MIDDLE SCHOOL (AGES 10-13) STUDENTS' UNDERSTANDING OF STATISTICS

Tim Jacobbe, Steve Foti, and Douglas Whitaker
University of Florida, United States
fotisj@ufl.edu

This paper will present results from the administration of the LOCUS assessments to measure students' statistical understanding in grades 6-8 (ages 10 – 13). 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 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.

COMMON CORE EXPECTATIONS

The CCSS for mathematics introduce statistics and probability into the curriculum beginning in grade 6. After the students' first encounter with the subject, the standards require them to have developed a basic understanding of variability and to be able to summarize and describe data distributions. Students should be able to determine whether or not a context is statistical in nature and also be able to make basic graphical displays of data. In grade 7, students are introduced to more advanced topics such as the use of random sampling to make inferences about a single population or to draw informal comparisons between two populations. They will also be introduced to chance processes for the purpose of developing, using, and evaluating probability models. In their final year of middle school, grade 8, students will learn how to find patterns of association in bivariate data by creating and interpreting scatter plots, using linear regression models, and by using frequency tables (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). For some perspective, the level of statistical understanding outlined above is on par with introductory classes at tertiary schools in 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

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 [© 2014 ISI/IASE]

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 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. In each school, multiple teachers administered the LOCUS assessment to their students. All four forms appeared in each classroom that participated in the pilot. Table 1 shows the demographic information for the 2,075 students who participated in the pilot.

Table 1. Demographic Information for Students in Pilot

Gender		Grade		English*	
Female	45.73 %	6	25.98 %	No	8.29 %
Male	49.98 %	7	31.71 %	Yes	87.57 %
Race		8	39.33 %	Ethnicity	
Not Hispanic	87.23 %	9	0.92 %	AmIndian/PacIslander	2.07 %
Mexican	2.70 %	10	0.00 %	African American	5.40 %
Puerto Rican	2.17 %	11	0.14 %	Asian	5.01 %
Cuban	1.06 %	12	1.93 %	White	74.65 %
Other Hispanic	3.33 %			Other	5.01 %

**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 39 total possible points (multiple-choice and constructed-response combined), the minimum score across all of the forms was 2, and the maximum score across all of the forms was 34. 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	15.80	13.83	16.68	16.34
Std. Deviation	4.71	4.32	4.96	5.83

Multiple Choice

For the multiple-choice part of the exam, the mean item difficulties of the four forms were between 0.50 and 0.59. The biserial correlations for the forms were between 0.39 and 0.44. 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	13	0.64
Collecting Data	19	0.61
Analyzing Data	29	0.41
Interpreting Data	15	0.58

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 B item for this area of the statistical problem solving process is shown below. In the pilot, 58% of the students were able to correctly answer the item.

Example 1: For their final project, students in a math class are required to answer a question by collecting data about students at their school. For which of the following questions could a random sample of students provide the best approximate answer?

(A) How many students attend the school?
(B) How many hours does each class at the school meet per year?
(C) How many text messages do students at the school send per week?
(D) Do students at this school have higher test scores than students in other schools in the district?

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 B item from the collecting data area is shown below. In the pilot, 60% of students answered this item correctly.

Example 2: An advertisement makes the claim: "Lighter shoes make you run faster." Of the following, which is the best way to investigate this claim?

(A) Choose the records of the top twenty runners who are wearing the lighter shoes and compare their times to run 400 meters before and after they began wearing the shoes.
(B) Choose twenty runners and select ten at random to wear lighter shoes and have the other ten wear heavier shoes to run 400 meters and compare their times.
(C) Choose twenty runners at random and have the women wear the lighter shoes and the men wear the heavier shoes to run 400 meters and compare their times.
(D) Choose to observe the results of 400-meter races for the next year and see how many winners are wearing the lighter shoes.

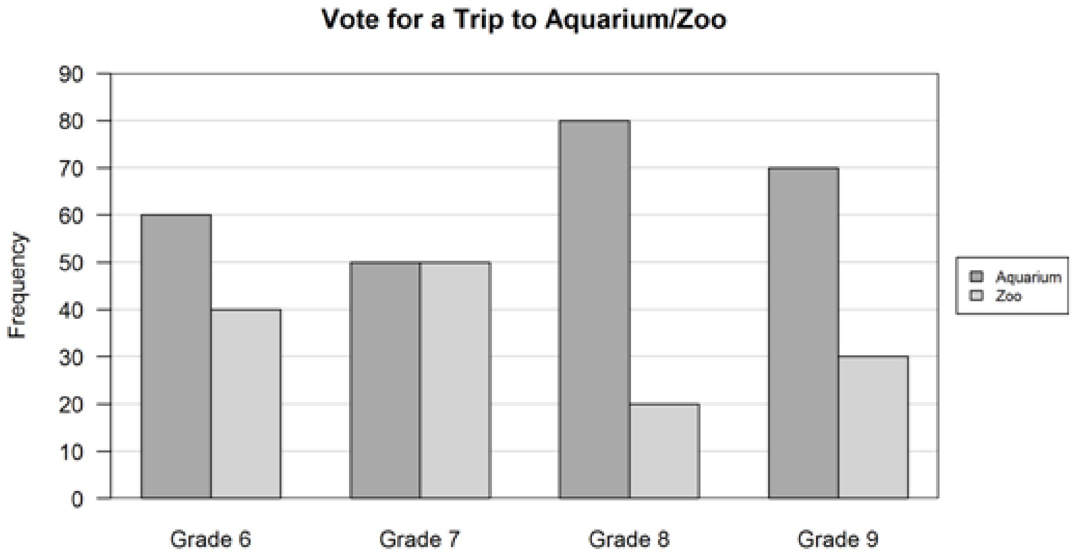
Items in the interpreting data area have the students answer an initial question by drawing conclusions from the data. An example of a Level A item from the interpreting data area is shown below. In the pilot, 68% of the students answered this item correctly.

Example 3: A recent study on habits of elementary and middle school students reported that elementary school boys typically watch 2 hours more TV per week than girls, and middle school boys typically watch 4 hours more TV per week than girls. Based on these data, can we conclude that in high school boys typically watch 6 hours more TV per week than girls?

(A) No, because data were not collected from high school students.
(B) No, because high school students can drive and don't watch as much TV.
(C) Yes, because 6 hours is the next number in the pattern.
(D) Yes, it is reasonable to predict TV habits of high school students from middle school students.

Items which required students to analyze data were substantially more difficult for the students. It is important to note, however, that items on the LOCUS assessment that are categorized as analyzing data items are more difficult than items found on typical statistics assessments. Usually, analyzing data questions 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 A item for this area of the process is shown below. In the pilot, only 21% of students answered this item correctly. Most students chose option (B) as the *most consistent* response, which reveals a misconception about reading this type of graphical display.

Example 4: A school is planning a field trip to the aquarium or to the zoo for students in grades 6 through 9. There are 100 students in each grade level and every student was asked which place he or she would prefer to visit. The bar graphs for the four grade levels are shown below.



In which grade level were the responses most consistent?

(A) Grade 6
(B) Grade 7
(C) Grade 8
(D) Grade 9

Constructed Response

The constructed response part of the assessment gave students more trouble. The mean scores, out of 4, on the four forms were 0.78, 0.58, 1.30, and 1.28, respectively. The highest mean score on any of the CR items across all forms was 1.9 and the lowest was 0.1. The polyserial correlations were between 0.57 and 0.65.

Table 4 show the item difficulties across all forms, 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 and analyzing data, and worst on interpreting data.

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

Process	Mean
Formulating Questions	1.74
Collecting Data	1.39
Analyzing Data	0.95
Interpreting Data	0.78

CONCLUSIONS

The LOCUS assessment provides anyone interested with the ability to evaluate a learner's conceptual understanding of statistics. Through the scoring process, this assessment reveals which areas of the statistical problem solving process the learner excels at, and which areas give them difficulty. The multiple-choice and the constructed-response items have both been built to reveal misconceptions and establish which level of statistical development the learner is at, according to the GAISE framework. The assessment can be used as a tool to not only evaluate students' understanding of statistics, but to also help identify what misconceptions need to be addressed. A more in-depth analysis of the pilot data will be used to inform the final forms of the assessment.

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.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics.* Washington, DC: Authors.