# DEVELOPING A STATISTICAL LITERACY ASSESSMENT FOR THE MODERN INTRODUCTORY STATISTICS COURSE

LAURA ZIEGLER
*Iowa State University*
*lziegler@iastate.edu*

JOAN GARFIELD
*University of Minnesota*
*jbg@umn.edu*

**ABSTRACT**

*The purpose of this study was to develop the Basic Literacy In Statistics (BLIS) assessment for students in an introductory statistics course, at the postsecondary level, that includes, to some extent, simulation-based methods. The definition of statistical literacy used in the development of the assessment was the ability to read, understand, and communicate statistical information. Multiple instruments were available to assess students in introductory statistics courses; however, there were no assessments available that focused on statistical literacy. Evidence of reliability and validity were collected during the development of the assessment. Evidence of reliability and validity was high; however, more items with high difficulty levels could increase the precision in estimating ability estimates for higher achieving students.*

***Keywords:*** *Statistics education research; Simulation-based inference; Item response theory*

## 1. INTRODUCTION

Statistical literacy has been described as an important learning outcome in introductory statistics courses (e.g., Garfield, delMas, & Zieffler, 2010). The need for teaching statistical literacy was emphasized in a special issue of the Statistics Education Research Journal (May, 2017) which was dedicated to statistical literacy. Guest editors of the special issue, Ridgway and Nicholson, concluded by saying statistics educators must take the time to directly promote statistical literacy (2017). Despite the attention paid to statistical literacy in education today there is little consensus regarding the definition of this learning outcome (Ben-Zvi & Garfield, 2004; Schield, 2017). Definitions of statistical literacy range from the context of basic skills (Garfield, delMas, & Chance, 2002) to critical thinking (Gal, 2002). There is a need for a new assessment to help researchers and instructors explore students' statistical literacy. In the process of developing an assessment of statistical literacy, there was a need for a clear working definition of statistical literacy to guide the development of such an assessment. A literature review was conducted to create an appropriate working definition.

***Brief literature review of statistical literacy*** One of the first published definitions of statistical literacy was made by Walker (1951) who examined definitions of general literacy. Based on those definitions, she suggested that statistical literacy is the ability to communicate statistical information. Many definitions of statistical literacy seem to align with current definitions of general literacy. Literacy, in general, has been defined as the ability to read and write ("Literacy," n.d.a; "Literacy," n.d.b). Several statistics educators have described statistical literacy as understanding and using the basic language of statistics (Garfield et al., 2005; Garfield & delMas, 2010; Garfield et al., 2002; Lehohla, 2002). Chick and Pierce (2013) claimed that statistical literacy encompasses general literacy, numeracy, statistics, and data presentation which includes the ability to reason with information presented in graphs and tables.

In contrast, statistical literacy has been defined by others as including higher order skills such as communicating, interpreting, and being critical of statistical information (Gal, 2002; Schield, 1999; Smith, 2002). Wallman (1993) also incorporated the ability to understand and critically evaluate statistical information in the real world but added that a statistically literate citizen should be able to appreciate contributions that statistical thinking provides to make decisions. Tiers of statistical literacy have also been suggested: understanding of basic statistical problems and terminology, being able to use the basics in the real world, and questioning statistical conclusions and results (Watson, 2011).

There are other terms that could include aspects of statistical literacy such as quantitative literacy and numeracy. Whereas definitions of these terms often include some statistical topics, they usually focus primarily on mathematical learning outcomes. For example, numeracy has been described as possessing mathematical skills as well as modeling, interpreting, evaluating/analyzing, communicating, and understanding relationships, data, and chance (OECD, 2012).

For the purposes of this paper and assessment project, a working definition of statistical literacy was defined as the ability to read, understand, and communicate statistical information. The type of statistical information that is relevant for statistical literacy (e.g., graphical representations, descriptive statistics, inferential statistics) is encountered in daily life, such as in media articles, and involves real contexts. The working definition draws on the various definitions of statistical literacy that align with the definitions of general literacy.

Using the specified definition of statistical literacy, existing assessments were examined to determine whether or not they contained items assessing statistical literacy. Multiple assessments measure students understanding of statistics in an introductory statistics course at the postsecondary level: Statistics Reasoning Assessment (SRA; Garfield, 2003), Statistics Concepts Inventory (SCI; Allen, Stone, Rhoads, & Murphy, 2004), Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS; delMas, Garfield, Ooms, & Chance, 2007), Assessment Resource Tools for Improving Statistical Thinking Topic Scale tests (ARTIST; Garfield et al. 2002), Quantitative Reasoning Test-Version 9 (QR-9; Sundre, Thelk, & Wigtil, 2008), Goals and Outcomes Associated with Learning Statistics (GOALS; Garfield, delMas, & Zieffler 2012), and Reasoning about *P*-values and Statistical Significance (RPASS-10; Lane-Getaz, 2017). Each assessment contains some items measuring statistical literacy, but all of these assessments also measure higher order outcomes, such as being able to make connections and reason about statistics. All but one, the GOALS assessment, were designed for students in an introductory statistics course taught with normal-based methods, or what has been termed the *consensus curriculum* (Cobb, 2007).

In past studies, researchers have examined learning outcomes such as understanding basic terminology, interpreting data presented in tables and graphs, and understanding percentages and probabilities. Results from the studies were mixed; however, a majority of the studies demonstrated that students do not have a very high level of statistical literacy (Anderson, Gigerenzer, Parker, & Schulkin, 2014; Galesic & Garcia-Retamero, 2010; McLauchlan & Schonlau, 2016; Pierce & Chick, 2013; Schield, 2006; Turegun, 2011; Wade, 2009; Watson, 2011). There were studies that showed evidence that students struggled with being able to make interpretations of statistical results (Jones et al., 2000; Ridgway, Nicholson, McCusker, 2008; Yolcu, 2012). In two studies, researchers found students could successfully describe data (Jones et al., 2000; Sharma, Doyle, Shandil, & Talakia'atu, 2012), and students have been shown to be able to understand tables of percentages (Atkinson, Czaja, & Brewster, 2006). Unfortunately, the researchers of a majority of these studies did not take advantage of well-established assessments, such as the assessments mentioned previously. There was only one study where the researchers used an assessment (ARTIST Topic Scale) that was created for research purposes (Turegun, 2011). None of the studies included assessments that had evidence of validity or reliability. Overall, these studies provide some insight into students' statistical understanding, but there is more work to be done.

***Increased use of simulation based methods in introductory courses*** In recent years there has been much interest in the use of simulation-based methods in introductory statistics courses (e.g., Cobb, 2007; Rossman, 2007). It has been proposed that simulation-based methods are easily grasped (Cobb, 2007) and promote understanding (Hesterberg, Monaghan, Moore, Clipson, & Epstein, 2003). Simulation-based methods, such as randomization tests, are being taught in some introductory statistics courses in addition to or in lieu of normal-based methods, such as the *t*-test (e.g., Garfield et al., 2012;

Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011). Multiple new introductory statistics textbooks are being published that incorporate simulation-based methods (e.g., Catalysts for Change, 2013; Lock, Lock, Lock Morgan, Lock, & Lock, 2013).

More recently, researchers have been examining what statistical knowledge students gain in these courses (e.g., Garfield et al., 2012; Holcomb, Rossman, & Chance, 2011; Tintle, et al., 2011). The research has suggested that, when compared to students who complete a normal-based introductory statistics course, students in a simulation-based course have a better understanding of some statistical concepts, especially statistical inference. However, given the dearth in assessments for students in simulation-based introductory statistics courses, a majority of the studies examined used either instructor-made assessments or assessments designed for students in a normal-based introductory statistics course.

The only assessment available that was designed for research purposes in an introductory statistics course taught with simulation-based methods is the GOALS assessment, and this includes very few items that measure statistical literacy. New assessments need to be created to meet the needs of instructors teaching with these newer methods. Research involving students in simulation-based introductory statistics courses could benefit from an assessment of statistical literacy that has evidence of validity and reliability. In this paper, the development of a new assessment, The Basic Literacy In Statistics (BLIS) assessment, is described.

## 2. METHODS

The researchers created an assessment of statistical literacy that can be used to determine what statistical literacy skills students have in an introductory postsecondary statistics course that incorporates simulation-based methods. The content included in the assessment needed to be relevant to a variety of introductory statistics courses: courses that include only simulation-based methods, courses that emphasize simulation-based methods and parametric methods, and courses that focus on parametric methods but include simulation-based methods to help students gain a conceptual understanding of the content. A mixed-methods approach was used to collect evidence of reliability and validity for the BLIS assessment.

### 2.1. DEVELOPING A HIGH QUALITY ASSESSMENT

According to AERA, APA, and NCME (2014), reliability, validity, and fairness are characteristics of assessments that should be examined. This paper will examine the first two; reliability and validity. Definitions of these terms and how to examine them vary. In this sub-section, definitions of reliability and validity, as well as sources of low reliability and low validity, are described.

The reliability of an assessment refers to the consistency of scores when an individual repeatedly takes an assessment (AERA, APA, & NCME, 2014; Thorndike & Thorndike-Christ, 2010). In other words, if an assessment is administered to the same individuals multiple times, the results will be similar each time (Weathington, Cunningham, & Pittenger, 2010). Low reliability can be the result of measurement error. All assessments have some measurement error due to natural variability, but there are some sources of measurement error that can be minimized: instrument error, participant variability, researcher variability, and environmental variability (Weathington et al., 2010). Instrument error includes wording and organizational issues, participant variability includes fatigue and misunderstanding of items, researcher variability includes errors in recording, and environmental variability includes distractions and differences in testing locations. Evidence of reliability as it relates to sources of measurement error can be collected throughout the assessment development process. According to Buckendahl and Plake (2006), without examining the reliability evidence for an assessment, other evidence of validity may not be meaningful.

The definition of validity is not as clear as the definition of reliability because it refers to the interpretations of test scores and not the assessment itself (Weathington et al., 2010). Interpretations are affected by how much variability there is in participant responses, and therefore reliability is needed in order to have validity (Kane, 2013). A unified view of validity is described in the definition specified by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014); "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed

uses of tests" (p. 11). The unified view takes the perspective that there are not distinct types of validity and focuses on validity as related to interpretations of test scores and uses of test scores for applied purposes instead of the test itself (Messick, 1989). Kane (2013) emphasizes an argument-based approach should be used not only to validate interpretation of test scores but also to validate the use of test scores. Multiple pieces of evidence must be gathered and collated into one validity argument for the reasonableness of inferences and uses (Cook, Brydges, Ginsburg, & Hatala, 2015).

## 2.2. DEVELOPMENT OF THE BLIS ASSESSMENT

The development of BLIS required several phases, as outlined in Table 1. These phases will be briefly described.

*Table 1. Overview of assessment development, data collection, and analysis*

| Phase | Product and analysis | Data collected |
|---|---|---|
| Phase I | Preliminary test blueprint, evaluate expert review of preliminary test blueprint, BLIS Test Blueprint-1 | Review of textbooks, expert review of preliminary test blueprint |
| Phase III | BLIS-1, evaluate students' cognitive interview data | Cognitive interviews with students |
| Phase IV | BLIS-2, evaluate students' responses from pilot | Pilot test |
| Phase V | BLIS-3, instructor survey, factor analysis, reliability analysis, analysis based on item response theory | Field test |

***Phase I: Test blueprint development*** Prior to developing the assessment, a test blueprint was created outlining the set of topics to be included and number of items. The 26 topics of statistical literacy that were included were chosen to be of interest to instructors who included simulation-based methods in their introductory statistics course. The list of topics was created based on a review of introductory statistics textbooks (Gould & Ryan, 2013; Catalysts for Change, 2013; Lock et al., 2013; Tintle et al., 2013) that incorporated simulation-based methods. In light of the goal of the instrument, which was to assess students' statistical literacy abilities in a simulation-based introductory statistics course, topics that appeared to be more important (e.g., confidence intervals) were included and topics that were less important (e.g., probability rules) were not included. Topics were considered to be less important if they were not emphasized in two or more of the four textbooks that were reviewed.

Statistical literacy learning outcomes were specified for each topic. A total of 54 learning outcomes were devised based on the review of textbooks. Each learning outcome corresponded with one item. Multiple learning outcomes were related to being able to identify, describe, translate, interpret, read, and compute, which are words that have been associated with items measuring statistical literacy (Garfield et al., 2010). A majority of the learning outcomes focused on being able to describe and interpret. For example, one topic is *descriptive statistics*. An appropriate learning outcome would be that a student has the "ability to interpret a standard deviation in the context of data."

After the preliminary test blueprint was created, six statistics educators reviewed the preliminary test blueprint by rating and commenting on how important each learning outcome was in determining the statistical literacy of a student. The review was conducted to provide evidence of validity; that is, the learning outcomes specified in the preliminary test blueprint captured the intended construct: statistical literacy. Two reviewers were experts in the field of statistics education, two were statisticians who teach introductory statistics using simulation-based methods, and two reviewers were experts in statistics assessment development. All of the reviewers were authors of textbooks for introductory statistics students, and four of the reviewers emphasized simulation-based methods in their textbooks.

Based on reviewers' feedback, modifications were made to create the BLIS Test Blueprint-1. Learning outcomes with low ratings were examined in detail and were removed or modified. Suggested new learning outcomes were examined to see if they aligned with the definition of statistical literacy used to develop the assessment. Learning outcomes that did align and did not overlap with other learning outcomes already included in the test blueprint were added. After the 37 learning outcomes

were finalized and organized into nine topics, the BLIS Test Blueprint-1 was complete. See Table 2 for a summary of the test blueprint with example learning outcomes.

*Table 2. BLIS test blueprint-1 summary*

| Topic | Num of Learning Outcomes | Example Learning Outcome |
|---|---|---|
| Data production | 8 | Understanding of the difference between a sample and population |
| Graphs | 3 | Ability to describe and interpret a dotplot |
| Descriptive statistics | 5 | Ability to interpret a standard deviation in the context of the data |
| Empirical sampling distributions | 3 | Understanding that an empirical sampling distribution shows how sample statistics tend to vary |
| Confidence intervals | 3 | Understanding that a confidence interval provides plausible values of the population parameter |
| Randomization distributions | 3 | Understanding that sample statistics in the tails of a randomization distribution are evidence against the null hypothesis |
| Hypothesis tests | 8 | Understanding of the purpose of a hypothesis test |
| Scope of conclusions | 2 | Understanding of the factors that allow a sample of data to be generalized to the population |
| Regression and correlation | 2 | Ability to match a scatterplot to a verbal description of a bivariate relationship |

***Phases II-IV: Assessment development*** The BLIS assessment was developed using an iterative process as recommended by Kane (2013). In Phase II, the preliminary version of the assessment was written using items from existing instruments as well as items developed specifically for the BLIS assessment. Nineteen existing selected-response items measuring statistical literacy were chosen from the CAOS test (delMas et al., 2007), ARTIST Topic Scale tests (Garfield et al., 2002), ARTIST item database (Garfield et al., 2002), and an early version of the GOALS assessment (Garfield et al., 2012). For the 18 learning outcomes that did not have existing items available, new items were created using item writing recommendations from the *Standards* (AERA, APA, & NCME, 1999) and Haladyna, Downing, and Rodriguez (2002). The new items were a combination of selected-response items and constructed-response items in the preliminary version of the assessment. The reason for starting with constructed-response items in early versions of the assessment was to use student responses in the pilot of the assessment to discover plausible incorrect answers to make meaningful distractors as recommended by Haladyna et al. (2002), Garfield and Franklin (2011), and Thorndike and Thorndike-Christ (2010).

Several important considerations were made when choosing and creating items. First, items included a real-world context as recommended by Gal (1998) and Garfield et al. (2005). Secondly, in order to create items to measure statistical literacy, the wording was carefully chosen to ensure the primary outcome being assessed was statistical literacy. Key words that could be used to assess statistical literacy were provided by Garfield et al. (2010) and were based on the key words mentioned by delMas (2002) as well as Garfield, delMas, and Chance (2003). Key words to include when assessing statistical literacy were: *identify*, *describe*, *translate*, *interpret*, *read*, and *compute*. These words were used not only to help determine which items in existing assessments measure statistical literacy, but also to create new items. More emphasis was placed on descriptions and interpretations and less emphasis was placed on computing because many courses have shifted focus away from computations to interpretations (Chance, 1997).

To create the BLIS-1 assessment, the preliminary assessment was reviewed by the same six reviewers who examined the test blueprint to provide further evidence of validity. For each item, reviewers were asked how much they agreed or disagreed with the following statement: "The assessment item measures the specified learning outcome." They rated their agreement on a 4-point scale: 1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, and 4 = *strongly agree*. Reviewers were also given

the option to provide additional comments improving the assessment. Modifications were made to reflect the suggestions of the reviewers, and this resulted in the BLIS-1 assessment.

In Phase III, six students completed the BLIS-1 assessment in cognitive interviews which were used to develop the BLIS-2 assessment. The students were from two universities in four courses; all of which were taught simulation methods. The students who participated in the interviews varied in statistical ability and background. Students were asked to talk about what they are doing and thinking while taking the BLIS-1 assessment. Students' responses were used to make additional changes to the assessment items in order to create the BLIS-2 assessment.

The pilot test was conducted in Phase IV with the BLIS-2 assessment to see whether the selected-response options were all viable options and to develop selected-response options for the constructed-response items. There were 76 students enrolled in three introductory statistics courses who took the assessment. The introductory statistics courses included a graduate-level course taught with the *Statistics: Unlocking the Power of Data* (Lock et al., 2013) textbook at the University of Minnesota, an undergraduate-level course taught with the CATALST curriculum (Garfield, et al., 2012) at the University of Minnesota, and an undergraduate-level course taught with *Statistics* (McClave & Sincich, 2007) at Winona State University.

The students' results from the pilot test were examined to determine what changes needed to be made to the assessment. First, the 16 selected-response items were examined to see if the distractors were chosen by students (see Table 3). All distractors were chosen by at least one student for each item except for Item 4. In Item 4, the question described a study and the student was asked to select which type of study was conducted: observational, experimental, or survey. No students chose the survey option. It was decided that the option should still be kept because one of the reviewers emphasized that observational studies and experiments were not the only study designs, and that surveys should be included.

*Table 3. Percentage of students (n = 76) who chose each selected-response option for the 16 pilot test items*

| Item | a | b | c | d | no response |
|---|---|---|---|---|---|
| 3 | 59.2* | 15.8 | 7.9 | 17.1 | 0.0 |
| 4 | 7.9 | 92.1* | 0.0 | | 0.0 |
| 5 | 82.9* | 14.5 | 2.6 | | 0.0 |
| 6 | 1.3 | 19.7 | 67.1* | 11.8 | 0.0 |
| 10 | 6.6 | 7.9 | 11.8 | 72.3* | 1.3 |
| 14 | 9.2 | 51.3* | 35.5 | 1.3 | 2.6 |
| 15 | 7.9 | 10.5 | 18.4 | 60.5* | 2.6 |
| 16 | 21.1* | 22.4 | 27.6 | 26.3 | 2.6 |
| 21 | 7.9* | 6.6 | 52.6 | 31.6 | 1.3 |
| 22 | 28.9 | 51.3* | 18.4 | | 1.3 |
| 27 | 43.4 | 42.1* | 3.9 | 9.2 | 1.3 |
| 31 | 13.2 | 81.6* | 3.9 | | 1.3 |
| 34 | 34.2 | 61.8* | 2.6 | | 1.3 |
| 35 | 17.1* | 11.8 | 7.9 | 61.8 | 1.3 |
| 36 | 90.8* | 5.3 | 2.6 | | 1.3 |
| 37 | 25.0 | 11.8 | 59.2* | 2.6 | 1.3 |

*Notes*. Items with no results presented for selected-response Option D represent an item that did not have an Option D. * indicates correct answer.

There were multiple items that had a relatively low percentage of students who chose the correct answer in the pilot test. Considering the pilot was administered half-way through the semester, instructors were questioned to see if students had learned about the statistical content included in the low scoring items. Items 16 and 21 included content that students had not been taught yet, so they were not changed. For Item 35, students did better in the graduate-level course compared with students in the undergraduate-level courses. The item was taken from the CAOS assessment, and in an analysis conducted by delMas et al. (2007), only 37.9% ($n$ = 715) of students answered the item correctly on a posttest, which was comparable to students in the graduate-level course in this study. An updated

version of the item was included in an early version of the GOALS assessment. It was decided to include this updated version of the item in the BLIS-3 assessment.

For the 21 constructed-response items, student responses were examined to convert the items to selected-response for the third version of the BLIS assessment (BLIS-3) as well as to make additional changes. In order to create the selected-response options, students' responses were grouped by similar responses. The incorrect responses that were submitted most often and appeared to measure misconceptions related to the learning outcomes were included as distractors. One item was deleted after it appeared to not measure the intended learning outcome: understanding that every model is based on assumptions that limit our scope of inferences. It was decided not to replace the deleted item because a similar learning outcome was already being assessed: understanding of the factors that allow a sample of data to be generalized to the population. The last change involved splitting one item into two items. In the question, students were given a research question and were asked what the null hypothesis and alternative hypothesis were. The item was split to create a testlet where the first question asked students to select the correct null hypothesis statement and the second question asked students to select the correct alternative hypothesis statement.

## 2.3. PHASE V: FIELD TEST ADMINISTRATION

Instructors who taught an introductory statistics course at the college-level, which included Advanced Placement Statistics courses, were recruited to administer the BLIS-3 assessment to their students. Characteristics of the courses students were enrolled in are shown in Table 4. All courses, except for two, were in the United States and most incorporated simulations in the curricula. Most students took the assessment out of class and received extra credit for completing the assessment.

*Table 4. Course characteristics as reported by instructors (n = 34 courses)*

| Characteristic | Count | % | Characteristic | Count | % |
|---|---|---|---|---|---|
| Country | | | Credit[b] | | |
| United States | 32 | 94.1 | Assignment | 5 | 16.1 |
| Canada | 1 | 2.9 | Extra credit | 23 | 74.2 |
| Spain | 1 | 2.9 | No credit | 3 | 9.7 |
| Type of institution[a] | | | Time of semester[b] | | |
| High school | 3 | 9.1 | Beginning | 0 | 0.0 |
| 2-year/technical college | 6 | 18.2 | Middle | 6 | 19.4 |
| 4-year college | 12 | 36.4 | End | 25 | 80.6 |
| University | 12 | 36.4 | Simulation methods[bc] | | |
| Prerequisites[bc] | | | Bootstrapping | 10 | 32.3 |
| No mathematics | 4 | 12.9 | Randomization tests | 14 | 45.2 |
| High school algebra | 16 | 51.6 | Probability simulations | 24 | 77.4 |
| College algebra | 10 | 32.3 | Other simulations | 23 | 74.2 |
| Calculus | 3 | 9.7 | None of the above | 4 | 12.9 |
| Setting[d] | | | | | |
| In class | 5 | 16.7 | | | |
| Out of class | 25 | 83.3 | | | |

*Note*. Ten students completed the assessment but did not provide their instructor code so their results are not included in this table.

[a]One instructor did not respond. [b]Three instructors did not respond. [c]Percentages do not add up to 100 because instructors could choose more than one option. [d]Four instructors did not respond.

Student demographics from the field test are included in Table 5. A majority of students (87.9%) was under the age of 25 and most students (87.4%) were in college.

*Table 5. Demographic characteristics of students who took BLIS-3 in the field test*

| Characteristic | *n* | % | Characteristic | *n* | % |
|---|---|---|---|---|---|
| Gender | | | Race | | |
| Female | 533 | 58.3 | White | 710 | 82.8 |
| Male | 382 | 41.7 | Black or African American | 24 | 2.8 |
| Age | | | Amer. Indian or Alaska Native | 7 | 0.8 |
| ≤ 19 | 364 | 39.9 | Asian | 66 | 7.7 |
| 20-24 | 438 | 48.0 | Pacific Islander | 6 | 0.7 |
| 25-29 | 55 | 6.0 | Other | 44 | 5.1 |
| 30-34 | 21 | 2.3 | Class | | |
| 35-39 | 15 | 1.6 | High school | 114 | 12.6 |
| 40-44 | 9 | 1.0 | Freshman/first year | 190 | 21.1 |
| 45-49 | 5 | 0.5 | Sophomore | 291 | 32.3 |
| 50-54 | 3 | 0.3 | Junior | 166 | 18.4 |
| 55+ | 2 | 0.2 | Senior | 90 | 10.0 |
| International or foreign national student | | | Graduate student | 31 | 3.4 |
| Yes | 44 | 4.8 | Other | 20 | 2.2 |
| No | 868 | 95.2 | | | |

***Descriptive statistics for the BLIS-3 assessment*** Descriptive statistics were computed for the BLIS-3 assessment. First, the percentages of students who chose each selected-response option were computed. The percentages were examined to see if any of the items had a higher percentage of students who selected a particular distractor than the correct option. Then, total scores were computed and a plot of students' total scores was created.

***Collecting further evidence of the reliability of the BLIS assessment*** To check for score reliability, analyses based on CTT and IRT were used. Coefficient alpha, based on CTT, was computed because it is a measure of internal consistency used when the items have varying difficulty levels and provides reliability evidence for the raw scores of the assessment (Thorndike & Thorndike-Christ, 2010).

The BLIS assessment has five testlets, each with two items. Testlets are items that share a common stem (Downing, 2006). Coefficient alpha was computed using testlet scores. 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.

To check for reliability at different ability levels and at the item level, analysis based on IRT was conducted. The required assumptions needed to conduct analyses based on IRT were checked as outlined by Raykov and Marcoulides (2010). Unidimensionality was assessed using a single-factor confirmatory factor analysis (CFA).

The generalized partial credit (GPC) model (Muraki, 1992) was chosen to model the BLIS data. Initially, the Rasch model, 2 parameter logistic model, partial credit model, and GPC model were compared. The Rasch model, 2 parameter logistic model, and partial credit model were fit with the approximate marginal Maximum Likelihood using the *ltm* package in R (Rizopoulos, 2006). The GPC model was fit with the standard EM algorithm with fixed quadrature using the *mirt* package in R (Chalmers, 2012). Model fit indices including AIC, BIC, and log likelihood provided evidence that the GPC model had the best fit.

The GPC model incorporates testlet scores in the analysis and provides a person ability parameter, $\theta$, and a difficulty parameter, $b$. For the GPC model, the probability of a randomly chosen person with a particular ability level, $\theta$, scoring a 0, 1, or 2 on a particular item or testlet $j$ is given by

$$\text{GPC: } P_{jk}(\theta) = \frac{exp\sum_{v=1}^{k} a_j(\theta - b_{jv})}{\sum_{c=1}^{m_j} \exp[\sum_{v=1}^{c} a_j(\theta - b_{jv})}$$

where $a_j$ is the slope parameter for each item or testlet, $b_{jk} = 0, j$ is the item or testlet number ($j$ = 0, 1, …, $n$), and $k$ is the category ($k$ = 0, 1, …, $m$).

***Collecting further evidence of the validity of the BLIS assessment*** CFA and analyses based on IRT were used to measure validity. CFA has been used to determine whether or not an assessment measures one underlying construct (Thorndike & Thorndike-Christ, 2010). Therefore, CFA was used in this study to provide evidence that the assessment is measuring the one intended construct, statistical literacy. Next, analyses based on IRT were conducted to determine the extent to which the assessment had internal validity (Thorndike & Thorndike-Christ, 2010). Item difficulties were calculated to display whether or not the assessment included items of varying difficulty levels.

## 3. RESULTS

The results from the analyses of the data collected during the large-scale field test are presented in this section.

### 3.1. RESULTS FROM FIELD TEST

***Descriptive statistics*** First, the percentage of students, out of 940, who chose each selected-response option for each item, was computed (see Table 6). All items, except for Items 21 and 35, had the highest percentage of students choose the correct option. Both Items 21 and 35 had a low percentage correct on the pilot test as well as the field test.

*Table 6. Percentage of students (n = 940) who chose each selected-response option for all 37 items administered in the field test*

| Item | a | b | c | d | Item | a | b | c | d |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 11.7 | 6.9 | 81.4* | | 20 | 16.2 | 27.4 | 48.7* | 7.7 |
| 2 | 6.2 | 7.9 | 49.9* | 36.1 | 21 | 19.9* | 4.3 | 39.0 | 36.8 |
| 3 | 56.9* | 16.9 | 3.6 | 22.6 | 22 | 27.3 | 64.7* | 8.0 | |
| 4 | 9.9 | 89.6* | 0.5 | | 23 | 10.4 | 23.2 | 49.3* | 17.1 |
| 5 | 90.0* | 9.1 | 0.9 | | 24 | 61.3* | 14.3 | 24.5 | |
| 6 | 1.1 | 12.4 | 76.8* | 9.7 | 25 | 53.4* | 21.6 | 25.0 | |
| 7 | 32.2 | 14.7 | 11.1 | 42.0* | 26 | 53.3* | 35.0 | 11.7 | |
| 8 | 40.0 | 18.7 | 41.3* | | 27 | 41.3 | 46.6* | 5.9 | 6.3 |
| 9 | 70.3* | 26.4 | 3.3 | | 28 | 9.7 | 17.4 | 72.9* | |
| 10 | 9.3 | 7.3 | 8.3 | 75.1* | 29 | 12.7 | 10.3 | 64.3* | 12.8 |
| 11 | 23.9 | 36.9 | 39.1* | | 30 | 5.9 | 17.4 | 15.4 | 61.3* |
| 12 | 13.9 | 21.1 | 3.0 | 62.0* | 31 | 20.6 | 71.8* | 7.6 | |
| 13 | 44.7* | 31.4 | 6.2 | 17.8 | 32 | 12.7 | 12.6 | 63.5* | 11.3 |
| 14 | 4.0 | 48.0* | 47.0 | 1.0 | 33 | 63.4* | 27.3 | 9.3 | |
| 15 | 5.1 | 4.4 | 14.0 | 76.5* | 34 | 25.9 | 70.9* | 3.3 | |
| 16 | 37.4* | 15.3 | 27.0 | 20.2 | 35 | 27.1* | 11.4 | 11.5 | 50.0 |
| 17 | 17.7 | 12.3 | 70.0* | | 36 | 89.1* | 8.5 | 2.3 | |
| 18 | 27.4 | 53.0* | 19.6 | | 37 | 19.7 | 11.1 | 64.0* | 5.2 |
| 19 | 4.7 | 18.2 | 16.8 | 60.3* | | | | | |

*Notes.* Only students who completed the entire assessment are included in this table. Items with no results presented for selected-response Option D represent an item that did not have an Option D.
* indicates correct answer.

Recall the BLIS-3 assessment contained five testlets. One testlet included two items, Items 29 and 30, which shared the same learning outcome: ability to determine a null and alternative hypothesis statement based on a research question. Item 29 asked students to provide the null hypothesis for a particular research question and Item 30 asked students to provide the alternative hypothesis for the same research question. A majority of students either answered both items correctly or both items incorrectly (87.7%). Therefore, it was decided that these items would be scored together as one item (Item 29/30) for the remainder of the analyses. Students who answered one or both items incorrectly received a score of 0 and students who answered both items correctly received a score of 1. Partial credit was not given if students answered only one item correctly.

The other four testlets had item pairs with different learning outcomes. Testlets have been claimed to create local dependence for items within a testlet (Downing, 2006). Therefore, it was decided to score the items together. 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.

Students' total scores were examined for the 36 items. The average total score was 21.41 out of 36 ($s$ = 6.25, $N$ = 940), or 59.5%. See Figure 1 for a visual representation of students' total scores for the 36 items.

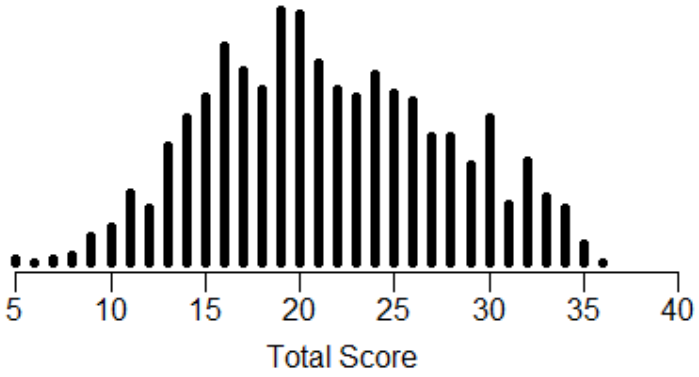


*Figure 1. Dotplot of students' total scores for the 36 items on the BLIS-3 assessment*

***Reliability*** Coefficient alpha was computed to check for the overall reliability of the BLIS-3 assessment at the test level. The value of coefficient alpha was 0.83, which is above the recommended value of 0.8 for "very good" reliability (Kline, 2011).

In order to conduct analyses based on IRT, the assumption of unidimensionality was examined using CFA. The scree plot of eigenvalues was examined for the BLIS-3 assessment that consisted of 32 item and testlet scores. The scree plot of eigenvalues showed evidence that the BLIS-3 assessment consisted of one factor (see Figure 2).

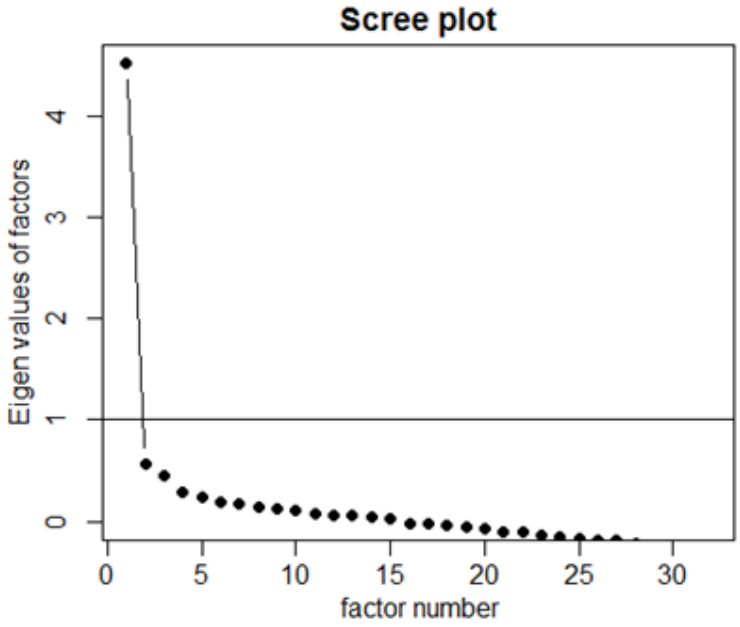


*Figure 2. Scree plot of eigenvalues for the BLIS-3 assessment*

Further evidence of unidimensionality was collected by examining the factor loadings and model fit indices. Each of the factor loadings were positive, indicating good fit for a single-factor model (see Table 7).

*Table 7. Factor loadings for one-factor CFA model with 32 individual items and four testlets*

| Item/ testlet | Loading | SE | *p*-value | Item/ testlet | Loading | SE | *p*-value |
|---|---|---|---|---|---|---|---|
| 1 | 0.476 | 0.037 | 0.000 | 18/19 | 0.603 | 0.028 | 0.000 |
| 2 | 0.316 | 0.041 | 0.000 | 20 | 0.605 | 0.034 | 0.000 |
| 3 | 0.441 | 0.038 | 0.000 | 21 | 0.378 | 0.057 | 0.000 |
| 4 | 0.341 | 0.050 | 0.000 | 22 | 0.296 | 0.041 | 0.000 |
| 5/6 | 0.425 | 0.037 | 0.000 | 23/24 | 0.582 | 0.030 | 0.000 |
| 7 | 0.570 | 0.036 | 0.000 | 25/26 | 0.559 | 0.029 | 0.000 |
| 8 | 0.086 | 0.045 | 0.054 | 27 | 0.492 | 0.039 | 0.000 |
| 9 | 0.350 | 0.041 | 0.000 | 28 | 0.549 | 0.037 | 0.000 |
| 10 | 0.387 | 0.041 | 0.000 | 29/30[a] | 0.780 | 0.025 | 0.000 |
| 11 | 0.205 | 0.044 | 0.000 | 31 | 0.498 | 0.036 | 0.000 |
| 12 | 0.187 | 0.043 | 0.000 | 32 | 0.476 | 0.036 | 0.000 |
| 13 | 0.355 | 0.041 | 0.000 | 33 | 0.551 | 0.035 | 0.000 |
| 14 | 0.283 | 0.042 | 0.000 | 34 | 0.493 | 0.035 | 0.000 |
| 15 | 0.510 | 0.037 | 0.000 | 35 | 0.556 | 0.043 | 0.000 |
| 16 | 0.602 | 0.039 | 0.000 | 36 | 0.593 | 0.037 | 0.000 |
| 17 | 0.583 | 0.033 | 0.000 | 37 | 0.554 | 0.034 | 0.000 |

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

Model fit indices examined indicated good fit (see Table 8). Fit indices referenced include: the Tucker-Lewis Index (TLI), Bentler's Comparative Fit Index (CFI), and the Root Mean Square Error of Approximation (RMSEA). Hu and Bentler (1999) recommended a cutoff value close to 0.95 or higher for the TLI and CFI, and a cutoff value close to 0.06 or lower for the RMSEA.

*Table 8. Fit indices for one-factor CFA model*

| Fit indices | |
|---|---|
| CFI | 0.952 |
| TLI | 0.968 |
| RMSEA | 0.027 |

A majority of the items had good item fit (see Table 9). The signed chi-square statistic ($S\text{-}X^2$; Orlando & Thissen, 2000) was computed for each item. Items 11 and 21 displayed poor fit and Items 3 and 27 were somewhat poor.

*Table 9. $S\text{-}X^2$ for the GPC model with 32 individual items and four testlets*

| Item/ testlet | $S\text{-}X^2$ ($p$) | Item/ testlet | $S\text{-}X^2$ ($p$) | Item/ testlet | $S\text{-}X^2$ ($p$) |
|---|---|---|---|---|---|
| 1 | 15.089 (0.891) | 13 | 28.242 (0.297) | 27 | 39.232 (0.019) |
| 2 | 31.013 (0.189) | 14 | 16.501 (0.899) | 28 | 23.006 (0.401) |
| 3 | 38.407 (0.031) | 15 | 28.387 (0.163) | 29/30[a] | 9.301 (0.952) |
| 4 | 19.499 (0.553) | 16 | 21.566 (0.163) | 31 | 23.394 (0.380) |
| 5, 6 | 28.011 (0.755) | 17 | 20.344 (0.500) | 32 | 24.561 (0.430) |
| 7 | 26.751 (0.221) | 18, 19 | 39.579 (0.356) | 33 | 29.712 (0.098) |
| 8 | 13.532 (0.979) | 20 | 34.44 (0.044) | 34 | 20.997 (0.521) |
| 9 | 25.406 (0.440) | 21 | 74.874 (0.000) | 35 | 23.581 (0.314) |
| 10 | 24.285 (0.388) | 22 | 36.004 (0.072) | 36 | 15.745 (0.610) |
| 11 | 53.062 (0.001) | 23, 24 | 37.226 (0.505) | 37 | 17.397 (0.687) |
| 12 | 30.571 (0.245) | 25, 26 | 42.877 (0.270) | | |

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

Person fit was examined by using the standardized statistic Zh (Drasgow, Levine, & Williams, 1985). Out of 940 students, only fifteen students had poor fit with a Zh-value less then negative two, the smallest being -2.62. There were four students who had overfit with a Zh-value greater than two, the largest being 2.28.

***Validity*** Parameter estimates and standard errors were examined to collect further evidence of validity (see Table 10). Item difficulties ranged from the least difficult at -3.69 to the most difficult at 2.32 with more items with lower difficulties than items with higher difficulties. In order to interpret the slope parameter estimates, Baker's (2001) proposed categorization of discrimination was referenced. Baker indicated that items with slope parameter estimates less than 0.65 have low to very low discrimination, between 0.65 and 1.34 have moderate discrimination, and above 1.34 have high to very high discrimination. Based on that recommendation, a majority of the slope parameter estimates have moderate to very high discrimination; however, there are seven items with low to very low discrimination. One of those seven items is the item with the highest difficulty, and because there are very few items with high difficulty, this item should be included in the assessment. The other six of the seven items should be examined in more detail to determine how useful they are in the BLIS assessment.

*Table 10. Item parameters for the GPC model with 32 individual items and four testlets*

| Item/ testlet | Slope parameter (SE) | Difficulty 1 (SE) | Difficulty 2 (SE) | Item/ testlet | Slope parameter (SE) | Difficulty 1 (SE) | Difficulty 2 (SE) |
|---|---|---|---|---|---|---|---|
| 1 | 1.08 (0.14) | -1.66 (0.17) | | 18, 19 | 1.06 (0.09) | -1.13 (0.10) | 0.44 (0.09) |
| 2 | 0.54 (0.08) | -0.001 (0.13) | | 20 | 1.26 (0.11) | 0.02 (0.07) | |
| 3 | 0.82 (0.09) | -0.41 (0.10) | | 21 | 0.73 (0.10) | 2.11 (0.26) | |
| 4 | 0.77 (0.15) | -3.08 (0.51) | | 22 | 0.54 (0.09) | -1.20 (0.21) | |
| 5, 6 | 0.79 (0.10) | -3.69 (0.41) | -1.30 (0.16) | 23, 24 | 1.01 (0.09) | -1.17 (0.11) | 0.59 (0.09) |
| 7 | 1.11 (0.10) | 0.33 (0.08) | | 25, 26 | 0.90 (0.08) | -0.95 (0.11) | 0.56 (0.10) |
| 8 | 0.15 (0.07) | 2.32 (1.18) | | 27 | 0.93 (0.10) | 0.15 (0.09) | |
| 9 | 0.66 (0.10) | -1.43 (0.21) | | 28 | 1.23 (0.13) | -1.05 (0.10) | |
| 10 | 0.78 (0.11) | -1.60 (0.20) | | 29/30[a] | 2.23 (0.20) | -0.27 (0.05) | |
| 11 | 0.35 (0.08) | 1.30 (0.33) | | 31 | 1.07 (0.12) | -1.08 (0.11) | |
| 12 | 0.32 (0.08) | -1.58 (0.42) | | 32 | 0.94 (0.10) | -0.72 (0.10) | |
| 13 | 0.62 (0.08) | 0.36 (0.12) | | 33 | 1.16 (0.12) | -0.62 (0.08) | |
| 14 | 0.48 (0.08) | 0.17 (0.15) | | 34 | 1.05 (0.12) | -1.05 (0.11) | |
| 15 | 1.14 (0.13) | -1.30 (0.13) | | 35 | 1.07 (0.11) | 1.11 (0.11) | |
| 16 | 1.18 (0.11) | 0.53 (0.08) | | 36 | 1.78 (0.23) | -1.69 (0.14) | |
| 17 | 1.35 (0.13) | -0.86 (0.08) | | 37 | 1.19 (0.12) | -0.65 (0.08) | |

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

## 4. DISCUSSION

The purpose of this study was to develop a new assessment of statistical literacy (called the BLIS assessment) to be used in an introductory statistics course that incorporates, to some extent, simulation-based methods. Evidence of reliability and validity of the BLIS assessment are discussed here.

### 4.1. RELIABILITY

Reliability evidence was collected throughout the development of the BLIS assessment. Considering the precautions taken in the wording of items and the results from the statistical analysis, the BLIS-3 assessment appears to have high reliability. Many changes in the wording of items on the preliminary version of the assessment were made based on the expert reviews, and in each of the following versions of the assessment, fewer changes were needed based on student data. Only a handful of items in the BLIS-2 assessment required changes in wording. This suggests that the wording of the items in the BLIS-3 assessment is of high quality. In the end, coefficient alpha was high (0.83) indicating good reliability.

Examining the item information functions, test information function, and standard error of measurement function from the GPC model, it appears that the precision in estimating students' abilities is highest when their abilities are closest to 0. Also, precision is higher for students with lower ability levels than students with higher ability levels because there are more items with low difficulty levels than items with high difficulty levels. As a result, reliability may be lower for students with higher abilities.

### 4.2. VALIDITY

Multiple pieces of evidence were gathered to create a validity argument for the reasonableness of inferences. The BLIS assessment was developed to make inferences about students' statistical literacy in an introductory statistics course that incorporates, to some extent, simulation-based methods. Validity evidence was collected through expert reviews, cognitive interviews with students, a small-scale pilot test, and a large-scale field test.

The first consideration in collecting evidence of validity was to ensure the BLIS assessment was measuring important aspects of statistical literacy by creating a test blueprint. The preliminary test blueprint was based on textbooks used in introductory statistics courses that incorporated simulation-based methods (Gould & Ryan, 2013; Catalysts for Change, 2013; Lock et al., 2013; Tintle et al., 2013), which was then reviewed by six experts. The reviewers rated the importance of each learning outcome for the assessment. The learning outcomes that were rated highest were included in the final version of the BLIS test blueprint. Furthermore, the reviewers provided feedback on what they felt was missing from the test blueprint, and this resulted in four new learning outcomes.

Multiple steps were taken to provide evidence that the BLIS assessment was measuring one construct, statistical literacy. The items designed to measure the learning outcomes in the test blueprint were reviewed by the same six experts. There was only one learning outcome that was identified as not measuring the intended learning outcome of statistical literacy, so a new item was developed. For the remaining items, reviewer feedback was used to refine the items to increase the validity evidence that the assessment was measuring statistical literacy.

The cognitive interviews and pilot test provided additional evidence that the assessment was measuring statistical literacy rather than measuring irrelevant content. Items were modified if students' responses did not appear to match the intended learning outcomes. One item was deleted after the pilot test was conducted because it was clear that the item was not measuring the intended learning outcome.

Data collected in the field test was used to conduct a single-factor CFA to present evidence that the BLIS assessment measures one construct. The scree plot of eigenvalues and fit indices, including the TLI, CFI, and RMSEA, showed evidence of a single factor. Considering the results from the expert reviews, cognitive interviews, and pilot test, it can be inferred that there is validity because the BLIS assessment measures one factor, statistical literacy.

An analysis based on IRT was conducted to see the extent to which validity exists for students of different ability levels. Using the GPC model, it was discovered that there are more items with low difficulty levels than items with high difficulty levels, meaning there is more evidence of validity when examining students with low ability levels compared with students with high ability levels. Results suggested that inferences made from the BLIS assessment have reasonably high validity.

### 4.3. LIMITATIONS

The results from this study showed evidence of reasonably high reliability and validity of the BLIS assessment; however, there are limitations to the claims that can be made. Not all sources of measurement error that can affect reliability were examined. Participant variability can include test fatigue and lack of test taker motivation (Weathington et al., 2010), and these factors were not examined in this study. In the field test, most students received credit for completing the assessment; however, when communicating with the instructors, some students were given credit for completion, regardless of how well they did on the assessment. Receiving credit only for completion could affect students' motivation to try to do well on the assessment. Environmental variability is another source of measurement error that was not taken into account. A majority (83.3%) of students took the assessment outside of class suggesting there could be measurement error related to environmental issues.

The methods used to collect the data could affect generalizability. Students who participated were not selected randomly. Further, data was collected during the middle and end of the semester indicating that validity evidence does not exist for using the BLIS assessment as a pretest. According to Thorndike and Thorndike-Christ (2010), a separate validity argument is needed for a pretest and posttest because the inferences to be made for each type of test are different. Therefore, if it is of interest to use the BLIS assessment as a pretest, data needs to be collected from students at the beginning of an introductory statistics course to collect evidence of validity.

### 4.4. IMPLICATIONS FOR TEACHING

The BLIS assessment could be useful for evaluating student learning in a wide range of introductory statistics courses. The field test data included responses from students in courses that had simulation in the curriculum throughout the semester as well as students in courses that had very little simulation. Based on the reliability and validity evidence collected, the assessment could ideally be used towards the end of a course. The results could inform instructors about which topics their students struggle with and which topics they understand. Instructors could use the results to help their students as well as to make changes in their future courses.

Introductory statistics instructors could use the results from the BLIS assessment to guide them to determine topics that are more difficult for students and to evaluate the effectiveness of teaching or curriculum change to improve student understanding of these topics. For example, there were two items that students performed poorly on in the field test, Items 21 and 35. Students chose one of the distractors in these items more frequently than those who chose the correct option. For Item 21, the learning outcome was: understanding that a confidence interval for a proportion is centered at the sample statistic. Students were asked to interpret the center of the confidence interval and approximately 75% of students chose an interpretation that included being 95% confident. Item 21 was also the item with the highest $S\text{-}X^2$ indicating the item had poor fit. The learning outcome for Item 35 was: understanding of the factors that allow a sample of data to be generalized to the population. Results from Item 35 suggested that some students incorrectly believed that a random sample of size 500 was too small to make a generalization from a sample to a population when estimating a proportion.

### 4.5. IMPLICATIONS FOR FUTURE RESEARCH

New items should be developed for the BLIS assessment for three reasons. First, recall that one item was deleted after the pilot test because it was determined that the item was not measuring the intended learning outcome: understanding that every model is based on assumptions which limit our scope of inferences. This was a learning outcome that was suggested by one of the expert reviewers, so a new item should be written for a future version of the assessment. The second reason why new items are needed is because there are more items with lower difficulty levels than items with higher difficulty levels. More difficult items are needed to increase the precision in estimating ability levels for higher achieving students. Lastly, there were a few items that had poor fit with the GPC model. These items should be examined in more detail to determine why they had poor fit.

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) recommended that three foundational characteristics of an assessment should be examined; validity, reliability, and fairness. Future research should look at fairness of the BLIS assessment. For example, Differential Item Functioning (DIF) has not been examined. Student demographic information collected in the field test could be used to examine if DIF exists for any of the items. Items could be examined to see if males and females perform differently. Race, language, and age are other variables that could be examined.

Item and person fit should be examined in more detail. The items with poor fit should be investigated to see why they had poor fit. Cognitive interviews could be conducted to better understand the poorly fit items. The demographics of the students with bad person fit should be studied to see if they have any demographic characteristics in common or to see if there is anything that makes those students standout compared to the students who did not have poor fit.

The data collected from instructors who administered the BLIS-3 assessment to their students in the field test could be used to conduct additional analyses. In particular, analysis could be conducted to

examine how the extent to which simulation was incorporated in the courses was related to the performance of students on the assessment. The invitation that was sent to the instructors asked for participants who "teach an introductory statistics course at the postsecondary level that includes simulation, to some extent, in the curriculum." However, in the survey, three instructors said they did not use bootstrap confidence intervals, randomization tests, probability simulations, or any other simulations to help understand statistical topics. More investigation could be conducted to understand the relationship between the amount of simulation included in the curriculum and students' responses on the BLIS assessment.

The BLIS assessment provides a valuable addition to the statistics education community considering the psychometric properties examined. Overall, future studies would benefit from the use of the BLIS assessment.

## ACKNOWLEDGEMENTS

The authors thank Dr. Michelle Everson for her great contributions to the work presented in this paper. The authors also thank the statistics education community for their encouragement and support, especially those at the University of Minnesota and those who administered the BLIS assessment to their students. Lastly, the authors thank the editors and anonymous referees for their valuable feedback.

## REFERENCES

Allen, K., Stone, A., Rhoads, T. R., & Murphy, T. J. (2004, June). The Statistics Concepts Inventory: Developing a valid and reliable instrument. *Proceedings of the 2004 American Society for Engineering Education Annual Conference and Exposition* (pp. 1–15). Salt Lake City, UT.

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

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Anderson, B. L., Gigerenzer, G., Parker, S., & Schulkin, J. (2014). Statistical literacy in obstetricians and gynecologists. *Journal for Healthcare Quality, 36*(1), 5–17.

Atkinson, M. P., Czaja, R. F., & Brewster, Z. B. (2006). Integrating sociological research into large introductory courses: Learning content and increasing quantitative literacy. *Teaching Sociology, 34*(1), 54–64.

Baker, F. B. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Retrieved from http://echo.edres.org:8080/irt/baker/final.pdf (Original work published in 1985)

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi& J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3–15). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Buckendahl, C., & Plake, B. S. (2006). Evaluating tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 725–738). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Catalysts for Change (2013). *Statistical thinking: A simulation approach to modeling uncertainty* (2nd ed.). Minneapolis, MN: CATALYST Press.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. doi: 10.18637/jss.v048.i06

Chance, B. L. (1997) Experiences with authentic assessment techniques in an introductory statistics course, *Journal of Statistics Education, 5*(3).
[Online: http://www.amstat.org/publications/jse/v5n3/chance.html]

Chick, H. L., & Pierce, R. (2013). The statistical literacy needed to interpret school assessment data. *Mathematics Teacher Education and Development*, 15(2).

[Online: https://files.eric.ed.gov/fulltext/EJ1018712.pdf]

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

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education, 49*, 560–575. doi: 10.1111/medu.12678

delMas, R. C. (2002). Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education, 10*(3).
[Online: http://www.amstat.org/publications/jse/v10n3/delmas_discussion.html]

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.
[Online: https://iase-web.org/documents/SERJ/SERJ6(2)_delMas.pdf]

Downing, S. M. (2006). 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.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

Gal, I. (1998). Assessing statistical knowledge as it relates to students' interpretation of data. In D. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K–12*, (pp. 275–295). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

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

Galesic, M., & Garcia-Retamero, R. (2010). Statistical numeracy for health: A cross-cultural comparison with probabilistic national samples. *Archives of Internal Medicine*, *170*(5), 462–468.

Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, *2*(1), 22–38.
[Online: http://iase-web.org/documents/SERJ/SERJ2(1).pdf]

Garfield, J., Aliaga, M., Cobb, G., Cuff, C., Gould, R., Lock, R., … Witmer, J. (2005). Guidelines for assessment and instruction in statistics education (GAISE) project: College report. Alexandria, VA: American Statistical Association.
[Online: http://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., & Chance, B. (2002). The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project. NSF CCLI grant ASA- 0206571.
[Online: https://apps3.cehd.umn.edu/artist/index.html]

Garfield, J., delMas, R., & Chance, B. (2003, April). *The web-based ARTIST: Assessment Resource Tools for Improving Statistical Thinking*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

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., & Franklin, C. (2011). Assessment of learning, for learning, and as learning in statistics education. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 133–145). New York: Springer.

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

Hesterberg, T., Monaghan, S., Moore, D. S., Clipson, A., & Epstein, R. (2003). Bootstrap methods and permutation tests [Companion Chapter]. In D. S. Moore, G. P. McCabe, W. M. Duckworth, & S. L. Sclove, *The Practice of Business Statistics*. New York: W. H. Freeman and Company.

Holcomb, J., Rossman, A., & Chance, B. (2011). Exploring student understanding of significance in randomization-based courses. In *Proceedings of the 58th World Statistical Congress, Dublin, Ireland* (pp. 880–889). The Hague, The Netherlands: International Statistical Institute.

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

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000). A framework for characterizing children's statistical thinking, *Mathematical Thinking and Learning, 2*, 269–307.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. doi: 10.1111/jedm.12000

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

Lane-Getaz, S. J. (2017). Is the *p*-value really dead? Assessing inference learning outcomes for social science students in an introductory statistics course. *Statistics Education Research Journal*, *2*(1), 22–38.
[Online: https://iase-web.org/documents/SERJ/SERJ16(1)_LaneGetaz.pdf]

Lehohla, P. (2002). Promoting statistical literacy: A South African perspective. In B. Phillips (Ed.), *Proceedings of the sixth International Conference on Teaching Statistics,* Cape Town, South Africa Voorburg, The Netherlands: International Statistical Institute.

Literacy [Def. 1]. (n.d.a). In *Cambridge Dictionaries Online.* Retrieved March 9, 2013, from http://dictionary.cambridge.org/dictionary/american-english/literacy?q=literacy

Literacy. (n.d.b). In *Merriam-Webster Online.* Retrieved March 9, 2013, from http://www.merriam-webster.com/dictionary/literacy

Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2013). *Statistics: Unlocking the power of data.* Hoboken, NJ: Wiley.

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

McLauchlan, C., & Schonlau, M. (2016). Statistical literacy in the classroom: Should introductory statistics courses rethink their goals? *Statistics, Politics and Policy, 7*(1–2), 99–115.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13–103). New York: American Council of Education and Macmillan.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176.

OECD (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the OECD Survey of Adult Skills*. Paris: OECD Publishing. doi: 10.1787/9789264128859-en

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–64.

Pierce, R., & Chick, H. (2013). Workplace statistical literacy for teachers: Interpreting box plots. *Mathematics Education Res. Journal*, *25*(2), 189–205. doi: 10.1007/s13394-012-0046-3

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory*. New York: Routledge.

Ridgway, J., Nicholson, J., & McCusker, S. (2008). Reconceptualising "statistics" and "education." In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 311–322). New York: Springer.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.
[Online: http://www.jstatsoft.org/v17/i05/]

Rossman, A. (2007, May). *Seven challenges for the undergraduate statistics curriculum*. Presentation at United States Conference on Teaching Statistics, The Ohio State University.
[Online: http://www.statlit.org/pdf/2007RossmanUSCOTS.pdf]

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

Schield, M. (2006). Statistical literacy survey analysis: Reading graphs and tables of rates and percentages. In A. Rossman & B. Chance (Eds.), *Proceedings of the seventh International Conference on Teaching Statistics,* Salvador, Bahia, Brazil. Voorburg, The Netherlands: International Statistical Institute.

Schield, M. (2017). GAISE 2016 promotes statistical literacy. *Statistics Education Research Journal, 16*(1), 50–54.
[Online: https://iase-web.org/documents/SERJ/SERJ16(1)_Schield.pdf]

Sharma, S., Doyle, P., Shandil, V., & Talakia'atu, S. (2012). Developing statistical literacy with Year 9 students: A collaborative research project. In C. Smith (Ed.), *Proceedings of the British Society for Research into Learning Mathematics*, *32*(3), 167–172. Cambridge, England: The University of Cambridge.

Smith, M. (2002, December). *The rocky road to statistical literacy*. Presentation at the ACCOLEDS Training, Vancouver, Canada.
[Online: https://www.slideserve.com/caden/the-rocky-road-to-statistical-literacy]

Sundre, D., Thelk, A., & Wigtil, C. (2008). The Quantitative Reasoning Test, Version 9 (QR 9): Test Manual. *Harrisonburg, VA: Center for Assessment and Research Studies.*
[Online: http://www.madisonassessment.com/uploads/qr-9_manual_2008.pdf]

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Boston, MA: Pearson.

Tintle, N. Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., & VanderStoep, J. (2013). *Introduction to statistical investigations*. New York: Wiley.
[Online: http://www.isi-stats.com/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).
[Online: https://amstat.tandfonline.com/doi/abs/10.1080/10691898.2011.11889599]

Turegun, M. (2011). *A model for developing and assessing community college students' conceptions of the range, interquartile range, and standard deviation.* (Unpublished doctoral dissertation). University of Oklahoma, Norman, OK.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/dissertations/dissertations.php]

Wade, B. A. (2009). *Statistical literacy in adult college students* (Unpublished doctoral dissertation). The Pennsylvania State University, University Park, PA.

Walker, H. M. (1951). Statistical literacy in the social sciences. *The American Statistician*, *5*(1), 6–12.

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association, 88*(421), 1–8.

Watson, J. M. (2011). Foundations for improving statistical literacy. *Statistical Journal of the IAOS, 27*(3&4), 197–204.

Weathington, B. L., Cunningham, C. J., & Pittenger, D. J. (2010). *Research methods for the behavioral and social sciences*. Hoboken, NJ: Wiley.

Yolcu, A. (2012). *An investigation of eighth grade students' statistical literacy, attitudes towards statistics and their relationship* (Unpublished doctoral dissertation). Middle East Technical University, Ankara, Turkey.

LAURA ZIEGLER
Iowa State University
Room 3409, Snedecor Hall
2438 Osborn Drive
Ames, IA 50011
USA