HIGH SCHOOL (AGES 14-18) STUDENTS' UNDERSTANDING OF STATISTICS

Tim Jacobbe, Steve Foti, Catherine Case, and Douglas Whitaker
University of Florida, United States
jacobbe@coe.ufl.edu

This paper will present results from the administration of the LOCUS assessments to measure students' statistical understanding in grades 9-12 (ages 14 – 18). The development of these assessments utilized an Evidence Centered Design (ECD) (Mislevy & Riconscente, 2006) approach to establish their content validity. After an iterative development process, these assessments were administered to over 2,000 students in the United States. Student performance in each of the four areas of the statistical problem solving process - formulating questions, collecting data, analyzing data, and interpreting results - will be discussed, and examples of multiple-choice and constructed-response items will be provided.

INTRODUCTION TO LOCUS

The release and widespread adoption of the Common Core State Standards (CCSS) have dramatically increased the expectations for teaching statistics in grades 6 through 12 in the United States. The development of many of the standards for teaching statistics was based on the American Statistical Association's *Guidelines for Assessment and Instruction of Statistics Education* (GAISE) (Franklin et al., 2007). With the increased expectations for teaching statistics comes the demand for tools to properly assess the conceptual understanding of learners of statistics. Most large-scale assessments, however, still emphasize procedure (Friel et al., 1997; Konold, 1995). The goals of the LOCUS project focus on the development and implementation of instruments to measure current levels of *conceptual* understanding in relation to expectations set forth in the CCSS. These assessments also serve as an example for testing industries as they work toward models of national assessment for the CCSS initiative. While the LOCUS assessment relates to the expectations of the CCSS, it is not limited to circumstances involving these standards. LOCUS remains a useful tool in assessing learners' levels of statistical development as outlined in the GAISE framework outside of the CCSS and the United States.

TEST DEVELOPMENT PROCESS

The GAISE framework identifies three levels of statistical development (Levels A, B, and C) that students progress through in order to develop statistical understanding. Grade ranges for these levels are intentionally unspecified; however ideally Levels A, B, and C would correspond with elementary (Grades K-5/Ages 5-11), middle (Grades 6-8/Ages 12-14), and high school (Grades 9-12/Ages 15-18), respectively. "Without such experiences, a middle [or high] school student who has had no prior experience with statistics will need to begin with Level A concepts and activities before moving to Level B" (Franklin, et al., 2007, p.13). One of the assessments developed in this project addresses content from both Level A and Level B, while the other exam addresses content from Level B and Level C. Levels A and B were combined into one assessment to remain consistent with the CCSS. Level B and C were combined to provide the ability to place students on a continuum. For example, a student who does not do exceptional on the Level C items could be considered a Level B student, with some confidence.

The assessments were developed using an evidence-centered assessment design model (ECD) (Mislevy et al., 2003). There are five layers involved in the ECD process: domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery. Using the GAISE framework and the CCSS, the advisory board and the test development committee (TDC) were able to establish the conceptual assessment framework. Once the framework was established, the TDC began to write prototype items, which consisted of both multiple-choice (MC) and constructed-response (CR) questions. As part of the development process, scoring rubrics for grading the constructed-response questions were also drafted and reviewed. The scoring rubrics are intended to not only evaluate responses on their statistical accuracy, but also in terms of evidence related to characteristics of the level (A, B, or C) of the response. Throughout the course of three TDC meetings, the items were reviewed, discussed, and

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

modified. After another meeting for review and alignment to the evidence model, the item pool was used to assemble 8 pilot forms. Once created, these pilot forms were reviewed and revised by the TDC, advisory board, and the joint NCTM/ASA committee. These final pilot forms were printed and administered in spring 2013. Jacobbe et al. (2014) provides further details regarding the ECD process used in creating the LOCUS assessment.

PILOT ADMINISTRATION

The pilot forms were administered to students at schools in 6 different states, all of which have adopted the CCSS and had a representative on at least one of the committees associated with the project to assist in the delivery of the assessments. Table 1 shows the demographic information for the 1,249 students who participated in the pilot.

Table 1. Demographic information for students in pilot

Gender		Grade		English*	
Female	46.28 %	9	12.81 %	No	13.29 %
Male	51.40 %	10	11.29 %	Yes	84.23 %
Race		11	27.14 %	Ethnicity	
Not Hispanic	87.11 %	12	48.76 %	AmIndian/PacIslander	1.12 %
Mexican	3.92 %			African American	4.72 %
Puerto Rican	2.32 %			Asian	6.08 %
Cuban	1.44 %			White	75.82 %
Other Hispanic	2.00 %			Other	5.92 %

**first language*

Once all of the pilot assessments were returned, the constructed-response items were graded using the scoring rubrics during a week-long scoring session involving the investigators on the project, the TDC, and some graduate students interested in statistics education.

RESULTS

Overall

Out of 41 total possible points (multiple-choice and constructed-response combined), the minimum score across all of the forms was 0, and the maximum score across all of the forms was 40. The means and standard deviations for total scores for all forms are shown in Table 2. The mean scores in the pilot show that the forms were all similar in difficulty with the exception of form 2, which appeared to be slightly more difficult. The internal reliabilities (stratified alpha) for the forms were between 0.70 and 0.77.

Table 2. Means and standard deviations for the 4 test forms.

	Form			
	1	2	3	4
Mean	16.00	19.45	16.97	18.21
Std. Deviation	7.46	8.03	7.65	7.56

Multiple-Choice

For the multiple-choice part of the exam, the mean item difficulty across the four forms was 0.52. The mean biserial correlation across the four forms was 0.46. Table 3 shows the item difficulties across all forms, organized by the area of the statistical problem solving process.

Table 3. Item Difficulties by Statistical Problem Solving Process - MC

Process	N	Mean
Formulating Questions	8	0.68
Collecting Data	22	0.60
Analyzing Data	28	0.51
Interpreting Data	26	0.40

The students performed best on items that were in the formulating questions area. These items involve the planning process used in a statistical problem, including having the students determine if the problem is statistical in nature, decide what data is needed to answer the question, or consider what inferences can be drawn from the data. An example of a Level C item for this area of the statistical problem solving process is shown below. In the pilot, 19% of the students were able to correctly answer the item.

Example 1: A 13-year study of 1328 adults randomly selected from a population carefully monitored the personal habits and health conditions of participants. Personal habits included tobacco use and coffee consumption. Health conditions included incidence of stroke. Which of the following questions about this population CANNOT be answered using data from this study?

(A) Are coffee drinkers more likely to smoke than adults who do not drink coffee?
(B) Does coffee consumption cause a reduction in the incidence of stroke?
(C) Do coffee drinkers have fewer strokes than adults who do not drink coffee?
(D) What percentage of the population are coffee drinkers?

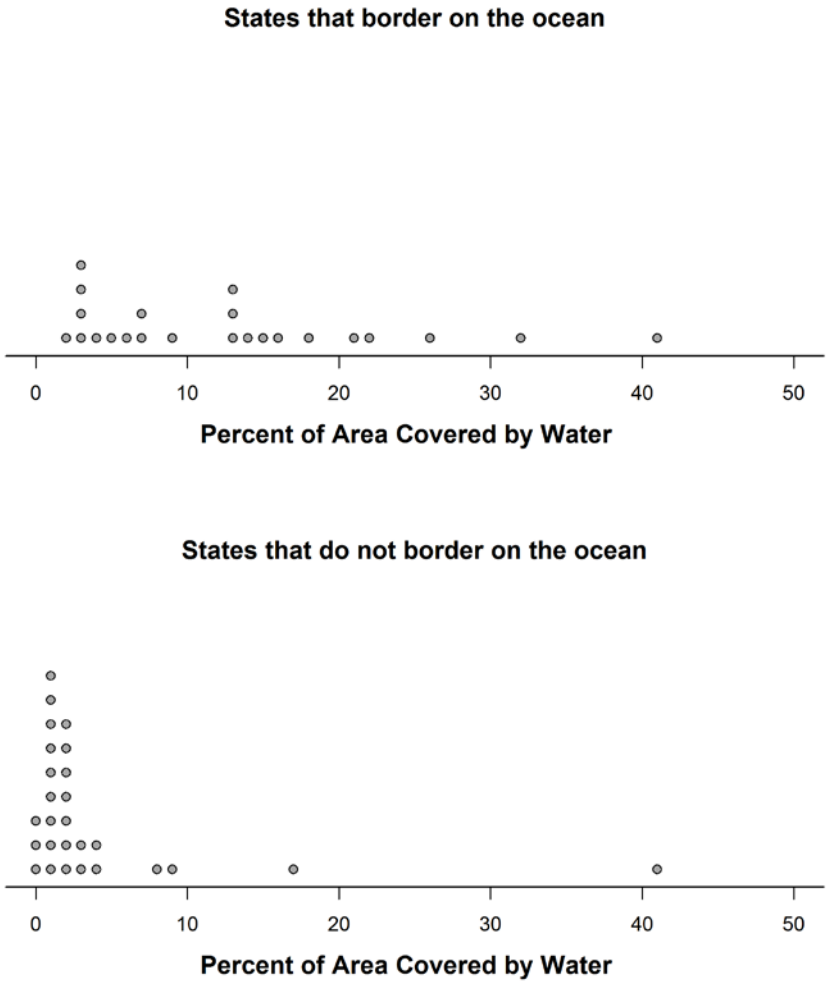
Students performed slightly worse on items that were in the collecting data and the interpreting data areas. Items in the collecting data area involve executing or implementing sampling or experimental assignment of treatments techniques. An example of a Level C item from the collecting data area is shown below. In the pilot, 62% of students answered this item correctly.

Example 2: Lee wants to answer the question, "What proportion of sophomores at my high school plan to take a foreign language class during the next school year?" Which of the following methods would best allow Lee to answer his question?

(A) Randomly select 50 students from the high school and ask them if they intend to take a foreign language class next year.
(B) Randomly select half of the foreign language teachers in the high school and ask them how many students are taking their classes this year.
(C) Randomly select half of the sophomores taking Spanish this year and ask them if they intend to take Spanish next year.
(D) Randomly select 40 sophomores from the high school and ask them if they intend to take a foreign language course next year.

The third category of questions involved the analysis of data. In the United States, analyzing data questions typically require the student to make calculations to find a number, such as the mean. LOCUS requires students to analyze data through a statistical lens to show that they understand what the data is telling them. An example of a Level C item for this area of the process is shown below. In the pilot, only 56% of students answered this item correctly.

Example 3: Carlton found data on the percent of area that is covered by water for each of the 50 states in the U.S. He made the dotplots below to compare the distributions for states that border an ocean and states that do not border an ocean.



Which of the following is the best statistical reason for using the median and interquartile range (IQR), rather than the mean and standard deviation, to compare the centers and spreads of these distributions?

(A) The mean and standard deviation are more strongly influenced by outliers than the median and IQR.
(B) The median and IQR are easier to calculate than the mean and standard deviation.
(C) The two groups contain different numbers of states, so the standard deviation is not appropriate.
(D) The two distributions have the same shape.

Items in the interpreting data area have the students answer an initial question by drawing conclusions from the data. An example of a Level C item from the interpreting data area is shown below. In the pilot, 68% of the students answered this item correctly.

Example 4: A survey of 625 randomly selected students was conducted to determine student opinion about music. The survey reported that 36 percent of the students surveyed preferred country/western music. The survey estimate had a margin of error of 4 percentage points. A margin of error is reported because

(A) Sample proportions vary from sample to sample.
(B) Students may intentionally respond incorrectly.
(C) Students may misunderstand the survey questions.
(D) The people doing the survey may have recorded results incorrectly.

Constructed Response

In addition to the 21 multiple-choice questions, each form of the assessment also had 5 constructed-response questions where students had to write in a response. A team of individuals that included project staff, members of the test development team, and graduate students, graded these responses. Many of the scorers had prior experience with the Advanced Placement Statistics grading process.

Table 4 shows the item difficulties across all forms (maximum score of 4.00), organized, again, by the areas of statistical problem solving process. The students continued to perform the best on items that required them to formulate questions. Students performed slightly worse on collecting data, next worse on analyzing data, and worst on interpreting data.

Table 4. Item Difficulties by Statistical Problem Solving Process - CR

Process	Mean
Formulating Questions	1.49
Collecting Data	1.30
Analyzing Data	1.32
Interpreting Data	1.25

The following provides an example of a constructed response question that was on the pilot exam:

Example 5: A public library is currently open from 9 a.m. to 5 p.m. on Saturdays. The director is considering whether or not to keep the library open until 8 p.m. on Saturdays. A library employee develops a one-question survey. The question is, would you use the library between the hours of 5 p.m. and 8 p.m. on Saturdays? The survey was administered using two different methods.

In Method 1, 100 individuals were selected at random from a list of people who have library cards at that library.

(a) To what population can the results of Method 1 be generalized?

In Method 2, the survey was given to all 25 individuals who were in the library at 4:45 p.m. on a particular Saturday. The results of the two surveys are summarized in the table below.

Method	Yes	No
1	30	70
2	15	10

(b) Create a graphical display that allows you to compare the results of the two surveys.

(c) Why do you think the two methods produced such different results?

The mean score on this particular item was 1.8 out of a possible 4 points with a standard deviation of 1.3. The distribution of scores was 0 (23.7%), 1 (10.4%), 2 (28.4%), 3 (25.9%), 4 (5.1%), while 6.6% of the students omitted the question. Sample response to this question will be provided at the conference.

CONCLUSIONS

This paper presents preliminary data from the pilot administration of the LOCUS assessments aimed to measures students' statistical understanding at the high school (ages 14-18) level. The data obtained through the pilot administration was utilized to construct representative forms that will be administered in the spring of 2014 with aim of created two equated forms for research and teaching purposes. These forms will be made available by January 2015 and will provide a valid and reliable tool to assess the conceptual understanding of school-based statistics. Please visit http://locus.education.ufl.edu for more information about this project as well as to obtain access to the assessments.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. DRL-1118168. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report*. Alexandría VA: American Statistical Association.

Friel, S. N., Bright, G. W., Frierson, D., & Kader, G. D. (1997). A framework for assessing knowledge and learning in statistics (K-8). In I. Gal and J. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 55-63). Amsterdam, The Netherlands: IOS Press (on behalf of ISI). http://iase-web.org/Books.php?p=book1

Jacobbe, T., Case, C., Whitaker, D., Foti, S. (2014). Establishing the content validity of the LOCUS assessment through evidence centered design. In K. Makar & R. Gould (Eds.), *Proceedings of the 9th International Conference on Teaching Statistics.*

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education, 3*(1), 1-9.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development.* Mahwah, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives, 1*(1), 3-62.