THE LOCUS ASSESSMENT AT THE COLLEGE LEVEL: CONCEPTUAL UNDERSTANDING IN INTRODUCTORY STATISTICS

Tim Jacobbe, Douglas Whitaker, Catherine Case, and Steve Foti
University of Florida, USA
whitaker@ufl.edu

LOCUS is a collection of assessments designed to measure conceptual understanding of statistics at the levels hypothesized by the GAISE framework. The original target population of the assessments is students in grades 6-12, but a version (pre/post) was constructed to meet the needs of introductory statistics instructors and students. With the implementation of the Common Core State Standards, the statistical knowledge of enrolled students is expected to change, and the LOCUS assessment for introductory statistics will be able to measure that change and possibly serve as a placement test. The LOCUS assessment is compared to the CAOS instrument and the ARTIST project. Preliminary results of the validity study with introductory college students are presented with attention given to the utility of the LOCUS assessment for placement purposes.

INTRODUCTION

In 2003, the American Statistical Association (ASA) began work on two documents that provide guidelines for the instruction and assessment of statistics at the K-12 and introductory college level: the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework* (Franklin et al., 2007) and the *GAISE College Report* (ASA, 2005). The Levels of Conceptual Understanding in Statistics (LOCUS) project has produced assessments that characterize the understanding of statistics for students in grades 6-12 according to the levels hypothesized in the GAISE framework. With widespread adoption of the Common Core State Standards for Mathematics (CCSS-M; National Governors Association Center for Best Practices [NGACBP] & Council of Chief State School Officers [CCSSO], 2010) in the United States, students entering college have the potential to possess greater knowledge of statistics than they do now. Introductory undergraduate statistics courses would need to adapt in response to changes in student knowledge. This paper describes the development of the LOCUS family of assessments, presents the ongoing validation work with students enrolled in introductory statistics courses, and compares these assessments with commonly used instruments: the Comprehensive Assessment of Outcomes in a First Statistics course (CAOS) instrument and Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project.

THE LOCUS PROJECT

LOCUS is an NSF-Funded (DRL-1118168) project focused on the development of assessments of statistical literacy. The intent of these assessments is to provide teachers, educational leaders, assessment specialists, and researchers with a valid and reliable assessment of statistics consistent with the Common Core State Standards (CCSS). These materials are designed to serve as exemplars for the development of high stakes assessments or curriculum materials. LOCUS assesses students' understanding across levels of development (i.e., A, B, or C) as identified in the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE).

The LOCUS project is on-track to produce two assessments: an assessment for students at the A/B level (targeted at grades 6-9) and an assessment for students at the B/C level (targeted at grades 10-12). For both the A/B and B/C assessments, forms are being created for 'pre' and 'post' administrations along with items and materials for public consumption. Additionally, two versions are being constructed: one with 23 multiple choice and 5 constructed response items (intended to be administered in two 45-minute class sessions or one 90-minute session), and one with 30 multiple choice items (intended to be administered in one 60-minute session).

Overview of Process Used to Develop the LOCUS Assessments

Evidence-centered design (ECD) is a process for creating assessments that seeks to make explicit and transparent the evidentiary reasoning used to make claims and inferences about the larger domain of knowledge, skills, and abilities being assessed through the use of observable

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

evidence supporting the claims and inferences (Mislevy, Steinberg, & Almond, 2003; Mislevy & Riconscente, 2006; Huff & Plake, 2010). At its heart, the validity of an assessment is an argument that the assessment is suitable for a specific purpose and that interpretations of results are meaningful within the established context (Kane, 2006). The formalized evidentiary reasoning process precipitated by the use of ECD is one component of the validity argument (Mislevy & Riconscente, 2006).

ECD formalizes the process of evidentiary reasoning as a means of establishing assessment validity by defining a series of layers that make up the assessment process: domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery (Mislevy & Riconscente, 2006). Throughout, artifacts are produced which document the activities in the ECD process and ensure that each layer follows logically from the work done in previous layers. In practice, ECD is not a linear sequence of activities but an iterative process with modifications made within and across layers (Huff, Steinberg, & Matts, 2010). The implementation of ECD is expected to be tailored to each project's goals, resulting in considerable variation among projects operating within the overall ECD framework (Brennan, 2010). Furthermore, a specific application of the ECD framework need not be a full implementation (Brennan, 2010).

Use of ECD in the Development of the LOCUS Assessments

The LOCUS project utilized an expert advisory board and two test development committees (TDCs – one each for the A/B and B/C assessments) to engage the five layers of the ECD framework. The expert advisory board participated in the domain analysis, domain modeling, and conceptual assessment framework layers, and the TDCs were used in the conceptual assessment framework, assessment implementation, and assessment delivery layers. The domain analysis was initially informed by a three dimensional framework: the two dimensional GAISE framework (Process of Statistics x Levels of Development) and the statistical content topics from the CCSS-M. This initial work was then used in the domain modeling layer to create a mapping between the two dimensional GAISE framework and the CCSS-M; this mapping substituted for the more conventional (in ECD) narrative argument for the type of understanding that should be assessed. This narrative, together with input from the TDC members, was used by the advisory board to finalize the conceptual assessment framework, referred to as the evidence model.

The TDCs used the evidence model to iteratively create items individually and at group meetings. When issues arose, the advisory board was consulted. After pilot forms were assembled, the TDCs and advisory board reviewed them. The pilot forms were also reviewed by the National Council of Teachers of Mathematics (NCTM) and ASA Joint Committee on Curriculum in Statistics and Probability. The feedback on the pilot forms was used to revise the items and pilot forms.

A timeline of the conceptual assessment framework and pilot assessment implementation layers of the ECD process are provided in Table 1. The timeline demonstrates the strong commitment made to iterative creation and refinement of the assessment as well as movement between the layers evidenced by repeated consultations with the advisory board. The pilot assessments were administered to 3,324 students in grades 6-12 (2,075 students took Form A/B, 1,249 took Form B/C), with 285 students enrolled in AP Statistics classes among that number.

A scoring session for the pilot assessment was held after the pilot administration. The participants in the scoring session were several members of the TDCs and graduate students with an interest in statistics education. Draft rubrics and items were created by the TDCs simultaneously. During the scoring session, these rubrics were modified to attend not only to the statistical accuracy of responses but also to level (A, B, or C) of response and common misconceptions. The rubrics were modified again at the following meeting of the TDCs.

Consonant with the iterative nature of the ECD process, the items and rubrics are still preliminary pending final review by the TDCs, advisory board, and NCTM-ASA Joint Committee. The construction of the operational forms was completed this year, concluding the assessment implementation phase and segueing into the assessment delivery phase, scheduled for April-May 2014 with students in grades 6-12. Final work on the pre/post forms for the A/B and B/C assessments targeting students in grades 6-12 should be complete by August 2014. The purpose of this paper is to justify the potential use of the LOCUS assessments with college students.

Table 1. Timeline of conceptual assessment framework and pilot assessment implementation layers.

Time Period	Event
September – November 2011	Initial mapping of GAISE and CCSS-M
November 2011	Initial review of mapping by EAB
January 2012	Initial creation and review of evidence statements by EAB for the Evidence Model
February 2012	Evidence Model reviewed and finalized by EAB
March – October 2012	Iterative creation and review of items by TDCs at group meetings and individually. EAB consulted regarding issues that arose.
October – December 2012	Assembly of 8 pilot forms
December 2012 – January 2013	Simultaneous review of pilot forms by TDCs, EAB, and JC
February 2013	TDCs revise items and forms using feedback from simultaneous review.
February – April 2013	Layout and printing of pilot forms
April – June 2013	Pilot administration (approximately 3,500 students)

Note: EAB = Expert Advisory Board; TDCs = Test Development Committees; JC = NCTM-ASA Joint Committee on Statistics and Probability

CURRENT INSTRUMENTS IN USE

ARTIST Project

The ARTIST project was developed to provide instructors of tertiary-level introductory statistics courses across the disciplines with resources for assessing students' statistical thinking and literacy (Garfield & delMas, 2010). Garfield and delMas (2010) describe the ARTIST website as providing: (1) access to an item bank containing more than 1,000 items, (2) online tests that align with key areas in a first statistics course, (3) the CAOS and START tests (detailed below), and (4) general resources about alternative forms of assessment in statistics. The ARTIST website has proved popular with tertiary and secondary instructors: of the more than 1,400 instructors that have used the item bank, 76% are from colleges and universities and 20% are from high schools (Garfield & delMas, 2010). As part of the LOCUS project's initial item development process, the ARTIST item bank was considered as possible source; after consideration, no items from the ARTIST project were ultimately included.

CAOS Instrument

The CAOS instrument is an important and popular component of the ARTIST project designed to address the lack of a standard instrument aligned with the consensus introductory statistics curriculum. The CAOS test is a 40-item, multiple choice, online assessment developed with the ARTIST item bank as an initial item pool, with items revised or created as needed (delMas, Garfield, Ooms, & Chance, 2007). A subset of 14 CAOS items form the aforementioned START test. The assessment went through multiple revisions during the design process and was validated using a large-scale sample of American students (delMas et al., 2007). By 2012, more than 5,000 students have taken the CAOS test and are used as a comparison group in research studies (Garfield, delMas, & Zieffler, 2012).

However, as the CAOS instrument was aligned with the content of the introductory statistics curriculum, alternative statistics courses may resort to using only the pieces of the test relevant to their modified curriculum (e.g. Garfield et al., 2012). Using substantially modified versions of the test jeopardize valuable psychometric properties, and a partial CAOS test used as one of several forms of research assessment in alternative tertiary-level statistics courses illustrate its place as the sole research assessment for tertiary introductory statistics courses may not be appropriate.

CURRICULA COMPARISONS

The ASA's GAISE reports provide a convenient starting point for comparing statistics curricula: Pre-K-12 and introductory statistics. In the United States, the widespread adoption of the CCSS-M suggests that that the Pre-K-12 statistics curriculum be further divided into primary (Pre-K-5) and secondary (6-12) level. Statistical understanding may be more appropriately viewed along a continuum and, as detailed below and suggested by Franklin et al. (2007), there is considerable overlap in material, particularly between late-secondary and introductory statistics.

Common Core State Standards for Mathematics

As part of the Domain Modeling phase of the ECD process employed in LOCUS, the specific pieces of the CCSS-M germane to data, statistics, and probability were mapped onto the GAISE framework. While the CCSS-M's coverage of the GAISE framework is not complete (e.g. Groth & Bargagliotti, 2012), there is at least some coverage in each of the resultant twelve constituent 'cells' of the two-dimensional framework.

The CCSS-M begins with "Standards for Mathematical Practice" – broad areas of expertise that should be the target of development among students of mathematics (NGACBP & CCSSO, 2010). Among these standards are four that have particular relevance for statistics: "Make sense of problems and persevere in solving them", "Construct viable arguments and critique the reasoning of others", "Model with mathematics", and "Use appropriate tools strategically". The remainder of the CCSS-M document is composed of content area "strands" that are consistent across several grade levels. In grades 1-5, the "Measurement and Data" strand contains some statistics material but is largely lacking the type of statistical content envisioned by the GAISE framework (Bargagliotti, 2012). In grades 6-12, the "Statistics and Probability" strand contains considerable statistical material corresponding at levels A, B, and C of the GAISE framework (but not exhausting the material recommended by it).

While the CCSS-M material does not exhaust the material in an introductory statistics course – far from it – much of the material nevertheless appears in introductory statistics courses. Common data representations such as bar charts, histograms, and boxplots, simple linear regression models, and the use of probability models as the basis for decisions are among the topics shared by introductory statistics and the CCSS-M. The LOCUS assessments attend to these and other topics, and therefore can serve as an ideal placement test for introductory statistics in addition to a measuring students' conceptual understanding at the end of a course.

Introductory Statistics

There is an element of curricular coherence to the introductory statistics course offered by a myriad of teaching units at the university level reflecting a consensus in the statistics community noted by Scheaffer (1997) and acknowledged by the ASA with the publication of the *GAISE College Report* (2005). This curricular coherence does not preclude considerable variability across courses and institutions with the labels *traditional* and *reform* sometimes being used. However, while Zieffler et al. (2008) note that there is not an operational definition of a *traditional statistics course*, the items on the CAOS assessment were developed so that they would be understandable to "students completing any introductory statistics course" (delMas et al., 2007, p. 30).

The content of traditional introductory statistics courses is typically comprised of three units receiving roughly equal attention: statistical design and descriptive, probability and sampling distributions, and statistical inference (Malone, Gabrosek, Curtiss, & Race, 2010; Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011). While the order of presentation can vary, a sketch of the typical topics covered is: summarizing data numerically and graphically; association, particularly Pearson's correlation coefficient; data collection, emphasizing sampling and experimental design; probability, including probability distributions and sampling distributions; confidence intervals and hypothesis tests, for comparing means and proportions for one and two groups based on a normal-theoretic framework; contingency tables and the Chi-Squared test; and inference for regression analysis (e.g., Agresti & Franklin, 2013). Tintle, Topliff, VanderStoep, Holmes, and Swanson (2012) give a tabular comparison between the consensus curriculum and their randomization curriculum, which notably begins immediately with an introduction to statistical inference. The CATALST curriculum – another randomization-based approach – is organized into three units: “Chance Models and Simulation”, “Models for Comparing Groups”, and “Estimating Models using Data” (Garfield et al., 2012, p. 887).

Furthermore, Scheaffer (1997) notes that the content of the AP Statistics course parallels the content of introductory statistics courses at the college level in the United States. The AP Statistics program has proved popular with over one million students sitting for the exam since its inception in 1997 (Rodriguez, 2012b), thereby precipitating a sizeable population of students enrolling in college that already know introductory statistics material. The lists of topics covered by the consensus introductory statistics curriculum above (e.g., Agresti & Franklin, 2013) and by the AP Statistics curriculum are essentially equivalent (College Board, 2010).

Implications for Future Curricula

While the implementation of the CCSS-M is far from a guarantee achieving its goal of (improving) the college and career readiness of students graduating from secondary school, even partial success in terms of increased data, statistics, and probability knowledge for students enrolling in college may necessitate substantial changes in the introductory statistics curriculum. Furthermore, while introductory statistics reform movement of the 1990s primarily focused on increase the role technology played in the classroom (e.g. through the increased use of graphing calculators; ASA, 2005), there is an increased focus on randomization-based statistics courses. Examples of randomization-based approaches include the CATALST course (Garfield et al., 2012), the work of Tintle et al. (2011, 2012), and the earlier work of Wardrop (1995).

The need to make changes to the undergraduate statistics curriculum in response to developments at the K-12 level is not new. Eight of the forty items on the CAOS instrument had greater than 60% correct responses on the pre-test from students in a nationally representative sample (delMas et al., 2007) indicating an appreciable amount of the introductory statistics curriculum being taught prior to college, and Rodriguez (2012a) attributes at least part of “soaring” (p. 4) number of undergraduates either majoring or minoring in statistics (and changes departments are making in response to this) to the success of the AP Statistics program.

VALIDATION OF LOCUS WITH UNDERGRADUATES

The LOCUS assessments are being piloted with students enrolled at a large research university in the United States in the first half of 2014. Further results should be available by July 2014. The results of this administration, along with an expert panel of university statistics instructors, will be the primary evidence used to justify the use of the assessments with students enrolled in college-level introductory statistics courses.

CONCLUSION

Throughout the development of the LOCUS assessments, careful attention has been paid to the validity argument for their use in grades 6-12. The LOCUS assessments are also appropriate for use at the introductory statistics level both as a placement test and as a research instrument to measure conceptual understanding of statistics. The validity for this use draws on the ECD process used in the assessments’ creation, the opinions of an expert instructor panel, and the performance of students enrolled in courses with the introductory statistics curriculum.

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

Agresti, A., & Franklin, C. A. (2013). *Statistics: The art and science of learning from data* (3rd ed.). Upper Saddle River, N.J: Pearson.

American Statistical Association. (2005). *GAISE College Report*. American Statistical Association.

Bargagliotti, A. E. (2012). How well do the NSF Funded Elementary Mathematics Curricula align with the GAISE report recommendations? *Journal of Statistics Education*, *20*(3), 1–26.

Brennan, R. L. (2010). Evidence-centered assessment design and the Advanced Placement Program®: A psychometrician's perspective. *Applied Measurement in Education*, *23*(4), 392–400. doi:10.1080/08957347.2010.510973

College Board. (2010). *AP Statistics Course Description*. Retrieved from http://apcentral.collegeboard.com/apc/public/repository/ap-statistics-course-description.pdf

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28-58.

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*. American Statistical Association. Retrieved from www.amstat.org/education/gaise

Garfield, J., & delMas, R. (2010). A web site that provides resources for assessing students' statistical literacy, reasoning and thinking. *Teaching Statistics*, *32*(1), 2–7.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883–898.

Groth, R. E., & Bargagliotti, A. E. (2012). GAISEing into the Common Core of Statistics. *Mathematics Teaching in the Middle School*, *18*(1), 38–45.

Huff, K., & Plake, B. (2010). Evidence-centered assessment design in practice. *Applied Measurement in Education*, *23*(4), 307–309. doi:10.1080/08957347.2010.510955

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, *23*(4), 310–324. doi:10.1080/08957347.2010.510956

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–153). Mahwah, N.J.: Lawrence Erlbaum Associates.

Malone, C. J., Gabrosek, J., Curtiss, P., & Race, M. (2010). Resequencing topics in an introductory applied statistics course. *The American Statistician*, *64*(1), 52–58.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, N.J.: Lawrence Erlbaum Associates.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). 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, D.C.: National Governors Association Center for Best Practices, Council of Chief State School Officers.

Rodriguez, R. (2012a, August). A major trend: The rise of undergraduate programs in statistics. *AmStat News*, (422), 3–4.

Rodriguez, R. (2012b, September). More than 1 million and counting: The growth of Advanced Placement Statistics. *AmStat News*, (423), 3–5.

Scheaffer, R. L. (1997). Discussion. *International Statistical Review/Revue Internationale de Statistique*, *65*(2), 156–158.

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal*, *11*(1), 21–40.

Tintle, N., VanderStoep, J., Holmes, V., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, *19*(1), 1–25.

Wardrop, R. L. (1995). *Statistics: Learning in the presence of variation*. Dubuque, Iowa: Wm. C. Brown.

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. *Journal of Statistics Education*, *16*(2), 1–23.