American Journal of Educational Research, 2015, Vol. 3, No. 12, 1528-1535
Available online at http://pubs.sciepub.com/education/3/12/8

DOI:10.12691/education-3-12-8

Project-based Approach to Teaching Standardized Test Construction: A Model for Pre-Service Education Students on How to Standardize a Test

David M. Gordon*, James C. Collins, Kelly L. Jewell

Special Education, University of Wisconsin – Whitewater, Whitewater, USA
*Corresponding author: gordond@uww.edu

Abstract The article is a descriptive case study of a project-based approach to simulate test construction in a pre-service undergraduate training program for future teachers. The curriculum for pre-service teachers (PSTs) in the training program involved learning how to develop tests, how to practice using standardized assessment instruments, and to how to interpret the results from testing. The PSTs completed an educational project to acquire hands on teaching skills in relation to test construction and standardizing a test of a sample population. This article discusses the use of a project-based instructional approach by having students develop a standardized test, rather than using a traditional lecture-based approach, to teach basic assessment skills. Results revealed that PSTs learned the necessary foundational knowledge required in the curriculum and enjoyed the hands-on approach to learning. Practical implications and areas in need of future research are discussed.

Keywords: *test construction, project-based learning, pre-service teachers*

Cite This Article: David M. Gordon, James C. Collins, and Kelly L. Jewell, "Project-based Approach to Teaching Standardized Test Construction: A Model for Pre-Service Education Students on How to Standardize a Test." *American Journal of Educational Research*, vol. 3, no. 12 (2015): 1528-1535. doi: 10.12691/education-3-12-8.

1. Introduction

When a pre-service teacher within the field of special education signs up to take an introduction to assessment class, many may not come in thrilled at the idea to learn about techniques to test for IEP development or implement an educational evaluation. This lack of motivation to learn could contribute to preserves teachers lack of knowledge about writing and interrupting the tests learn about in the class [7,12]. Asim, Ekuri, and Eni [1] found pre-service teachers were specifically not competent in how to construct multiple choice test items especially what they referred to as best-answer type (BAT). This is key when trying to develop preservice teachers on informal and formal assessment. Teacher preparation programs typically have a class in assessment that prepares teachers in a variety of areas, such as the nature of educational measurement, technical adequacy of tests, writing objective tests and test items, basic information related to standardized tests, statistical treatment of test scores, and how to report results [3]. To combat this lack of interest, universities and colleges are pushed to think of differentiated methods to incorporate into classes to include the diverse learners within the college environment. One such method is the use of project-based learning, which provides hands on learning for pre-services teachers so that they can "feel" their way through the experience while learning the essential skill within the area of special education assessment. This hands-on learning allows the preservice teachers to have a vested interested in their learning instead of just completing an assignment.

2. Literature Review

The use of project-based learning was best described by Blumenfeld et al. as activities driven by questions or problems with a culminating end product that cycled back to the beginning questions or problems. The design of project-based learning is for students to learn using a hands-on approach, rather than inundating students with lecture. The existence of this learning style dates back to 1975 when Adderley et al. wrote an article describing how project-based learning would benefit students in higher education. Authors described project-based learning as involving (a) the solution and the problem, (b) a student or groups of students with educational expectations, (c) an end product (i.e., thesis, report, design plan, or paper, (d) projects that continue for a length of time, and (e) educators who function as coaches or advisors rather than the lecturer. Within the project based learning assessment class, the instructor made sure to use these as guidelines throughout the entire class, while acting as a coach for the student. Another area researchers pulled information from was simulation. Project-based learning shares common features with a simulation. As Dewey described in bring the two designs together it both involve a learning activities through problems with an end solution.

Project-based approach is consistent with the Dewey philosophy of constructing learning. Building a better process for training of pre-service special education majors and minors in the assessment process is very important. The project-based approach uses simulations and real life examples to replicate the real-world experience the PSTs will get when they are practicing teachers. Muslevy described the process of a simulation as,

> "A careful design can support a person's learning by tailoring the features of situations to his or her skill level(s), allowing repeated attempts, and by providing timely feedback"(p. 1).

Within the project based learning, preservice teachers interacted with standardized test within multiple environments. Standardized tests are an important part of PSTs needed knowledge base because it's often their responsibility to administer national and state-developed high-stakes tests. Evaluation teams use results from standardized tests as a measuring tool in determination of eligibility for special education services [8]. The authors define standardized tests as;

> They are called *standardized* because they are administered and scored according to *specific* and *uniform* (i.e., standard) procedures….Because of these standardization of procedures, measurement error due to administration and scoring is reduced which increases score reliability. Standardized tests are constructed by test construction specialists, usually with the assistance of curriculum experts, teachers and school administrators...When standardized tests are used to compare performance to students across the country, they are called standardized norm-referenced tests, and when they are used to determine whether performance meets or exceeds criteria like state standards, they are called standardized criterion-referenced tests (p. 370).

2.1. Assessment Course Curriculum

This descriptive case study took place at college of education in the special education department of a program in the upper Midwest. PST's are required to take two assessment classes for the majors and one for the minors in special education. In the assessment classes for pre-service educators at the university, the curriculum is aligned with the textbook adopted by the department. The students are taught methods of administering, scoring and interpreting standardized test. In Kubiszyn and Borich [8], the process of standardized testing is the final third of the test book the department adopted for this course. The curriculum also includes the process of writing [14], administering standardized tests, as well as finding the measures of central tendency, variability, normal distribution and converting score. Additionally, the material identifies methods of correlation, validity and reliability, accuracy and error. Finally, the text and curriculum in the assessment classes includes interpreting the results for making educational recommendations [15].

This paper describes a project-based instructional approach that was presented to multiple undergraduate assessment classes for special education majors and minors. Students were taught basic assessment knowledge, followed by active participation in the process of test construction, standardization, and use of data to evaluate psychometric properties of the assessment. Other sections of the same class were taught the curriculum of developing and interpreting a standardized test in a more traditional method of practice with descriptive and inferential statistics, formulas for calculating normal distribution and correlations along with providing evidence for validity and reliability. This simulation allowed the PSTs to be introduced to these concepts and use data they collected to draw conclusions about the uses and value of these tests. Future research will evaluate the effectiveness of these two approaches.

3. Method

3.1. Participants

Participants in the project included 62 students who were enrolled in a semester-long introductory assessment course. Students were either in a project based learning class (25) or a lecture-based class (37). The lecture-based class was considered business as usual and the instructor, second author, taught the class using the same material the previous semester. The project spanned one semester. The primary author facilitated this simulation/project-based learning within the context of a course offered in a licensure track program for special education majors and minors (see Table 1).

Table 1. Simulation Participants

Number of Participants	N = 75	
Gender	Male = 12%	Female = 88%
Age	Mean = 21.27	
Special Education Majors	N = 28	37.33%
Special Education Minors	N = 47	63.67%

All of the participants were studying in a special education licensure program to teach in the state of Wisconsin. The special education majors represented (37.33%), while minors in special education, with a major in curriculum and instruction, consisted of (62.67%) of the participants. For majors who successfully complete the program, they will be eligible to hold a teaching license in the area of "cross categorical" with an emphasis in learning disabilities and emotional-behavioral disorders or intellectual disabilities. The minors who successfully complete the program will be eligible to hold a teaching license in their area of study in curriculum and instruction. Most of the PSTs were female (88%) and most were typical age college sophomore or junior standing with the mean age 21.26 years old. All of the PSTs were ethnically, Caucasian and most have lived most of their lives in Wisconsin.

3.2. Materials and Procedures

In the Spring 2014, the Introduction to Assessment and Diagnosis class participated in a simulation to construct a test and then to standardize the test. At the beginning of the semester, the students were told they were going to construct a test of achievement. The subject of the test was on the facts and history of the state of Wisconsin. A group earlier named this test *Knowledge of Wisconsin* (KOW).

3.2.1. Evaluation of Learning

To evaluate the project-based learning versus the lecture based learning method the instructors used a Likert Scale questionnaire, which evaluated students' level of knowledge in 8 areas (Objective Test, Item Analysis, Test Data, Central Trends, Variability, Correlation, Validity, and Reliability). At the end of the class students could rate their level of learning in each area from 1 to 6.

3.3. Description of Project

3.3.1. Test Construction

Students, within the project based learning class, gathered information on the state from a variety of sources, including the Internet and middle and high school textbooks. In a whole class discussion, three primary domains of questions were developed: *Basic Facts*, *Famous Wisconsinites*, and *History*. Each of the students had to write at least three questions in each domain for the pool. The PSTs met in small groups to discuss and enhance the language of the different questions they brought in. One issue raised in the debate at this point was "regionalism." Some of the questions were so narrow in focus only a person in a particular region might know the answer but less likely for even a long-term resident of the state. The PSTs helped design how the paper version of the test and the directions for administering the test would be given.

3.3.2. Item Pool

After introducing the students to writing objective test items, specifically writing multiple-choice items, students were able to identify the components of multiple-choice items (e.g., stem, correct answer, and distractors). Next, students began developing the item pool. Students brought the items into the university class and, in small groups, began reviewing their work. Questions were initially vetted in small groups to edit language and to make certain that there was one correct answer and that the distractors were plausible, but clearly incorrect. Reference materials and Internet searches on different library and historical societies were completed to ensure the accuracy of the test questions. Fifty questions were selected by the class based on questions ranging in difficulty to topics covered the different domains. Questions were reviewed to avoid redundancy and to have a mix of questions in different domains.

3.3.3. Item Analysis

Using the pool of fifty questions, each question was discussed for inclusion in the final version of the KOW and the level of perceived difficulty. The test pool was then put on an electronic survey instrument that asked the PSTs to review the questions and anonymously address if the question was worthy of being in the final version of the KOW, and to rank the questions on perceived levels of difficulty. Each of the PSTs ranked the questions on a Likert scale from easy to difficult. Questions that made the final version of the KOW had to receive an 80% or higher approval from the class. A version of the KOW was put together with questions from this pool that had varying levels of difficulty based on the PSTs ratings.

The ranking order of the PSTs of the different items led to a good discussion of construct validity, which is defined as the degree to which a test accurately measures a hypothetical construct [8]. The intent of the test construction was to have the questions get progressively harder the farther you went in the test. This introduced concepts like "suggested starting points," "basals," and "ceiling" into the dialogue. As Figure 3 illustrates, the results of the KOW demonstrated to the PSTs that their rating and order of questions did get progressively more challenging and the version of the test had a fair mix of easy to difficult questions.

3.3.4. Administering the KOW to a Sample Population

One critical part of the standardization process was to take the test out to practice administering it and collecting data from a representative sample population. A set of standard directions and a sample items were added for the PSTs to introduce when administering it to participants. This was an important part of the curriculum to demonstrate, model, and practice the standardized administration of a test. Students practiced in class with their peers and they were directed to administer a thirty-question version of the KOW to at least five individuals who live in the state of Wisconsin. Along with administering this to five people, the PSTs had to collect demographic information to be used to see if there was correlation between factors such as age, years living in the state, etc. They were given a two-week window to administer the test, score, and collect the demographic information. All but one of the PST's was able to meet this deadline. The PSTs entered their subjects' information in an electronic survey instrument for data analysis purposes.

Table 2. Demographics of Simulation Participants

Sample Size	N=389		
Gender	Male = 40.1%	Female = 55.8%	
Age	Mean = 28.05	Range min = 7.0, max = 80.2	
Years Living in Wisconsin	24.87	Range min = 1.0, max = 80.2	
Highest Grade in School	12.49	Range min = 1.0, max = 22.0	
Ethnicity		Frequency	Percentage
	Caucasian	358	92.0%
	African American/Black	5	1.3%
	Hispanic/Latino	8	2.1%
	Middle Eastern	1	.3%
	Native American/Alaskan	1	.3%
Job		Frequency	Percentage
	No Job	55	14.1%
	One Job	127	32.6%
	Two or More	156	40.1%
Language Spoken at Home		Frequency	Percentage
	English	367	94.3%
	Spanish	1	.2%
	Multilingual	5	1.3%

3.3.5. Sample Population

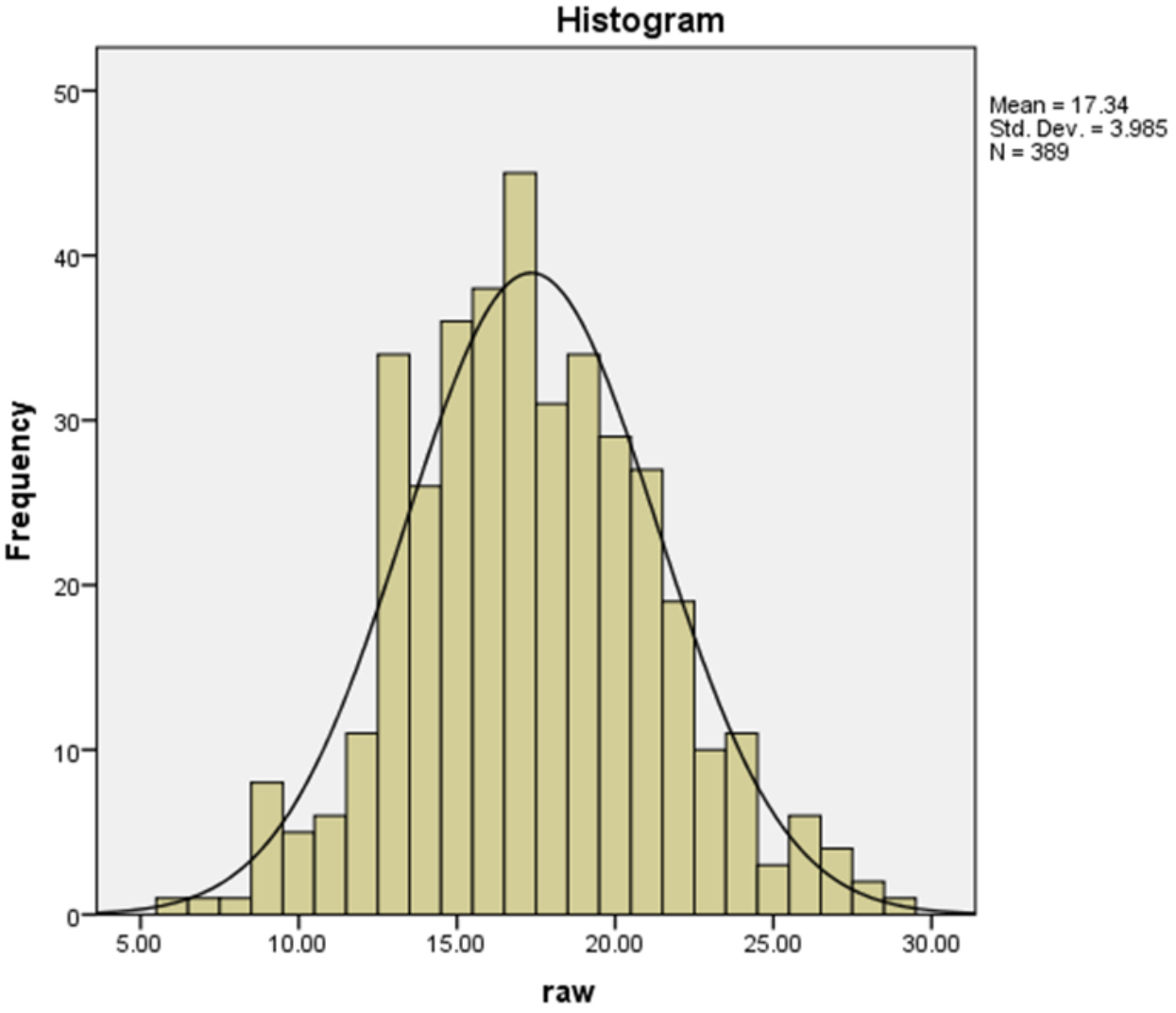
The PSTs were instructed to administer the KOW to at least five individuals. This created a database that students used to learn about and apply concepts related to assessment. The nature of where they found subjects to test was left up to them and is reflected in the demographics of the pool of test subjects (see Table 2). However, most of the PSTs reported testing people within their circle including friends, families and neighbors. The PST's collected 389 tests results.

3.3.6. Standardization of the KOW

After raw scores from the KOW were obtained and compiled, students were shown how to convert scores to standard scores and why this process is necessary to effectively interpret the results. The KOW gave the PSTs the opportunity to write, administer and score a test. The standardization process was done to frame the interpretation of the results.

3.3.7. Teaching Measures of Central Tendency

The database allowed for a lot of options to teach the concept of measures of central tendency. All of the PST's in spring 2014 reported that they had been taught these concepts in previous math courses at the university, but reported that the applied approach used in this project was helpful and provided more depth of understanding.



N	389
Mean	17.3445
Median	17
Mode	17
Range	23
Variance	15.881
Std. Deviation	3.98510
Std. Error of Mean	.20205

Figure 1. Measures of Central Tendency and Initial Statistics – Based on Raw Scores on Knowledge of Wisconsin (KOW) Test – 2012-2014

The process of determining measures of central tendency was introduced and the results of the KOW were used to demonstrate the values of mean, median and mode (see Figure 1). Conveniently and likely due to the large sample size, the raw scores of the KOW produced a normal distribution of scores with a mean, median and mode that were nearly identical. The PST's each calculated the arithmetic mean on the sample tests they gave to practice calculating the mean.

3.3.8. Analyzing the Results of the KOW

While more advanced statistical analysis for the KOW were presented, the primary emphasis was to use the data to demonstrate how to convert raw scores into standard scores, determine measures of variance, percentile ranks and standard deviation scores. This introduction to inferential statistics was to supplement the curriculum and e-textbook used in the class.

3.3.9. Norming the Results

Raw scores were converted to standard scores, which produced a relatively normal distribution. Norming of the test was done in SPSS by calculating the standard deviation and then converting raw scores to z-scores. The bulk of the scores (77%) were between -1 *z* score and +1 *z* score and the remaining scores fell outside this range. The PSTs were able to use the mean and the standard deviation to observe how their subjects scored on the KOW.

3.3.10. Interpreting the Results

To interpret results on the KOW, the PSTs were introduced to correlation coefficients and how these can be used to understand the results. Students then constructed a scatterplot of data (see Figure 2), which was used to assist in the discussion and understanding of positive and negative correlations. The PSTs were able to discuss how the results of the KOW can be interpreted and to make predictions based on factors in this case the years of living in Wisconsin.

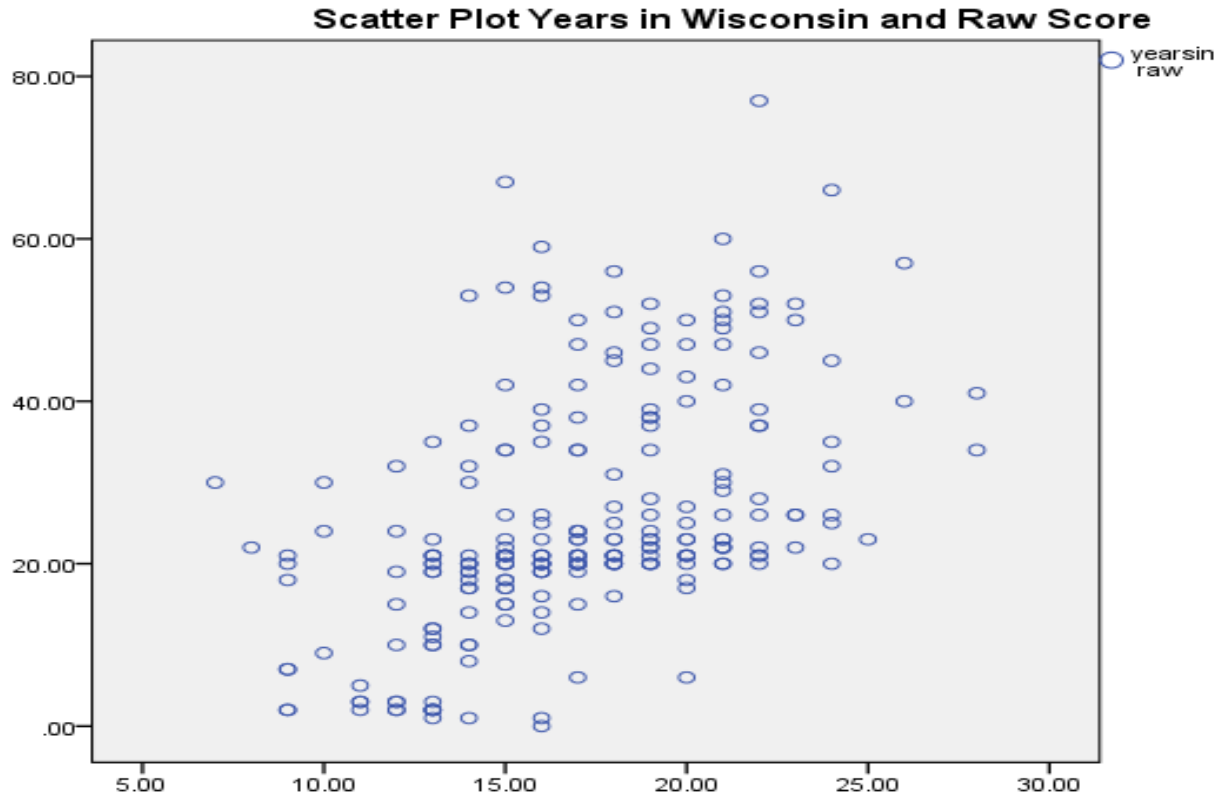


Figure 2. Graphic of Scatterplot of Raw Scores and Years Living in Wisconsin on Knowledge of Wisconsin (KOW) Test – 2014

3.3.11. Technical Adequacy of the KOW (Validity and Reliability)

An important part of the assessment curriculum for PSTs is to understand the concepts of test validity and test reliability. The purpose of this part of the curriculum is to present methods for the PSTs to select quality assessment instruments when testing their students. The Kubiszyn and Borich [8] text presents validity (defined briefly as, "does the test measure what it is supposed to measure?" page 329) and reliability (defined briefly as, "does the test yield the same or similar score ranking (all other factors being equal) consistently?" page 329) as the standards for determining the adequacy of a test instrument. There are three primary forms of measuring test validity.

In the case of the KOW the PSTs developed the item pools to create the test instrument with they address content and construct validity. In content validity the test items are examined based on some defined content. For content validity the PST needed to the appropriateness of the types of items included, the completeness of the item sample, and, the way in which the items assess the content. In Figure 3 the question statistics are presented that demonstrated on the progressive nature of question difficulty. In this case the content was based on information on the state of Wisconsin. This was discussed in the initial construction of the test as the PSTs had to acquire information from different domains about the state and then to rank them in order from easy to difficult. As seen in the graphic, the results indicated that the questions

did get progressively more difficult. The outliers were questions six and eleven. Later versions of the KOW can reflect this information based on the data collected in this sample.

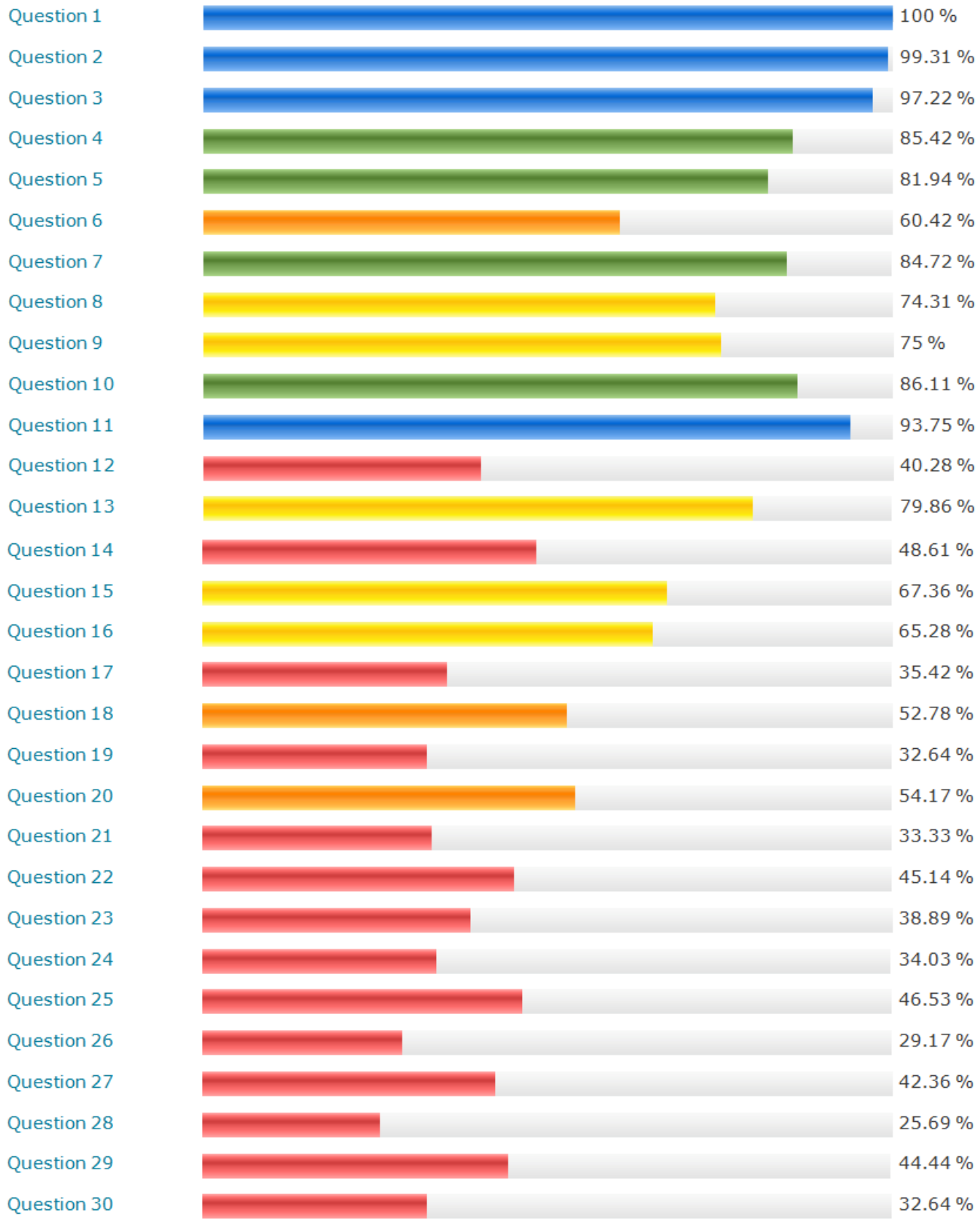
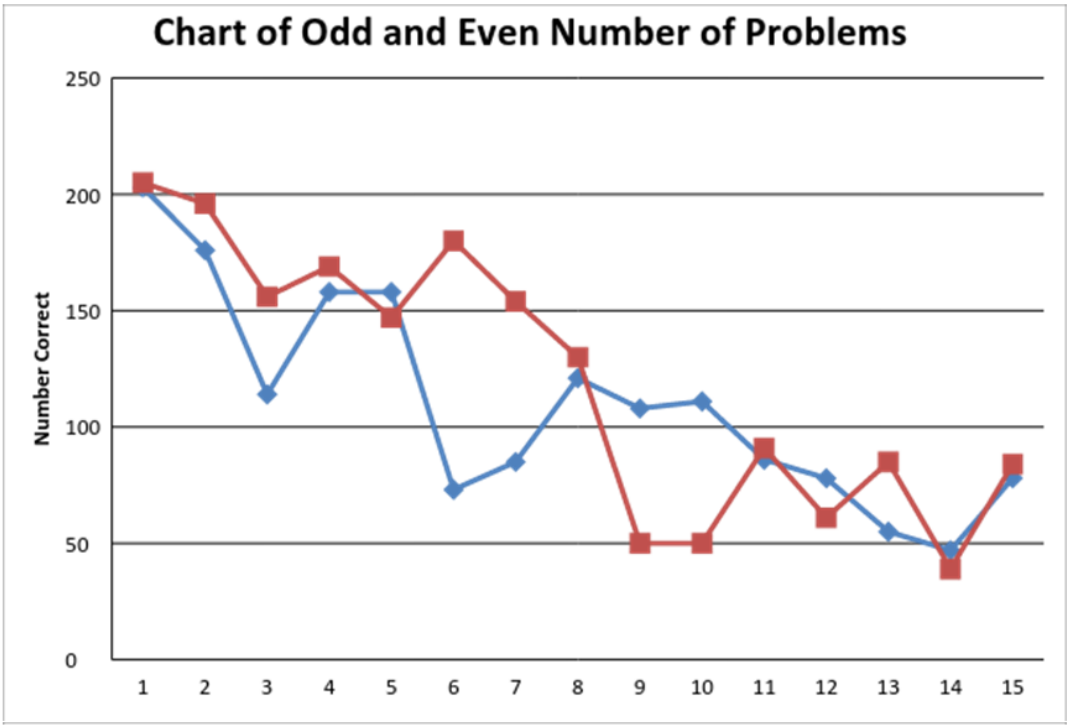


Figure 3. Graphic of Question Statistics Spring 2014

The PSTs were introduced to four primary methods for determining reliability. The KOW was used to demonstrate "split half reliability" by comparing the results on the even versus the odd problems. Figure 4 shows the split half reliability for the KOW over a couple years of administration. The PSTs were able to graph the scores for the visual of the split half to show internal consistency. Using SPSS, the a negative correlation of -0.67 demonstrated a "strong relationship."



Even	Odd	Correlation between scores on Even and Odd
Average Degree of Difficulty – 3	Average Degree of Difficulty – 3	r = -0.67
% Correct – 53.87%	% Correct – 58.16%	Indicates a "strong relationship" between scores

Figure 4. Graphic Split Half Reliability of the KOW 2011-2014

4. Results

While the KOW correlation demonstrated correlation between the number of years living in Wisconsin and performance on the test, between the two classes there no significant differences when students self rated their learning in any questions covered in the classes. However, for question 1 on Objective Test $t(60) = -1.47$. Question 2 on Item Analysis $t(60) = 0.75$. Question 3 on Test Data $t(60) = 1.02$. Question 4 on Central Trends $t(60) = 0.73$. Question 5 on Variability $t(60) = -1.55$. Question 6 on Correlation $t(60) = 1.36$. Question 7 on Validity $t(60) = -0.58$. Question 8 on Reliability $t(60) = 0.49$. This demonstrates the value of a hands on approach within a teacher education program.

5. Discussion

Although there was no specific quantitative data from on semester demonstrating advance learning in the project based class versus the lecture-based class. The results from the study demonstrated many benefits to using simulations within an assessment classroom for special education majors and minors. In the comment section of the Likert Scale a student wrote

> "I really enjoyed creating this test and have learned a few things along the way to be able to create different tests in the future. I put a lot of thought and effort into this assignment and so it made creating it and giving the test more fun. I fully participated fully in the creation of the test because how often do you get to create and administer a test to whomever as a college student. The one thing that I really learned from creating this test was the importance of creating questions that were measuring what you wanted to be measured. It is easy to come up with test questions, but if they don't match what you are testing then they are pointless and sometimes frustrating to include in a test. After creating this test, I feel confident creating tests and administering them to my future students. I will create tests that measure exactly what I want my students to learn and nothing more. I will make questions that provide only one correct answer and questions that are objective, limiting bias. I will include questions that are "easy" and "hard" so students aren't bored or frustrated with the test."

Using the project-based method allowed pre-service teachers to work independently, take on responsibility for their learning, and construct knowledge using an engaging hands-on approach. Growing evidence suggests that is demonstration outside of education, that when give activities related to job experiences with feedback then they are more likely to use this feedback and apply to future behaviors. Impara, Plake abd Fager [5] stress, "Teacher education programs must take the initiative to provide their students with skills and knowledge in assessing student performance" (p.117).

5.1. Limitations

With a study where the research and the instructors are the same people, there are going to be several limitations.

The limitations in this study involve the pre-service teachers' level of performance, sample size for norming, and centralism of sample size.

5.1.1. Performance

Using a project-based approach as part of the class is a strength, yet also a limitation because PSTs received a grade for engaging in the work. This encouraged many pre-service teachers to participate at an optimal level because there was a connection to the grade in the class. Although many pre-service teachers ended up finding more enjoyment and satisfaction beyond the grad that they received, it must be noted as a potential limitation.

5.1.2. Norming Limitations

The challenge of using the KOW simulation was norming the test and being able to create a database sufficient enough to introduce and teach how standardized tests use inferential statistics. The first group of students who administered the KOW to a sample population only had a sample size of 42, which created a challenge to determine if the test could be normed. With subsequent groups adding to the sample, the test shows promise as an instrument to determine scale and standard scores and to calculate percentile rank. As an initial instrument for PSTs, it provides a demonstration as to how these scores are factored and how to use the scores to make determination of students' progress, eligibility, and aptitude.

5.1.3. Sample Size

The university is located in the southern portion of Wisconsin and our students were on campus while completing this study. They did have access to their families and friends to answer the questions, yet we cannot ensure we had an ample representation of the entire state. The issues, such as regionalism of the questions on the KOW and diversity of the sample population, adversely affected the standardization process.

6. Conclusions and Areas of Future Research

Using project-based learning is beneficial for pre-service teachers to understand the small (inside classroom) evidence of student learning and big picture (norming of a test) when it is used within a college assessment course. As higher education is trying to find new ways to reform and redevelop, taking a hands-on approach to teaching this assessment class was a step in the right direction. Future research needs to evaluate if this can be replicated within another university with similar results. Another project would be to evaluate the efficacy of project-based learning for informal learning assessments inside the classroom.

References

[1] Asim, A.E., Ekuri, E.E., and Eni, E.I. (2013). A diagnostic study of pre-service teachers' competency in multiple-choice item development. *Research in Education, 89,* 13-22.

[2] Brickell, H.M. (1976). Needed: Instruments as good as our eyes. *Journal of Career Education 2*(3), 56-66.

[3] Blood, D.F., and Budd, W.C. (1972). *Educational measurement and evaluation*. New York: Harper & Row.

[4] Brownell, M.T., Ross, D.R., Colón, E.P., and McCallum, C.L. (2003). *Critical features of special education teacher preparation: A comparison with exemplary practices in general teacher education.* (COPSSE Document Number RS-4). Gainesville, FL: University of Florida, Center on Personnel Studies in Special Education.

[5] Impara, J.C., Plake, B.S., and Fager, J.J. (1993). Teachers' assessment background and attitudes toward testing. *Theory into Practice 32*(2), 113-117.

[6] Kasparek, J., Malone, B., and Schock, E. (2004). *Wisconsin history highlights: Delving into the past.* Madison, Wisconsin: Wisconsin Historical Society Press.

[7] Koetsier, C.P., & Wubbels, J.T. (1995). Bridging the gap between initial teacher training and teacher induction. *Journal of DEducation for Teaching, 21*(3), 333-345.

[8] Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement: Classroom application and practice* (10th ed.). Hoboken, NJ: John Wiley & Sons, Inc.

[9] Mayo, S.T. (1967). *Pre-service preparation of teachers in educational measurement. Final report* (Project No. 5- 0807). Washington, D.C Office of Education (DHEW), Bureau of Research.

[10] Mislevy, R.J. (2011). *Evidence-centered design for simulation-based assessment CRESST Report 800*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

[11] Murphy, K.R., and Davidshofer, C.O. (1991). *Psychological Testing Principles and Applications* (4th ed.), Upper Saddle River, New Jersey: Prentice Hall.

[12] Nenty, H. J., Adedoyin, O. O., Odili, J. and Major, T. E. (2007). Primary teachers' perception of classroom assessment practices as means of providing quality primary education by Botswana and Nigerian teachers.*Educational Research and Reviews, 2* (4), 74-81.

[13] Parker, D.C., McMaster, K.L. and Burns, M.K. (2011). Determining an instructional level for early writing skills. *School Psychology Review 40* (1) 158-167.

[14] Yu, H. (2008). Conceptualized technical writing assessment to better prepare students for workplace writing: Student-centered assessment instruments. *Journal of Technical Writing and Communication 88*(3), 265-284.

[15] Zenisky, A. L., and Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement Issues and Practice 31*(2), 21-26.