ESTABLISHING THE VALIDITY OF THE LOCUS ASSESSMENTS THROUGH AN EVIDENCED-CENTERED DESIGN APPROACH

Tim Jacobbe, Catherine Case, Douglas Whitaker, and Steve Foti
School of Teaching and Learning, University of Florida, Gainesville, FL, USA
jacobbe@coe.ufl.edu

This paper presents the systematic process utilized by the Levels of Conceptual Understanding in Statistics (LOCUS) project to establish content validity for assessments measuring students' statistical understanding in grades 6-12 (ages 11-18). Evidence Centered Design (ECD) was used to develop assessments aligned with the United States' Common Core State Standards in Mathematics (CCSSM) as well as the Guidelines for Assessment and Instruction in Statistics Education (GAISE). The ECD process began with a domain analysis based on CCSSM, GAISE, and learning trajectories from statistics education research and subsequently added layers articulating claims about student proficiency and observable evidence to support those claims. The ECD approach formalized the evidentiary reasoning by which performance on LOCUS can be used to support valid inferences about the larger domain of statistical understanding.

BACKGROUND

Purpose of the LOCUS Assessments

Over the past 25 years, inclusion of statistics at the school level and recognition of the importance of statistical literacy have been gaining momentum in the United States. These ongoing efforts to promote statistical literacy are exemplified by the American Statistical Association's (ASA) *Guidelines for Assessment and Instruction in Statistics Education (GAISE)* (Franklin, et al., 2007) which identify three levels of statistical understanding (Levels A, B, and C) ideally achieved at the school level. Influenced by the GAISE document, the widely adopted Common Core State Standards in Mathematics (CCSSM) (NGACBP & CCSSO, 2010) have considerably raised the expectations for statistics learning in grades 6-12 (ages 11-18) across the country. However, inclusion in the standards does not guarantee that statistics will be taught at a level sufficient to produce statistically literate citizens. Gal & Garfield (1997) note that teaching of statistics often reflects the way it is assessed on large-scale assessments, with an emphasis on procedural knowledge rather than conceptual understanding. Further, increased emphasis on statistics in the precollege curriculum warrants research investigating students' understanding of statistics and the effectiveness of instructional interventions, purposes for which existing large-scale assessments and instructor-designed exams are typically ill-suited (Gal & Garfield, 1997). Continued progress in the field of statistics education demands a valid and reliable assessment of statistical understanding

The Levels of Conceptual Understanding in Statistics (LOCUS) project, funded by the National Science Foundation (DRL-1118168), is addressing these assessment issues by addressing three broad goals:

- Develop instruments to assess levels of statistical understanding as initially defined in the GAISE framework and the CCSSM
- Provide a characterization of grade 6-12 students' current level of statistical understanding
- Provide a tool for researchers and teachers to assess growth in statistical understanding

LOCUS differs from traditional large-scale assessments in a few critical ways. First, LOCUS intends to measure conceptual understanding at all stages of the statistical problem-solving process (Franklin, et al., 2007), de-emphasizing rote calculations. Second, LOCUS takes into consideration the distinction between mathematical and statistical reasoning, noting that statistical reasoning is inextricably linked to context and must at every stage account for variability (Cobb & Moore, 1997; delMas, 2005).

Use of Evidence-Centered Design to Establish Validity

The goal of educational assessment is to use students' performance on a relatively small set of tasks as evidence to support valid inferences about a larger domain of knowledge, skills, 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

abilities (Mislevy, Steinberg, & Almond, 2003). Evidence-centered design (ECD), (Mislevy & Riconscente, 2006) makes explicit this exercise in evidentiary reasoning by systematically articulating the claims to be made about student proficiency and the observable evidence that supports those claims. Acknowledging that valid assessment is broader than writing quality items, ECD is a process that begins with an analysis of the domain of interest and subsequently adds layers detailing the assessment argument and the design of an assessment system that embodies that argument (Mislevy & Riconscente, 2006).

The central role of evidentiary reasoning in ECD (Mislevy, Steinberg, & Almond, 2003) is grounded in research in assessment validity (Kane, 2006; Messick, 1994). The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) defines assessment validity as "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of test scores," and based on this definition, Kane (2006) characterizes validity as an argument. As such, the intended interpretations or claims based on the test scores should be clearly stated, and the various assumptions and inferences connecting the test scores to those claims should be stated and justified (Kane, 2006). To justify inference from responses on a set of assessment tasks to competence in the domain of interest, Messick (1994) outlines a chain of reasoning that begins with the intended conclusions and works backward to item creation. First, the domain should be analyzed to determine which knowledge, skills, and abilities are valued and should be assessed; then test developers should consider which performances and behaviors exhibit the valued knowledge, skills, and abilities and how these performances and behaviors can be prompted through assessment tasks.

ECD formalizes the process of evidentiary reasoning for establishing assessment validity by defining a series of five 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).

Brief Overview of the Implementation of ECD for LOCUS

The developers of the LOCUS assessment used a modified evidence-centered design in order to create a robust validity argument for LOCUS as an assessment of statistical literacy for characterizing students' current understanding and measuring growth. The five layers of the ECD process were primarily carried out by three distinct teams of experts: an advisory board and two test development committees (TDCs). The principal assignment for the advisory board was to establish a blueprint for the assessment, a task which encompasses the first three layers of the ECD process and culminates in a detailed evidence model. ECD guarantees rigor in the creation of this evidence model, which is a critical piece of the assessment's validity argument.

After the evidence model was finalized, the TDCs were charged with writing and revising items for two versions of the test, Level A/B and Level B/C corresponding to the levels of development defined in the GAISE framework, and providing input for the assembly of the forms. The Level A/B assessments are designed for students in grades 6-9 (ages 11 – 15) while the B/C assessments are designed for students in grades 10 – 12 (ages 15 – 18). The pilot and operational forms were reviewed by the advisory board and the ASA-NCTM Joint Committee on Curriculum in Statistics and Probability to verify the assessments' alignment with the evidence model before they were administered.

Each member of the advisory board, the TDCs, and the project staff has an established record in the statistics education community and/or the development of large-scale assessments. The members of the advisory board are David Miller, a Professor of Research, Evaluation, and Measurement at the University of Florida specializing in item analysis and construction of assessments; Mike Shaughnessy, a Professor Emeritus in the Department of Mathematical Sciences at Portland State University and former President of the National Council of Teachers of Mathematics; Dick Sheaffer, a Professor Emeritus in the Department of Statistics at the University of Florida and the first chief reader for the AP Statistics exam; and Jane Watson, a Professor of Mathematics Education at the University of Tasmania and one of the most prolific researchers in

the statistics education community (Shaughnessy, 2007). The advisory board, TDCs, and project staff participated in the test development process through individual expert review of ECD artifacts and assessment items as well as through online and face-to-face meetings.

LAYERS OF THE ECD PROCESS

Domain Analysis

The first layer of ECD, domain analysis, involves "gathering substantive information about the domain of interest that has implications for assessments" and is an essential step towards establishing a coherent assessment argument (Mislevy & Riconscente, 2006). The process involves synthesis of information from various sources that may include national content standards, curriculum materials, and research-based conceptualizations of learning in order to determine what constitutes valued work and valued knowledge.

The two principal sources consulted during the domain analysis for the LOCUS assessment were the PreK-12 *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) framework and the CCSSM (NGACBP & CCSSO, 2010). The GAISE framework is a two-dimensional conceptual framework of statistical literacy. One dimension consists of the three levels of development: A, B, and C. The other consists of the four components of the statistical problem-solving process: formulating questions, collecting data, analyzing data, and interpreting results. In order to ensure that the LOCUS assessment would also be aligned with statistical content and skills required in schools in the United States under the Common Core State Standards, the GAISE framework was mapped onto the CCSSM. This mapping allowed the major claim of statistical literacy to be broken down into four claims corresponding to the components of the statistical problem-solving process. These claims were then further broken down into sub claims by the level of development and the particular content standard.

Domain Modeling

Building on information gathered in the previous layer, domain modeling entails the initial construction of an assessment argument, providing a structure for the more formal argument to come later (Mislevy & Riconscente, 2006). As part of the domain modeling for the LOCUS assessment, the advisory board reviewed and rated the initial mapping of the GAISE framework onto the CCSSM and provided an argument for the appropriateness of the mapping as a representation of the domain. This review served as the foundation for the evidence model created in the following layer.

Moreover, research on students' learning of statistical concepts and processes informed the domain analysis and modeling as the advisory board drew from models that have been presented in the statistics education literature. The advisory board considered comprehensive models of cognitive development in statistical reasoning as well as models characterizing students' understanding of specific statistical topics, such as measures of center and variation . The literature review by Jones, Langrall, Mooney, and Thornton (2004), synthesizes these various models of development in statistical reasoning.

Conceptual Assessment Framework

Based on the general analysis and assessment argument completed in the first two layers, the conceptual assessment framework consists of detailed specifications for the assessment. After articulating the claims to be made about students' statistical literacy by mapping the GAISE framework onto the CCSSM, the advisory board drafted an evidence model detailing the observable evidence that supports those claims. For each sub claim, the model included an evidence statement, possible work products, upper category observable features, and lower category observable features. That is, a formal description of what constitutes evidence of statistical literacy was provided for each content standard within each component of the statistical problem-solving process at each stage of development. In subsequent layers, items were designed to prompt these work products and observable features as evidence of student proficiency.

Once the extensive task of drafting the evidence model was complete, evidence statements were reviewed (rated 1, 2, or 3); every cell of the evidence model was revisited and many were

substantially revised before the evidence model was finally confirmed. The rigorous process of articulating student understandings in terms of observable performances and behaviors resulted in a clear description of the integration of content and skill to be measured (Ewing, Packman, Hamen, & Clark, 2010).

In addition to creating an evidence model, the advisory board made other detailed specifications for the assessment. For example, they specified the item format and decided the percentage of assessment items to be dedicated to each component of the statistical problem-solving process. The LOCUS assessments include both multiple choice and constructed response items with the prioritization of process components specified as follows:

- Formulating Statistical Questions: 10-15% on Level A/B and 15-20% on Level B/C
- Collecting Data: 25-30% on Level A/B and 20-25% on Level B/C
- Analyzing Data: 30-35% on Level A/B and 25-30% on Level B/C
- Interpreting Results: 25-30% on Level A/B and 30-35% on Level B/C

Again, the decisions made by the advisory board were informed by the research literature. Gal & Garfield (1997) argue that multiple choice or short answers questions alone may not reflect the inherent complexity of statistical problems. They also note that traditional assessments too often focus on the accuracy of computations or application of formulas, whereas the prioritization specified above better reflects the breadth of the statistical problem-solving process.

Assessment Implementation

The assessment implementation layer includes work common to all effective methods of test development such as writing and revising items, fitting psychometric models, and creating rubrics for scoring responses (Mislevy & Riconscente, 2006). Although ECD does not prescribe specific procedures for assessment implementation, these test development tasks play a critical role in validation, so at this layer, ECD is complemented by conventional best-practices such as those described in Schmeiser and Welch (2006). The key advantage is that the quality of artifacts with which trained item writers are presented are necessarily of a higher quality with regard to specificity and transparency than those of conventional item development (Ewing, et al., 2010).

Once the LOCUS evidence model and test specifications were confirmed, the TDCs began the iterative process of writing, reviewing and revising items. At each stage, items were reviewed not only for accuracy and clarity, but also to ensure alignment of the content and process to the evidence models. Additionally, considerations of fairness to different populations of students, such as readability and biases associated with particular items, were taken into account by TDC members with years of large-scale assessment experience and fairness training.

After being introduced to the evidence model, the TDCs carried out several cycles of writing and revision through face-to-face and online meetings. Through the process, some items were removed from the item pool, and many of those that remained underwent substantial revisions. At each iteration, writing assignments became more focused in order to create items matching test specifications for eight pilot forms.

After the pilot forms were assembled, the forms were simultaneously reviewed by the advisory board, the TDCs, and the ASA/NCTM Joint Committee on Curriculum in Statistics and Probability Committee. Each person completing the review was instructed to respond to the multiple choice and constructed response items for key verification, to verify the alignment of each item with the specified content category, to suggest specific revisions, and to provide an overall reaction to forms and process focus. These reviews were used to make final revisions to the items prior to the pilot administrations.

Pilot forms were administered in high-performing school districts in six different states where statistics was taught prior to CCSSM: Arizona, Colorado, Florida, Georgia, Ohio, and New Jersey. In total, 2,075 students took the Level A/B assessments and 1,249 students took the Level B/C assessments. The constructed response items were scored by TDC members and graduate students at the University of Florida according to detailed rubrics, and the scores of both the multiple choice and constructed response items were evaluated using psychometric analysis. Information gathered during the item analysis process—item difficulties, response patterns, item correlations with overall assessment scores, etc.—were then used to inform assembly of the operational forms. It is important to note that validity is related to interpretations and uses of test

scores not test *items*. A number of challenging psychometric tasks such as scaling and equating are necessary to produce valid scores, and these tasks are outside the scope of ECD (Brennan, 2010). In practice, ECD is supported by psychometric and test development models and methods not specific to ECD.

Informed by the analysis of pilot data, assembly of final forms began with the selection of the equating set—the set of ten multiple-choice items and two constructed-response items to appear on all A/B and B/C forms in order to allow comparisons of student performance across forms. The equating set was constrained to have a content distribution and overall difficulty similar to the overall assessment, among other considerations. Once this equating set was reviewed by the TDCs, the process began to create two similar forms of each version of the assessment to be used as a pre-test and a post-test for research. Assembly of these final forms required consideration of the distribution of difficulty across all process standards and selection of items that were similar in content and in difficulty to allow for strong comparisons between the pre-test and the post-test. The final forms were reviewed by the TDCs, the advisory board, and the ASA/NCTM Joint Committee before administration.

Assessment Delivery

The final layer of ECD describes how the administrator and the examinee will interact with the assessment (Mislevy & Risconcente, 2006). After the operational forms were finalized, they were administered in ten states and the results will be analyzed in the upcoming months. The operational forms (two at level A/B and two at level B/C) will be made available as secure assessments to be used as pre- and post-tests in research and to measure program/course effectiveness. Additionally piloted items that do not appear on the final forms of the LOCUS assessments will be made available online.

CONCLUSION

Validation of an assessment requires systematic evaluation of the network of inferences that connect performance on assessment tasks to a set of clearly articulated interpretations and uses (Kane, 2006). Use of the ECD approach formalized the evidentiary reasoning by which performance on the LOCUS assessment tasks can be used to support valid inferences about the larger domain of statistical understanding. In particular, efforts were made to establish the assessment's validity as a measure of statistical literacy to characterize students' current level of understanding of statistical topics and to assess growth. The mapping created in the domain analysis and modeling layers specified the various sub claims which constitute the larger construct of statistical literacy as articulated in the GAISE framework and aligned with the CCSSM. Further, the detailed evidence model created as part of the conceptual assessment framework provided justification for inference from observed student responses to the assessment claims.

Beyond the scope of ECD are a number of other test development tasks critical to assessment validity. After stating and justifying claims to be made about students' statistical literacy, LOCUS items were developed through an iterative process of writing and revision to ensure alignment with the evidence model and to control possible sources of extraneous variance. Operational forms were assembled according to detailed test specifications in order to guarantee that each form was appropriately representative of content standards and all components of the statistical problem-solving process. Psychometric analysis of pilot data was also considered in order to construct essentially parallel pre- and post-tests to be used in research.

Validation efforts always rely heavily on expert judgments about the plausibility of inferences connecting test performance to the proposed interpretations and uses (Kane, 2006). Together the members of the advisory board and the TDC represent a wealth of experience in statistics education and the development of large-scale assessments. Moreover, learning trajectories from statistics education research informed the test development process, as the advisory board and the TDC drew from their knowledge of the literature and their own experience.

The modified ECD approach utilized in the development of the LOCUS assessment provides well-documented evidence to support the inferences that link performance on the test to conclusions about statistical literacy. Specifically, LOCUS may be considered valid as a measure of students' current statistical understanding and as a research tool for assessing growth.

ACKNOWLEDGEMENT

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Brennan, R.L. (2010). Evidence-centered assessment design and the Advanced Placement program: A psychometrician's perspective. *Applied Measurement in Education*, *23*(4), 392-401.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801–823.

del Mas, R. C. (2005). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 79-95). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ewing, M., Packman, S., Hamen, C., & Clark, A. (2010). Representing targets of measurement within evidence-centered design. *Applied Measurement in Education*, *23*(4), 325-341.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Schaeffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) Report: A pre-K-12 curriculum framework*. Alexandria, VA: American Statistical Association.

Gal, I., & Garfield, J. B. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education*. Amsterdam; Washington, DC: IOS Press.

Jones, G. A., Langrall, C. W., Mooney, E. S., & Thornton, C. A. (2004). Models of development in statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 97–117). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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, NJ: Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *32*(2), 13-23.

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, NJ: Erlbaum.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3-67.

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

Schmeiser, C.B., & Welch, C.J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Washington DC: American Council on Education.