RECONCEPTUALIZING STATISTICAL LITERACY: DEVELOPING AN ASSESSMENT FOR THE MODERN INTRODUCTORY STATISTICS COURSE

Laura Ziegler
University of Minnesota, USA
sath0166@umn.edu

In light of changes to the introductory statistics course, namely the move to include simulation-based methods, new assessments are needed. The Basic Literacy In Statistics (BLIS) assessment was created to measure students' ability to read, understand, and communicate statistical information. The assessment was designed for students enrolled in introductory statistics courses at the postsecondary level that include some coverage of simulation-based methods in the curriculum. This paper describes the development of the instrument, including the validation of the test blueprint, pilot testing, and analysis of assessment data.

INTRODUCTION

The content taught in introductory statistics courses has changed a great deal over the past 10 years. In particular, simulation-based methods, such as randomization tests and bootstrapping, are being taught in addition to or in lieu of parametric tests (e.g., the *t*-test) in many tertiary-level introductory statistics courses (Garfield et al., 2012; Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011). Reasons for teaching simulation-based methods in an introductory statistics course include that the methods are more easily grasped by students (Cobb, 2007) and can be taught very early in a course—as early as the first day of class (Rossman, 2007).

Statistical literacy has been described as an important learning outcome in all introductory statistics courses (Garfield, delMas, & Zieffler, 2010). Definitions of statistical literacy vary from knowing the basic language of statistics (Garfield et al., 2007), to communicating, interpreting, and being critical of statistical information (Gal, 2002; Schield, 1999) and there is no consensus among researchers or teachers. The definition of statistical literacy adopted for the remainder of this paper is being able to read, understand, and communicate statistical information.

Using the definition of statistical literacy just adopted, there are a handful of assessments that have been used with students at the tertiary-level to measure statistical literacy. Each of these assessments has advantages and limitations. For example, the *Comprehensive Assessment of Outcomes in a First Statistics Course* (CAOS) test was created to assess students' statistical literacy and reasoning (delMas, Garfield, Ooms, & Chance, 2007). However, fewer than half of the items were judged to measure statistical literacy. The *ARTIST Topic Scales* are a set of 11 tests (Garfield, delMas, & Chance, 2002), which include more items that measure statistical literacy than the CAOS test. A majority of the literacy-based items in these scales, however, are limited to definitions and simple computations. In addition, validity and reliability evidence have not been collected for the *ARTIST Topic Scales*. The *ARTIST Topic Scales* and CAOS test are becoming outdated (having been written approximately 10 years ago) and do not include items related to simulation-based methods. The *Goals and Outcomes Associated with Learning Statistics* (GOALS) test is a more recent assessment designed to measure students' statistical reasoning after completing an introductory simulation-based statistics course (Garfield, delMas, & Zieffler, 2012). But, GOALS only includes three items that appear to assess statistical literacy.

It is clear there is a need for an assessment that measures students' statistical literacy—including being able to read, understand, and communicate statistical information—and that is aligned with the randomization and simulation methods. This assessment could be used for a variety of purposes. It could be used as a pretest to determine the level of literacy students have prior to taking an introductory statistics course. Additionally, researchers could also use the assessment to measure change in students' statistical literacy to evaluate different teaching methods or curricula. For example, do students have a higher level of statistical literacy when they are taught simulation-based methods followed by parametric methods, or when the teaching of simulation-based and parametric methods is consistently alternated? To address these questions, a new assessment of statistical literacy is needed. Lastly, instructors could also use such an

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

assessment formatively to provide students information about literacy-related topics that they do and do not understand.

The purpose of the study reported in this paper was to design an assessment, the *Basic Literacy In Statistics* (BLIS) assessment, which could be used to determine students' statistical literacy skills students in an introductory statistics course at the postsecondary level that includes some extent of simulation-based methods in the curriculum. A mixed-methods approach was used to collect evidence of the value, reliability, and validity of inferences for the BLIS assessment throughout the development process. The extent to which the BLIS assessment has value, reliability, and validity will be assessed using evidence collected via feedback from reviewers, student responses from cognitive interviews, results from a small-scale pilot test, and results from a large-scale field test.

TEST BLUEPRINT DEVELOPMENT

In order to create a preliminary test blueprint for the BLIS assessment, topics addressing statistical literacy were compiled from introductory statistics textbooks that utilized simulation-based methods to some extent (Catalysts for Change, 2013; Gould & Ryan, 2013; Lock, Lock, Lock Morgan, Lock, & Lock, 2013; Tintle et al., 2013). The topics chosen were related to the definition of statistical literacy presented earlier. They also met the goals of the assessment, including appropriateness for curricula with simulation-based methods in the curriculum (e.g., randomness, dotplots, randomization tests, and scope of conclusions).

For each topic, learning outcomes were specified using words that have been associated with items measuring statistical literacy such as *describe* and *interpret* (Garfield et al., 2010). For example, a learning outcome associated with one topic, *dotplots*, is "describing and interpreting the overall distribution of a variable as displayed in a dotplot, including referring to the context of the data". A total of 54 learning outcomes were identified in the preliminary test blueprint.

After the preliminary test blueprint was created, six statistics educators reviewed the test blueprint. The reviewers were purposefully chosen based on their expertise in teaching introductory statistics using simulation-based methods, assessment, and/or statistics education research. The reviewers were provided with the definition of statistical literacy, examples of what it would mean to be statistically literate, and the test blueprint. Using a 4-point scale, the reviewers rated the importance of each learning outcome in determining statistically literacy. Reviewers were also asked to describe important topics and learning outcomes that were not included in the test blueprint. If more than one reviewer suggested a topic, learning outcomes were created to be included in the final version of the test blueprint. For example, regression and correlation had not been included on the preliminary test blueprint and were added to the final test blueprint. Based on the ratings, some learning outcomes were also removed from the test blueprint. The final test blueprint had 37 learning outcomes.

ASSESSMENT DEVELOPMENT

Using the updated test blueprint as a guide, a preliminary version of the BLIS assessment was created using a combination of selected-response items from existing instruments (e.g., GOALS) and newly created constructed-response items. The constructed-response items will be converted to selected-response items after the BLIS assessment is piloted. This allows common incorrect student responses to be used as distractors for the selected-response options as recommended by Haladyna, Downing, and Rodriguez (2002). Each learning outcome was mapped to a single item, resulting in an assessment with 37 items.

Many characteristics of statistical literacy were considered when choosing and creating items. First, items needed to include a "real-world" context as recommended by Garfield et al. (2007). Textbooks were used to find "real-world" contexts (Gould & Ryan, 2013; Lock et al., 2013), as were published survey data and articles. Secondly, items required students to perform tasks rather than only provide definitions. Lastly, more emphasis was placed on items that required descriptions and interpretations than on items that required computing.

After the preliminary assessment was created, the six reviewers of the test blueprint also reviewed the assessment to determine how well the items aligned with the learning outcomes listed in the test blueprint. Using a 4-point scale, reviewers were asked to rate each item based on how

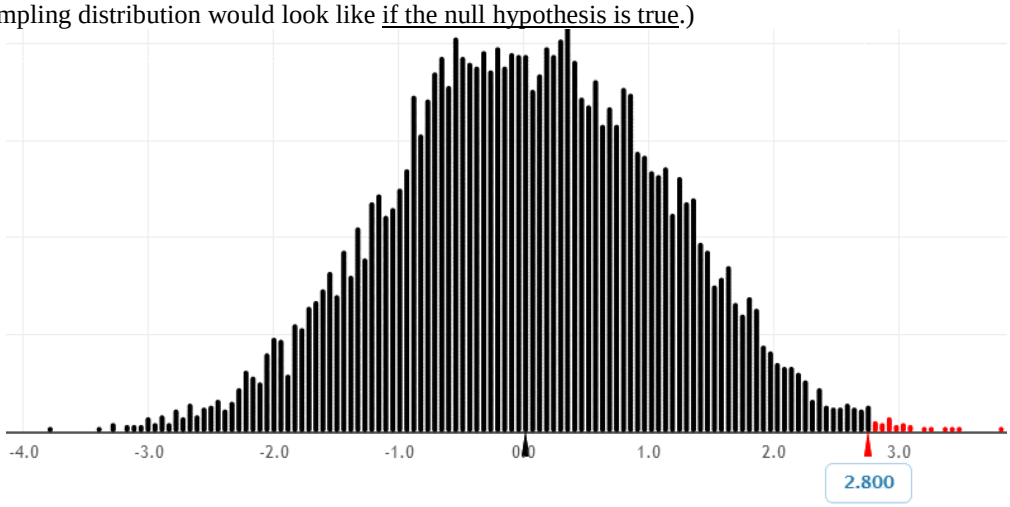
much they agreed that it measured the specified learning outcome. In addition, the reviewers were asked to provide any comments they had for each item. All reviewers provided suggestions for changes for multiple items (e.g., changes in wording) and the items were modified based on these suggestions. Figure 1 shows an example item and how it was changed based on reviewer feedback. There was only one item that had low agreement with the learning outcome. This item was designed to measure students' ability to interpret a percentage in the context of data. The item was subsequently replaced with a new item to better fit the learning outcome. After modifications were made to reflect the suggestions of the reviewers, the BLIS-1 assessment was complete.

Are people able to recall words better after taking a nap or taking a caffeine pill? A randomized experiment was conducted with 24 participants. Participants were shown a list of words in the morning. In the afternoon, half of the participants were randomly assigned to take a nap and the other half took a caffeine pill. The response variable was the number of words participants were able to recall 7 hours after being shown the list of words in the morning. The nap group recalled an average of 15.8 words and the caffeine group recalled an average of 13.0 words.

A randomization distribution was produced ~~graphed~~ by doing the following:

- From the original sample, the 24 participants were re-randomized to the nap group (n=12) or ~~and~~ caffeine group (n=12), without replacement.
- The mean difference in words recalled between the two re-randomized groups was computed [mean(nap group) – mean(caffeine group)] ~~for the re-randomized groups~~ and placed on the plot shown below.
- This was repeated 999 more times.

Below is the plot of the randomization distribution. (Note: This plot can be used as an estimate of what the sampling distribution would look like if the null hypothesis is true.)



(Words Recalled for Nap Group) – (Words Recalled for Caffeine Group)

The null hypothesis is there is no difference in the true mean number of words recalled for the nap group and caffeine group. Looking at the observed sample mean difference in number of words recalled for the nap group and the caffeine group of 2.8 on the plot, is there evidence against the null hypothesis? Explain your answer.

Figure 1. New assessment item with context based on an example from Gould and Ryan (2013) with modifications displayed based on reviewer feedback. Text that is crossed out was deleted from the preliminary version of the BLIS assessment and text that is underlined was added.

Six students participated in cognitive interviews from four different introductory statistics courses: a master's-level course taught with the *Statistics: Unlocking the Power of Data* (Lock et al., 2012) textbook, and two undergraduate-level courses including the CATALST course (Garfield, et al., 2012), which used the *Statistical Thinking: A Simulation Approach to Modeling Uncertainty* (Catalysts for Change, 2013) textbook, a course that used the *Introduction to*

Statistical Investigations (Tintle et al., 2013) textbook, and a course that used the *Investigating Statistical Concepts, Applications, and Methods* (Chance & Rossman, 2006) textbook. The interviews were conducted three to ten weeks after the students completed their introductory statistics course. During the interviews, students were asked to think aloud as they took the assessment by saying everything they were thinking about as they answered the questions.

Based on the cognitive interviews, 15 items were changed to create the BLIS-2 assessment. It was decided that three selected-response items taken from existing instruments did not include meaningful distractors so the items were converted to constructed-response items. There were eight items that had minor changes in non-statistical wording. The question in one constructed-response item was changed to be more specific in order to elicit more appropriate responses to measure the intended learning outcome. There was one selected-response item that included *none of the above* as one of the distractor options, which was deleted. Lastly, two items had slight modifications in statistical wording. For example, one selected-response item asked students to select the best interpretation of a graph. The correct answer included the range as the measure of variability, which was changed to include the standard deviation instead of the range.

The next stage in the development of the assessment was to administer the BLIS-2 assessment to students in introductory statistics courses that included simulation-based methods in a small-scale pilot. There were 76 students that participated who were enrolled in one of three introductory statistics courses: a master's-level course taught with the *Statistics: Unlocking the Power of Data* (Lock et al., 2012) textbook, and two undergraduate-level courses including the CATALST course (Garfield, et al., 2012), and a course taught with *Statistics* (McClave & Sincich, 2007). Based on an examination of the results, each of the selected-response options appeared to be viable options. However, there was one item that only 17.1% of students responded correctly. The item was taken from the CAOS assessment (delMas et al., 2007), however, an updated version of the item was included in an early version of the GOALS assessment (Garfield et al., 2012). Therefore, the item was changed to the updated version.

For the constructed-response items, student responses were examined to convert constructed-response items to selected-response for the third version of the BLIS assessment (BLIS-3). In order to create the selected-response options, students' responses were grouped by similar responses. The incorrect answers that were submitted most often and appeared to measure misconceptions related to the learning outcomes were included as distractors. Some of the answers were reworded to be similar in language, length, and style as the distractor options as recommended by Haladyna and Rodriguez (2013).

RESULTS FROM THE FIELD TEST

The BLIS-3 assessment was administered in a large-scale field test to introductory statistics students at multiple institutions. There were 940 students from 34 different introductory statistics courses who completed the entire assessment in 10 to 120 minutes. A majority of the courses were in the United States however one was in Canada and one was in Spain. Analyses based on Classical Test Theory (CTT) and Item Response Theory (IRT) were used to provide evidence of score reliability and validity. The value of coefficient alpha was .83, which is above the recommended value of .8 for "very good" reliability (Kline, 2011).

In order to conduct analysis based on IRT, the assumptions of unidimensionality and local independence were examined using confirmatory factor analysis (CFA). The scree plot of eigenvalues showed evidence that the BLIS-3 assessment consisted of one factor. Also, each of the factor loadings were positive and model fit indices showed good fit for a single-factor model (see Table 1). Hu and Bentler (1999) recommended a cutoff value close to .95 or higher for the Tucker-Lewis Index (TLI) and Bentler's Comparative Fit Index (CFI), and a cutoff value close to .06 or lower for the Root Mean Square Error of Approximation (RMSEA).

The BLIS-3 assessment contained five testlets, which were pairs of items that included a common stem (Downing, 2006). One testlet included an item asking students to identify the null hypothesis for a particular research question and the other item in the testlet asked students to identify the alternative hypothesis statement. For all analysis, this testlet was scored as a 0 for one or both being incorrect, and a 1 for both being correct.

Table 1. Fit indexes for one-factor CFA models

Fit indices	No testlets	Testlets
CFI	0.944	0.952
TLI	0.961	0.968
RMSEA	0.027	0.027

In order to check if the remaining four testlets resulted in a violation of the local independence assumption, the tetrachoric correlation residuals were examined. The correlation residuals for the item pairs in testlets were not much different in magnitude compared to other item pairs. However, another single-factor CFA was run taking into account the items that were in testlets. Each testlet was scored as a 0, 1, or 2, where a score of 0 indicated that both items in the testlet were incorrect and a score of 2 indicated that both items in the testlet were correct. The model fit indices indicated that the single-factor CFA model fit slightly better when incorporating testlet scores (see Table 1). Therefore, it was decided that including testlet scores was acceptable to meet the local independence assumption.

The Partial Credit Model (PCM) based on IRT was fit to the assessment data. A parametric bootstrap approximation to the Pearson chi-squared Goodness-of-Fit measure provided evidence that the partial credit model had good fit (p = .27). See Table 2 for the item difficulty estimates.

Table 2. Item parameters for the partial credit model with 32 individual items and 4 testlets

Item/ testlet	Difficulty 1 (SE)	Difficulty 2 (SE)	Item/ testlet	Difficulty 1 (SE)	Difficulty 2 (SE)
1	-1.704 (0.094)		18, 19	-1.093 (0.096)	0.424 (0.086)
2	-0.009 (0.078)		20	0.046 (0.078)	
3	-0.341 (0.79)		21	1.619 (0.094)	
4	-2.438 (0.115)		22	-0.721 (0.080)	
5, 6	-3.058 (0.216)	-1.045 (0.085)	23, 24	-1.112 (0.095)	0.566 (0.087)
7	0.366 (0.079)		25, 26	-0.858 (0.092)	0.514 (0.088)
8	0.402 (0.079)		27	0.147 (0.078)	
9	-1.016 (0.083)		28	-1.159 (0.085)	
10	-1.291 (0.087)		29/30[a]	-0.326 (0.078)	
11	0.506 (0.080)		31	-1.099 (0.084)	
12	-0.588 (0.080)		32	-0.662 (0.080)	
13	0.238 (0.079)		33	-0.657 (0.080)	
14	0.081 (0.078)		34	-1.045 (0.084)	
15	-1.375 (0.088)		35	1.150 (0.086)	
16	0.591 (0.080)		36	-2.390 (0.114)	
17	-0.999 (0.083)		37	-0.688 (0.080)	

[a]Items 29 and 30 were combined to make one item score

DISCUSSION AND FUTURE PLANS

The reviews of the test blueprint provided evidence that the learning outcomes measured in the BLIS assessment are of importance to statistics educators. In addition, there is preliminary evidence that the items in the BLIS assessment measure the learning outcomes described in the test blueprint.

The evidence of reliability and validity of inferences for the BLIS-3 assessment are promising. The assessment appears to be measuring a single construct when incorporating testlet scores. The partial credit model has good model fit.

Future plans include examining data collected from instructors during the field test. The data from the instructors will determine the value of the BLIS assessment for instructors and researchers. More statistical analyses need to be conducted to examine the validity of inferences for the assessment. For example, student results from the field test will be used to examine if differential item functioning exists for any of the assessment items.

REFERENCES

Catalysts for Change (2012). *Statistical thinking: A simulation approach to modeling uncertainty*. Minneapolis, MN: CATALYST Press.

Chance, B. L., & Rossman, A. J. (2006). *Investigating statistical concepts, applications, and methods*. Cengage Learning.

Cobb, G. W. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1). Retrieved from: http://repositories.cdlib.org/uclastat/cts/tise/vol1/iss1/art1

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, Retrieved from http://www.stat.auckland.ac.nz/serj

Downing, S. M. (2006a). Selected-response item formats in test development. In Downing, S. M., & Haladyna, T. M. (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1–25.

Garfield, J., Aliaga, M., Cobb, G., Cuff, C., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, R., Utts, J., Velleman, P., & Witmer, J. (2007). Guidelines for assessment and instruction in statistics education (GAISE) project: College report. Retrieved from: http://www.amstat.org/education/gaise/

Garfield, J., delMas, R., & Chance, B. (2002). The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project. NSF CCLI grant ASA- 0206571. Retrieved from https://apps3.cehd.umn.edu/artist/index.html

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinking in an introductory, tertiary-level statistics course. *ZDM – The International Journal on Mathematics Education, 44*(7), 883-898. doi: 10.1007/s11858-012-0447-5

Garfield, J., delMas, R., & Zieffler, A. (2010). Assessing important learning outcomes in introductory tertiary statistics courses. In Bidgood, P., Hunt, N., & Jolliffe, F. (Eds.), *Assessment methods in statistical education* (pp. 75-86). John Wiley & Sons, Ltd.

Gould, R., & Ryan, C. N. (2013). *Introductory statistics: Exploring the world through data*. Pearson.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-333.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York: Routledge.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Kline, R. B. (2011). *Principles and practice of structural equation modeling (3rd Edition).* New York, NY: The Guilford Press.

Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2012). *Statistics: Unlocking the power of data.* Hoboken, NJ: Wiley.

McClave, J. T., & Sincich, T. (2007). *Statistics* (11th ed.). Upper Saddle River, NJ: Pearson.

Rossman, A. (2007). Seven challenges for the undergraduate statistics curriculum in 2007. *Slides at http://www.statlit.org/pdf/2007RossmanUSCOTS6up.pdf; handout at www.statlit.org/pdf/2007RossmanUSCOTS.pdf*

Schield, M. (1999). Statistical literacy: Thinking critically about statistics. *Of Significance, 1*(1), 15-20.

Tintle, N. Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2013). Introduction to statistical investigations. Retrieved March 15, 2013, from http://www.math.hope.edu/isi/

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).