Rubric

<table>
<thead>
<tr>
<th>Short Answer Test Assessment Rubric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong> 4 points</td>
</tr>
<tr>
<td>Unacceptable 0 points</td>
</tr>
<tr>
<td>Needs Improvement 1 point</td>
</tr>
<tr>
<td>Satisfactory 2 points</td>
</tr>
<tr>
<td>Meets Expectations 3 points</td>
</tr>
<tr>
<td>Exceptional 4 points</td>
</tr>
<tr>
<td>Did not answer question.</td>
</tr>
<tr>
<td>Answers are partial or incomplete.</td>
</tr>
<tr>
<td>Key points are not clear.</td>
</tr>
<tr>
<td>Question not adequately answered.</td>
</tr>
<tr>
<td>Answers are not comprehensive or</td>
</tr>
<tr>
<td>completely stated.</td>
</tr>
<tr>
<td>Key points are not well supported.</td>
</tr>
<tr>
<td>Answers are accurate and complete.</td>
</tr>
<tr>
<td>Key points are stated and supported.</td>
</tr>
<tr>
<td>Answers are comprehensive,</td>
</tr>
<tr>
<td>accurate and complete.</td>
</tr>
<tr>
<td>Key ideas are clearly stated,</td>
</tr>
<tr>
<td>explained, and well supported.</td>
</tr>
</tbody>
</table>

| **Organization** (Answers are clearly thought out and articulated) 4 points |
| Did not answer question. |
| Organization and structure detract from the answer |
| Inadequate organization or development. Structure of the answer is not easy to follow. |
| Organization is mostly clear and easy to follow. |
| Well organized, coherently developed, and easy to follow. |

| **Writing Conventions** (Spelling, punctuation, grammar, and complete sentences) 4 points |
| Did not answer question. |
| Displays over five errors in spelling, punctuation, grammar, and sentence structure. |
| Displays three to five errors in spelling, punctuation, grammar, and sentence structure. |
| Displays one to three errors in spelling, punctuation, grammar, and sentence structure. |
| Displays no errors in spelling, punctuation, grammar, and sentence structure. |
APPENDIX H:
FAIR USE WORKSHEET FOR REPRODUCTION OF TEST ITEM SPECIFICATIONS

INSTRUCTIONS

Check all boxes that apply, and keep a copy of this form for your records. If you have questions, please contact the USF General Counsel or your USF Tampa Library Copyright Librarian.

Name: Heather Peltier Wright  Date: 04/30/2017

Class or Project: Exploring Teacher Assessment Literacy Through the Process of Training Teachers to Write Assessment Items

Title of copyrighted work: Test Item Specifications, Florida Department of Education

PURPOSE AND CHARACTER OF THE USE

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educational</td>
<td>Commercial</td>
</tr>
<tr>
<td>Teaching (including multiple copies for classroom use)</td>
<td>Entertainment</td>
</tr>
<tr>
<td>Research or Scholarship</td>
<td>Bad-faith behavior</td>
</tr>
<tr>
<td>Criticism, Parody, News Reporting or Comment</td>
<td>Denying credit to original author</td>
</tr>
<tr>
<td>Transformative Use (your new work relies on and adds new expression, meaning, or message to the original work)</td>
<td>Non-transformative or exact copy</td>
</tr>
<tr>
<td>Restricted Access (to students or other appropriate group)</td>
<td>Made accessible on Web or to public</td>
</tr>
<tr>
<td>Nonprofit</td>
<td>Profit-generating use</td>
</tr>
</tbody>
</table>

Overall, the purpose and character of your use: □ supports fair use or □ does not support fair use.

NATURE OF THE COPYRIGHTED MATERIAL

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factual or nonfiction</td>
<td>Creative or fiction</td>
</tr>
<tr>
<td>Important to favored educational objectives</td>
<td>Consumable (workbooks, tests)</td>
</tr>
<tr>
<td>Published work</td>
<td>Unpublished</td>
</tr>
</tbody>
</table>

Overall, the nature of the copyrighted material: □ supports fair use or □ does not support fair use.

AMOUNT AND SUBSTANTIVITY OF MATERIAL USED IN RELATION TO WHOLE

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small amount (using only the amount necessary to accomplish the purpose)</td>
<td>Large portion or whole work</td>
</tr>
<tr>
<td>Amount is important to favored socially beneficial objective (i.e. educational objectives)</td>
<td>Portion used is qualitatively substantial (i.e. it is the ‘heart of the work’)</td>
</tr>
<tr>
<td>Lower quality from original (ex. Lower resolution or bitrate photos, video, and audio)</td>
<td>Similar or exact quality of original work</td>
</tr>
</tbody>
</table>

LeEtta Schmidt, lmschmidt@usf.edu and Drew Smith dsmith@usf.edu
Reviewed by USF General Counsel 08/11/2015
Overall, the amount and substantiality of material used in relation to the whole □ supports fair use or □ does not support fair use.

**EFFECT ON THE MARKET FOR ORIGINAL**

<table>
<thead>
<tr>
<th>Likely Supports Fair Use</th>
<th>Likely Does Not Support Fair Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ No significant effect on the market or potential market for the original</td>
<td>□ Replaces sale of copyrighted work</td>
</tr>
<tr>
<td>□ No similar product marketed by the copyright holder</td>
<td>□ Significantly impairs market or potential market for the work</td>
</tr>
<tr>
<td>□ You own a lawfully acquired copy of the material</td>
<td>□ Numerous copies or repeated, long-term use</td>
</tr>
<tr>
<td>□ The copyright holder is unidentifiable</td>
<td>□ Made accessible on Web or to public</td>
</tr>
<tr>
<td>□ Lack of licensing mechanism for the material</td>
<td>□ Affordable and reasonably available permissions or licensing</td>
</tr>
</tbody>
</table>

Overall, the effect on the market for the original □ supports fair use or □ does not support fair use.

**CONCLUSION**

The combined purpose and character of the use, nature of the copyrighted material, amount and substantiality of material used in relation to the whole and the effect on the market for the original □ likely supports fair use or □ likely does not support fair use.

*Note: Should your use of copyrighted material not support fair use, you may still be able to locate and request permissions from the copyright holder. For help on this, please feel free to contact your Copyright Librarian.*

This worksheet has been adapted from:

Cornell University’s Checklist for Conducting A Fair use Analysis Before Using Copyrighted Materials:  
[https://copyright.cornell.edu/policies/docs/Fair_Use_Checklist.pdf](https://copyright.cornell.edu/policies/docs/Fair_Use_Checklist.pdf)


Smith, Kevin; Macklin, Lisa A.; Gilliland, Anne. A Framework for Analyzing any Copyright Problem. Retrieved from:  
[https://d396quszo4orc.cloudfront.net/cfel/Reading%20Docs/A%20Framework%20for%20Analyzing%20any%20Copyright%20Problem.pdf](https://d396quszo4orc.cloudfront.net/cfel/Reading%20Docs/A%20Framework%20for%20Analyzing%20any%20Copyright%20Problem.pdf)

LeEtta Schmidt, lmschmidt@usf.edu and Drew Smith dsmith@usf.edu  
Reviewed by USF General Counsel 08/11/2015
APPENDIX I:

INSTITUTIONAL REVIEW BOARD LETTER

December 15, 2016

Heather Wright
Educational and Psychological Studies
Tampa, FL 33612

RE: Not Human Subjects Research Determination
IRB#: Pro00026968
Title: Exploring Teacher Assessment Literacy Through the Process of Training Teachers to Write Assessment Items

Dear Ms. Wright:

The Institutional Review Board (IRB) has reviewed your application and determined the activities do not meet the definition of human subjects research. Therefore, this project is not under the purview of the USF IRB and approval is not required. If the scope of your project changes in the future, please contact the IRB for further guidance.

All research activities, regardless of the level of IRB oversight, must be conducted in a manner that is consistent with the ethical principles of your profession. Please note that there may be requirements under the HIPAA Privacy Rule that apply to the information/data you will utilize. For further information, please contact a HIPAA Program administrator at 813-974-5638.

We appreciate your dedication to the ethical conduct of research at the University of South Florida. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

John Schinka, Ph.D., Chairperson
USF Institutional Review Board