

EXAMINING THE ROLE OF CONTEXT IN STATISTICAL LITERACY ASSESSMENT

SAYALI PHADKE

*Pennsylvania State University
sayalip@psu.edu*

MATTHEW BECKMAN

*Pennsylvania State University
beckman@psu.edu*

KARI LOCK MORGAN

*Pennsylvania State University
klm47@psu.edu*

ABSTRACT

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report advocates for use of real data with context and purpose. This work contributes to the growing literature on assessing statistical literacy by investigating the influence of context as it relates to assessment performance among post-secondary introductory statistics students. We discuss the development of an isomorphic form of an existing assessment instrument, and report results which concluded that test takers demonstrated lower statistical literacy scores when assessment tasks incorporated real data from published studies as context when compared with functionally similar tasks such as those with a contrived data set and a realistic context.

Keywords: *Statistics education research; Statistical literacy; Real data; Context; Isomorphic assessment*

1. INTRODUCTION

The importance and role of statistical literacy has been discussed extensively in the statistics education literature (Ben-Zvi et al., 2018; Ben-Zvi & Garfield, 2004, 2008; Engel, 2017; Gal, 2002; Garfield et al., 2010; Gould, 2017; Utts, 2021; Wallman, 1993; Watson, 1998; Watson & Callingham, 2003; Weiland, 2017). Guiding documents that inform researchers and practitioners alike, such as the *GAISE College Report* (2016), *International Handbook of Research in Statistics Education* (Ben-Zvi et al., 2018), and *GAISE Pre-K–12 Report* (Bargagliotti et al., 2020), have highlighted cognitive outcomes, curriculum considerations, teaching practices, and assessments. As well, the American Statistical Association (ASA) seeks to “build a statistically literate society” as one of its objectives under the strategic goal of statistics education (<https://www.amstat.org/education>). In parallel, the PARIS21 Partnership in Statistics for Development in the 21st Century partnership, a consortium of global organizations that includes the United Nations, European Union, Organization for Economic Cooperation and Development, International Monetary Fund, and the World Bank, also considers statistical literacy to be a focus of its work (<https://www.paris21.org/>). Even though definitions of statistical literacy vary in some aspects (Sharma, 2017), mainstream conceptualizations of statistical literacy describe its critical role in promoting a citizenry that is more capable of understanding the world around them and making evidence-based decisions in their private and public lives. One marker of a statistically literate citizen is the ability to make sense of statistical insights encountered in the context of their day-to-day lives (Wilks, 1951).

The GAISE recommendations endorsed by the ASA, first in 2005 and again in 2016, advocated that how we teach contemporary statistics ought to “integrate real data with context and purpose” (2016, p. 6). This remark was not new thinking, rather it was an extension of a broader mantra that there is no statistics without context (Rao, 1975). As such, a considerable amount of work has discussed the value

of contexts and methods for utilizing contexts familiar to the students (Brown, 2016; Gal, 2019; Garfield et al., 2012; Lee & Tran, 2015; Ratnawati et al., 2020). Even still, the GAISE College Report (2016, p. 60–62) pointed out that data/context can be described according to a taxonomy, which includes (1) naked data with no apparent context presented; (2) realistic data that describes (or at least feigns) a believable context but has been blunted or manipulated in some way to emphasize a specific concept or outcome; (3) real data such that either the context has little compelling scientific value (e.g., summary statistics of exam scores), or has not cited a bona fide source; (4) real data from a real study make clear that the investigation involves authentic data and from real and compelling context with a provided reference.

Concurrently, studies focusing on improving statistical literacy among students at various levels have been conducted (Barbieri & Giacché, 2006; Carmichael, 2010; Ferligoj, 2015; Schield, 2004; Suhermi & Widjajanti, 2020; Watson, 2011). There are, however, relatively few published assessments of statistical literacy outcomes intended for research use (Sabbag et al., 2018; Sanchez, 2007; Ziegler & Garfield, 2018), and no studies were found to have investigated the impact of fidelity to real data presented in context during the assessment of statistical literacy.

Isomorphic assessment tasks are one possible mechanism for investigating the impact of modifying context of an assessment task, while preserving the conceptual learning objective. Specifically, a structurally isomorphic task is intended to mirror a base item in structure—e.g., concept, phrasing, and the function of corresponding distractors for selected response tasks as closely as practical—and differs only in superficial details, while measuring the same underlying concept, construct, or learning outcome (Gick & Holyoak, 1980; Millar & Manoharan, 2021; Williamson et al., 2002). Importantly, Lehrer and Schauble (2007) and Fay et al. (2018) highlighted that isomorphic assessment tasks cannot guarantee that respondents' cognitive processes in answering these tasks will be comparable. While there has been limited study of isomorphic tasks in statistics education, isomorphs have been studied in other STEM disciplines, including physics education (Barniol & Zavala, 2014; Kusairi et al., 2017, 2020; Lin & Singh, 2011; Luger & Bauer, 1978; Suganda et al., 2020) and computer science education (Millar & Manoharan, 2021; Parker et al., 2016).

There is, however, some evidence (e.g., Kusairi et al., 2017) that more practice on the base topics improves performance, as discussed by Lovett and Greenhouse (2000), although the benefit is not necessarily symmetric. In one study related to learning transfer across disciplinary boundaries, Bassok and Holyoak (1989) reported that training in mathematics facilitated transfer to physics but not the other way round. These studies of transfer using isomorphic tasks deployed a variety of types of assessments developed principally to serve the objectives of a single study and not necessarily for use by other researchers, student populations, or institutions. Parker et al. (2016) emphasized the importance of developing assessment instruments that undergo the rigorous process of collecting and evaluating reliability and validity evidence, and for subsequent researchers to adopt tools that have endured such scrutiny.

The Basic Literacy in Statistics (BLIS) assessment (Ziegler, 2014; Ziegler & Garfield, 2018) is one such assessment tool carefully developed for research use with a focus on statistical literacy outcomes. Ziegler (2014, p. 5) defined statistical literacy as the “ability to read, understand, and communicate statistical information.” In this paper, we investigate a modification of the BLIS assessment in order to study the impact of changes to context. The goal is not to introduce a novel assessment tool, instead we have created a parallel form that intends to utilize structural isomorphs of the corresponding BLIS tasks—i.e., a modification of BLIS called “MBLIS” hereafter—to answer the following research questions related to the impact of real data from published studies with contexts believed to be familiar to students while completing an assessment of statistical literacy:

- (RQ1) How does the functioning of the isomorphic items compare to the functioning of the original assessment?
- (RQ2) Is there evidence to suggest that test-takers respond to the underlying statistical question differently if the item is based on a modified context?

In Section 2 we discuss the development of MBLIS and design of the study implemented to gather reliability evidence and to develop a validity argument. Section 3 discusses results from the study data, concluding with a discussion of limitations and future research opportunities based on this work in Section 4.

2. METHODS

This section discusses the development of MBLIS, design of the study, and statistical methods used to analyze data from the study. Section 2.1 details the study design, Section 2.2 describes the sample, Section 2.3 discusses development of the MBLIS assessment form, and Sections 2.4 and 2.5 summarize the analytical methods associated with RQ1 and RQ2, respectively.

2.1. STUDY DESIGN

The BLIS and MBLIS assessment forms were administered to students in a large-enrollment introductory statistics course at a public research university in the Mid-Atlantic region of the eastern United States. The course adopts a simulation-based approach to introduce statistical inference, has no formal prerequisites, and serves a general education student audience with a wide range of academic interests.

This study was conducted in the last week of classes for the Spring 2021 semester and was administered outside of class. Students received an email from the course coordinator outlining the procedure as well as the incentive offered. Students could earn three extra credit points over and above the 1000 points possible for the overall course grade.

Students who consented to participate in the research study were randomly assigned to one of the assessment forms—BLIS or MBLIS—using the built-in randomization in Qualtrics. This gave us a baseline on the original assessment within the target population, facilitating a comparison of results across the results from the Ziegler (2014) field test and our study. At the end of the assigned assessment (BLIS or MBLIS), students were surveyed to self-report demographics as well as interest/engagement with topics such as diversity questions, immigration, public policy and governance, and their experience of interacting with items pertaining to the COVID-19 pandemic.

2.2. PARTICIPANTS

At the time this study was conducted, 1,653 students were enrolled in the course. Of these students, 1,532 opened the initial screening survey and 1,489 of them consented to participate in the research. We excluded the 69 students who left the survey completely blank. Consistent with the analysis conducted by Ziegler (2014), we only analyzed the 1,253 complete responses available from the study comprising of 638 responses on the original form and 615 on the parallel form. Approximately eighty two percent of the responses from participants who opened the survey were usable, which is equivalent to 76% of total enrollment at the time the study was conducted.

2.3. ASSESSMENT MODIFICATION

The original BLIS assessment included 37 tasks. According to the GAISE taxonomy, 19 tasks utilized real data with an overt link to a real study (e.g., scientific research), five applied real data without a clear link to “a compelling application of statistics” (e.g., class quiz scores), 12 used realistic data (e.g., contrived or hypothetical), and one (1) relied only on naked data with no apparent context described. A subset of six tasks from the BLIS assessment were presented verbatim on the MBLIS assessment to function as anchor tasks between the two forms. Two of the anchor tasks utilized realistic contexts and four utilized real data from a real study. The remaining 31 tasks from the BLIS assessment were modified for the MBLIS form.

All modified tasks for the MBLIS form were intended to function as structural isomorphs that assess the same concept or learning outcome while incorporating real data from real contexts. Preliminary work surveyed students regarding various topics with broad relevance such as climate change, immigration, race-related issues, and the COVID-19 pandemic. The COVID-19 pandemic emerged as a unifying context due to seemingly ubiquitous impact on daily life at the time of the study, presenting a unique opportunity for research in the form of a topic with nearly universal relevance. Of course, we acknowledge the devastating effects of the pandemic and the associated trauma endured by many. Therefore, while deciding to proceed with the COVID-19 pandemic as the unifying context for MBLIS task modifications, we actively avoided contexts and data pertaining to severe illness, loss of life, and

other serious health effects. For example, context choices included topics of collateral impact or disruption such as dental care choices, flight cancellations, restaurant visit frequency, and air pollution.

In order to prepare for isomorphic task modification, each item on the BLIS assessment was evaluated to identify structural features that should be preserved such as type(s) of variable(s), parameter(s) of interest, type of sample, type of study (observational versus experimental), and whether creation of the item required access to raw data or summary statistics or neither. Each original item was also categorized according to the taxonomy discussed in the *GAISE College Report* (2016): Naked data, Realistic data, Real data, and Real data from a real study. The primary purpose of considering this categorization was to analyze whether any observable effect is associated with the degree of change from the original data category to the modified category (real data from a bona fide research study). While a tension arises based on the constraint of prioritizing fidelity to real data from real studies for MBLIS, a concerted effort was made to favor studies and task modifications that preserved the distance of the sample statistic from the parameter, scale of the p -value, small sample size, and overall length (in characters) of the names of variables or context description.

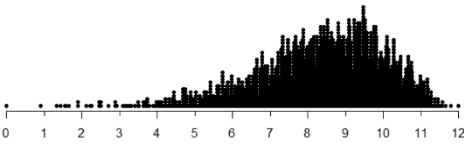
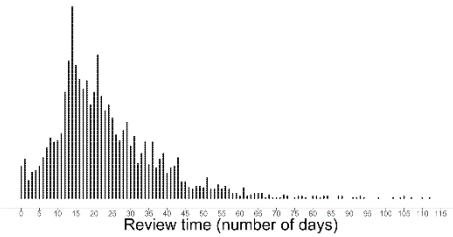
Table 1 demonstrates an item that was based on real data from a real study, leading to an isomorph that retained the structure of the original item very closely. This is an example of a pure isomorph. Naturally, BLIS tasks that utilized real data could be modified into isomorphs readily. Whereas tasks that utilized naked or realistic (i.e., hypothetical) data required more extensive modification in order to introduce context for the MBLIS counterpart where there had previously been little or none.

Table 1. (Left) BLIS task with real data from a real study; (Right) MBLIS task that cites real data from a real study

Original item stem	Modified item stem
The Pew Research Center surveyed a nationally representative group of 1,002 American adults in 2013. Of these adults, 21% have had an email or social networking account compromised. Identify the population about which the Pew Research Center can make inferences from the survey results and the sample from that population.	The Pew Research Center surveyed a nationally representative group of 12,648 U.S. adults in November 2020. Of these adults, 62% said they would be uncomfortable being among the first to get the vaccine for COVID-19. Identify the population about which the Pew Research Center can make inferences from the survey results and the sample from that population. <i>Source: Pew Research Center.</i>

In an extreme case (Table 2), the item was based on naked data in the original BLIS, meaning there was no apparent context provided at all, so any modification to introduce real data with real context was necessarily more extensive. For each modified item, the source link was provided at the end of the prompt. It was added on a separate line with the word “Source” followed by a very short key phrase identifying the source with a hyperlink. This was intended to underscore the authenticity and credibility of the contexts presented in the item without distracting the test-taker from the key task. During a think-aloud conducted prior to data collection, a respondent explicitly stated that this added legitimacy to the questions in the student’s mind.

Table 2. (Left) BLIS task with naked data; (Right) MBLIS task that cites real data from a real study

Original item	Modified item
<p>The distribution for a population of measurements is presented below.</p>  <p>A sample of 10 randomly selected values will be taken from the population and the sample mean will be calculated. Which of the following intervals is MOST likely to include the sample mean?</p> <ul style="list-style-type: none"> • 6 to 7 • 8 to 9 • 9 to 10 • 10 to 11 	<p>For scientific credibility, journal articles are reviewed by other scientists before publication. This process is called peer-review. Researchers collected data to study how the pandemic has affected the peer-review timelines for six Ecology journals. The plot below shows the distribution of number of days taken by all reviewers to review papers assigned to them.</p>  <p>A sample of 10 randomly selected values will be taken from the population and the sample mean will be calculated. Which of the following intervals is MOST likely to include the sample mean?</p> <ul style="list-style-type: none"> • 0 to 10 • 10 to 20 • 20 to 30 • 40 to 50 <p>Source: Research article.</p>

The BLIS assessment form administered was unchanged from the version provided in Ziegler (2014). For items that involved data visualization, plots were created using the *ggplot2* package (Wickham, 2016) in *R* (R Core Team, 2023). Even though some of the original visualizations were created using the *plotrix* package (Lemon, 2006), same aesthetics and scales were maintained in the modified visualizations.

2.4. BLIS AND MBLIS ASSESSMENT FORM COMPARISON

Recall RQ1: How does the functioning of the isomorphic items compare to the functioning of the original assessment?

Although the purpose of this study was not primarily intended to introduce and evaluate a novel assessment tool, it is meaningful to characterize instrument reliability and validity evidence to steer intuition about whether MBLIS and BLIS appear to measure underlying constructs as similarly as possible, such that the principle difference between comparable tasks is simply the nature of context utilization to the extent possible (Parker et al., 2016). Consequently, we gathered data to evaluate reliability evidence and develop a validity argument using expert reviews, think-aloud interviews, and a study adapted slightly to suit the development of an isomorphic assessment instead of a new instrument (American Educational Research Association et al., 2014). For example, one such “adaptation” requested that expert reviewers prioritize evaluating whether each MBLIS task was structurally isomorphic to the corresponding BLIS task.

The three expert reviewers recruited included the author of the original BLIS assessment, a senior statistics education researcher who supervised the design and development of the original BLIS assessment, and a statistics education researcher with experience in assessment development and educational measurement. The expert reviewers carefully reviewed the modified instrument with the prompt, “Please consider each modified item vis-a-vis the original item and comment on whether they

are comparable in measuring the underlying learning outcome.” The instrument was updated based on the expert feedback prior to implementation for student data collection. For the final assessment instrument in use, six out of the 37 total items were randomly selected to remain unchanged as anchors appearing verbatim on both assessment forms as a means for comparison of the student groups presented with each assessment form to screen for baseline disparity (i.e., internal-anchor design [Livingston, 2004]).

2.5. STUDENT ASSESSMENT PERFORMANCE ASSOCIATED WITH CHANGES OF CONTEXT

Recall RQ2: Is there evidence to suggest that test-takers respond to the underlying statistical question differently if the item is based on a modified context?

We seek to compare and contrast student assessment performance when tasks are modified from several different taxa of data/context (e.g., Naked, Realistic, Real, Real from Real Study) on the BLIS assessment to functionally isomorphic tasks on the MBLIS form that utilize data/contexts that conform to the Real Data from a Real Study class. To this end, we fit a generalized linear mixed model (GLMM) using the “glmer” function from the *lme4* R package (R Core Team, 2023; Bates et al., 2015) as shown in Equation 1.

Equation 1 illustrates the linear mixed effects logistic regression model used:

$$\begin{aligned} \text{logit}(Y_{ij}) = & \beta_0 + \beta_1(\text{WordCountContrast}_j) + \beta_2(\text{MBLIS}) + \\ & \beta_3(\text{Anchor}_j) + \beta_4(\text{Naked}_j) + \beta_5(\text{Real}_j) + \beta_6(\text{Realistic}_j) + \\ & \beta_7(\text{MBLIS}*\text{Anchor}_j) + \beta_8(\text{MBLIS}*\text{Naked}_j) + \beta_9(\text{MBLIS}*\text{Real}_j) + \\ & \beta_{10}(\text{MBLIS}*\text{Realistic}_j) + u_i + v_j, \end{aligned}$$

where the random effects due to ability of student i and difficulty of question j are assumed to follow independent Normal distributions:

$$\begin{aligned} u_i & \sim N(0, \sigma_u^2); \\ v_j & \sim N(0, \sigma_v^2). \end{aligned}$$

Our outcome variable is the log-odds of the response for question j by student i (i.e., Y_{ij}). Fixed effects include test form (MBLIS indicator variable), indicator variables for the data/context taxa associated with each question for the original BLIS form (Anchor, Naked, Realistic, Real, Real with Real Study), as well an adjustment based on the change in word count from each BLIS question to the corresponding MBLIS question. The latter adjustment for word count was modeled as a contrast between the word count in use on the BLIS and MBLIS form for each question (WordCount(MBLIS question j)—WordCount(BLIS question j)) as a proxy for the increased/decreased reading burden associated with the contexts accompanying the corresponding, functionally isomorphic items on the two assessment forms. Table 3 includes a summary of word counts within each group of items based on context taxa. Additionally, the model includes random effects for individual student ability and question difficulty.

Table 3: Average word counts for items in each category of context taxa

Item context on BLIS	Instrument	# of items	Median	Mean	SD
Real data with real study	BLIS	15	103	109.07	61.22
Real data	BLIS	5	75	72.2	45.25
Realistic data	BLIS	10	59.5	62.5	23.89
Naked data	BLIS	1	43	43	NA
Anchor items	BLIS	6	94.5	129.67	96.76
Real data with real study	MBLIS	15	96	107.87	50.32
Real data	MBLIS	5	78	75.8	34.19
Realistic data	MBLIS	10	82.5	78.3	27.8
Naked data	MBLIS	1	87	87	NA
Anchor items	MBLIS	6	94.5	129.67	96.76

Of particular interest for RQ2 is the interaction between the test form and context type, and the effect on task difficulty associated with the contrast between the context type represented on the original form of the BLIS assessment (Naked, Realistic, Real, Real with Real Study) and the MBLIS form after all tasks have been modified to represent Real with Real Study. While word count associated with each task is not of primary interest to the research question, it is included as a proxy for the incremental change in cognitive load attributed to increased reading burden when comparing the version on the BLIS form to the version on the MBLIS form for each task.

Note that standard item response modeling seeks to estimate student ability as well as item parameters (e.g., difficulty), but we model student and question as random effects in order to better isolate the influence associated with the change to context. Also, it bears mentioning that the model intercept, β_0 , is not only interpretable, but quite important as a basis for comparisons as it represents BLIS tasks that utilize real data from a real study, for which corresponding questions on MBLIS would have an identical word count (i.e., contrast = 0).

3. RESULTS

3.1. (RQ1) COMPARING BLIS AND MBLIS ASSESSMENT FORMS

In this section, we address the first research question (RQ1): How does the functioning of the isomorphic items compare to the functioning of the original assessment?

To ensure reliability of MBLIS and ensure its validity for the intended use, we considered two sources of evidence. Expert reviews contributed to the evidence for a validity argument, and we analyzed data from the study using the same metrics used for BLIS (Ziegler, 2014) to check for reliability and further support the validity argument. The goal was to ensure comparability of BLIS and MBLIS in the way a parallel form of an assessment would be expected.

Many of the experts' comments suggested a change in the structure of the original BLIS instrument, which was deemed out of scope for the purpose of this study. Of the 31 MBLIS items under consideration, 12 remained unchanged, 16 received minor changes, and one received major updates based on the expert reviews. The remaining two items were discussed at length and rewritten based on the expert reviews. One of these two items is presented in Table 4. The other isomorph that underwent a significant change based on expert reviews shared a key characteristic with the item in Table 4. The context in that item also had a binary outcome that could be intuitively assumed to be equiprobable. In both these cases, the original context of our choice was retained while rephrasing the item stem. The resulting instrument based on these changes was deployed in the study (see Section 2.1).

Table 4. *Implicit assumption changed*

Original item stem	Modified item stem
Two students are flipping coins and recording whether or not the coin landed heads up. One student flips a coin 50 times and the other student flips a coin 100 times. Which student is more likely to get 48% to 52% of their coin flips heads up?	Penn State University administrators surveyed all undergraduate students to capture feedback from the entire student body on several issues. As a result, they learned that 86% of all students planned to return in fall 2020. Despite knowing the proportion for all Penn State students as a whole, several instructors surveyed their own classes in order to be sensitive to the views of their students. One instructor had a class with 50 students and another instructor had a class with 100 students. Assuming both classes were representative of the entire student body at Penn State, which instructor was more likely to find that 84% to 88% of their students would plan to return in fall 2020?

Source: Adapted from Penn State News.

Additional data analyses confirmed that BLIS and MBLIS are comparable in terms of several key metrics. The coefficient alpha values (0.78 for BLIS and 0.77 for MBLIS) confirmed comparably high internal consistency among test items and local independence among testlet items. A testlet includes two or more items that share a common question stem. Principal Component Analysis (PCA) confirmed that the parallel form is performing comparably when considering the assumptions of unidimensionality in the assessment scale and local independence among items. Acceptability of the local independence assumption was also verified using single-factor confirmatory analyses. Various test and item information metrics were considered based on the partial credit IRT model. The test information metrics indicated that MBLIS contains a few more items measuring respondents at higher ability levels than BLIS, and also provides most information at a slightly higher ability level. Finally, item difficulty rankings and item characteristic curves displayed comparable ranges of difficulty, though slightly uneven on either side of zero indicating a reasonable mixture of item difficulty sustained by both forms.

3.2. (RQ2) STUDENT ASSESSMENT PERFORMANCE ASSOCIATED WITH CHANGES OF CONTEXT

In this section, we address the second research question (RQ2): Is there evidence to suggest that test-takers respond to the underlying statistical question differently if the item is based on a modified context? Before interpreting the mixed effects logistic regression model that estimates the effects associated with the two assessment forms and taxa of data/context present, we first look briefly at an overall comparison of holistic performance on the BLIS and MBLIS forms, and review summaries of the data.

Summary of assessment performance. First we compare overall performance. Figure 1 shows a distribution of total score by assessment form. Both assessments are scored as one point per correctly answered question with a highest possible total of 37 points. The violin plot on the top represent scores on MBLIS ($n = 615$ students) and the one on the bottom represent scores on BLIS ($n = 638$ students).

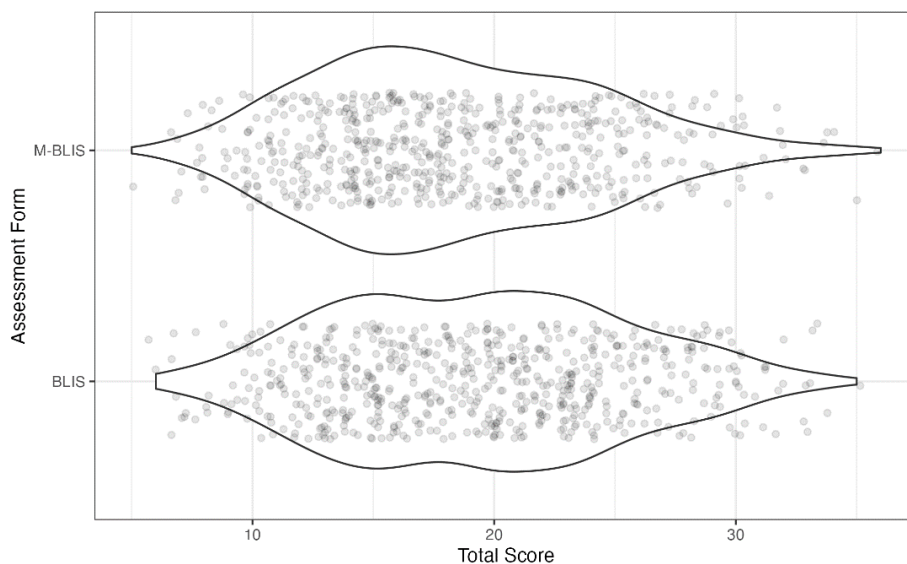


Figure 1. Violin plot comparison of total score associated with the BLIS form and the MBLIS form

Overall scores were comparable on both assessments, as suggested by the numerical summaries in Table 5. A two-sided t -test indicated statistically discernible evidence of a difference in mean scores for BLIS and MBLIS (p -value = 0.005; 95% CI: 0.28 to 1.58), yet the magnitude of difference—approximately one point out of 37 possible—adds to our evidence that the two forms appear comparable overall, as intended. For additional information, Table 8 in the Appendix contains percentage of correct responses per item. Table 9 in the Appendix tabulates selected responses, that is, percentages of respondents who chose each distractor.

Table 5. Summary statistics of total scores

Instrument	<i>n</i>	Mean	Median	<i>SD</i>	<i>IQR</i>
BLIS	638	19.31	19	5.98	8.75
MBLIS	615	18.38	18	5.78	9

We now compare and contrast student assessment performance based on the different context taxa (Naked, Realistic, Real, Real from Real Study) on the BLIS assessment. Table 6 shows the number of items associated with each taxa on the BLIS form, yet recall that all tasks are converted to real data from a real study for the MBLIS form. Note that the Anchor items and the items associated with real data from real studies on the BLIS form (and therefore on both forms) resulted in the smallest differential between BLIS and MBLIS results. The MBLIS tasks with real data from real contexts may have been slightly more challenging for students than corresponding BLIS tasks that had been associated with realistic data or real data. There is only one task on the BLIS form with naked data (no apparent context), so it is impossible to tell whether the large difference observed is meaningful. A complete table of differences for corresponding items on the BLIS and MBLIS forms appears in the appendix (Table 8), which shows that the difference observed for the Naked data task is indeed large, but it is not the only large difference nor is it the largest difference for any single item.

Table 6. Difference in average proportion of respondents correctly answering items in each context category

Item context on BLIS	# of items	Average proportion correct—BLIS	Average proportion correct—MBLIS	Difference
Real data with study	15	0.471	0.462	-0.009
Real data	5	0.657	0.619	-0.038
Realistic data	10	0.597	0.551	-0.046
Naked data	1	0.428	0.246	-0.182
Anchor items	6	0.428	0.433	0.005

Summary of statistical modeling results. Table 7 shows the fitted mixed effects logistic regression model associated with Equation 1. Of particular note is the statistically discernible evidence of an interaction between the assessment form (i.e., MBLIS indicator variable) and the BLIS context types (e.g., indicator variables).

Table 7. Coefficients of the logistic regression model and variances of random effects

Fixed Effects	Coefficient Estimate	Standard Error	Test statistic	<i>p</i> -value
(Intercept)	-0.148	0.191	-0.777	0.437
mblis	-0.042	0.051	-0.824	0.410
wordCountContrast	-0.003	0.001	-2.550	0.011
blisAnchor	-0.185	0.353	-0.525	0.599
blisNaked	-0.243	0.752	-0.323	0.746
blisRealistic	0.632	0.299	2.116	0.034
blisReal	0.948	0.379	2.503	0.012
mblis:blisAnchor	0.067	0.059	1.137	0.255
mblis:blisNaked	-0.727	0.145	-5.024	<0.001
mblis:blisRealistic	-0.136	0.057	-2.391	0.017
mblis:blisReal	-0.015	0.067	-2.227	0.026

Random Effects	Variance
student	0.479
question	0.526

For clarity, we rewrite the fixed effects portion of the fitted model associated with Equation 1 to illustrate the result associated with each data/context taxa represented on the BLIS form:

$$\text{logit}(\hat{Y}_{ij} / \text{Real with study}) = -0.148 - 0.042(\text{MBLIS}) - 0.003(\text{wdctContrast}_j)$$

$$\text{logit}(\hat{Y}_{ij} / \text{Anchor}) = -0.333 + 0.025(\text{MBLIS}) - 0.003(\text{wdctContrast}_j)$$

$$\text{logit}(\hat{Y}_{ij} / \text{Naked}) = -0.391 - 0.769(\text{MBLIS}) - 0.003(\text{wdctContrast}_j)$$

$$\text{logit}(\hat{Y}_{ij} / \text{Realistic}) = 0.484 - 0.178(\text{MBLIS}) - 0.003(\text{wdctContrast}_j)$$

$$\text{logit}(\hat{Y}_{ij} / \text{Real}) = 0.800 - 0.191(\text{MBLIS}) - 0.003(\text{wdctContrast}_j)$$

Before interpreting the effects associated with the two assessment forms and taxa of data/context present, it should be noted that there is a modest statistically noticeable effect associated with word count contrast, such that the odds of a correct response decreases by approximately $1 - \exp(-0.003) = 0.3\%$ for each additional word in the question prompt. Said another way, questions appear to be more challenging for students when more reading is required of them, on average (p -value = 0.011).

According to these data, there is not evidence of a statistically noticeable difference between the BLIS and MBLIS forms for the Anchor tasks (p -value = 0.255), nor the questions that utilize real data with a real study on both forms (p -value = 0.410). This suggests no discernable difference was observed for corresponding questions on the BLIS and MBLIS forms that prompted students with the same taxa of data/context after adjusting for word count differences.

For questions that were presented as either Realistic data or Real data (without a cited study) on the BLIS form, there was a statistically discernible difference in the odds of correct response on average for students who were presented structurally isomorphic questions that utilized real data that cite a real study on the MBLIS form, after adjusting for word count differences. Specifically, students were about $1 - \exp(-0.191) = 17.4\%$ less likely to correctly answer MBLIS questions that cite real data from a real study, when compared to peers presented with structurally isomorphic questions on the BLIS form that purport to include Real data without citing the actual study. Similarly, students were about $1 - \exp(-0.178) = 16.3\%$ less likely to answer MBLIS questions that cite real data from a real study, when compared to peers presented with structurally isomorphic questions on the BLIS form that reference ostensibly Realistic data.

For questions that were presented as Naked data (i.e., no apparent context) or students that completed the BLIS form, a large and statistically noticeable difference was observed in the data. While the result was consistent with the trend observed among the other context types, the BLIS form only included one question classified as Naked data so the effect should not be overstated.

4. DISCUSSION

The research studies used as contexts for the MBLIS form were intentionally focused on topics that students reported would be of interest to them. The data, however, suggested that students were more likely to answer the question correctly as presented in the BLIS form. Items on the BLIS form used contexts with varying degrees of reduced ties to real data with context and purpose as would have been advocated by the GAISE (2016) recommendations. Frankly, this result was a surprise to the research team for this study. We had expected context to play a meaningful role, but anticipated prior to data collection that students might be *more likely* to produce a correct response for questions that draw upon familiar contexts, especially when they invoke a topic that students are likely to have engaged with or at least thought about on their own terms outside of class. We expected the familiarity with context to remedy the notion of “suspension of sense-making” in which students are tempted to provide answers that they think a teacher expects (e.g., Carotenuto et al., 2021).

Underscoring this result is the lack of evidence for a statistically noticeable difference between two groups of tasks that function as active and passive controls in the study. For example, no compelling

difference was observed between BLIS or MBLIS results for a set of six anchor items that were reproduced verbatim, nor was a difference observed for questions that appear on both forms as Real data with real study. The latter is ostensibly an “active control” in the MBLIS form based on 19 items that had been updated following exactly the same process, expert scrutiny, etc, yet the context type (real data that cites a real study) was preserved for the BLIS and MBLIS forms. The internal anchor design (Livingston, 2004) of the assessment forms mentioned by the former is a kind of passive control since six identical items appear verbatim on both forms of the assessment. Neither resulted in conclusive evidence of a statistically noticeable difference between the students shown the BLIS form and the students shown the MBLIS form, yet the story was somewhat different when tasks involving other context types present in the BLIS assessment were reframed to incorporate real data that applies to a real study for the purpose of the MBLIS form.

Expert reviews and comparison of reliability and validity analysis provided favorable evidence that the development of corresponding tasks to appear on the MBLIS form successfully preserved the form and function of the tasks as structural isomorphs. Lehrer and Schauble (2007) and Fay et al. (2018), however, caution that use of analogical or isomorphic tasks cannot guarantee that respondents’ cognitive processes in answering these tasks will be comparable. Replication and additional study is needed to substantiate these results or investigate the mechanism underlying the noteworthy difference in performance associated with a change to the context type between structurally isomorphic tasks on the two assessment forms.

For a statistically literate individual, the ability to merge understanding of statistical constructs with the context at hand is essential. In fact, as there is no statistics without context (Rao, 1975). Statistical literacy is also inherently contextualized and students report high satisfaction with learning statistical concepts that are closely and explicitly linked to real data with real contexts (Brown, 2016). To be clear, the authors of this paper remain strong proponents of the GAISE (2016) recommendation in favor of real data with context and purpose. However, the transfer of statistical skills to new contexts is non-trivial. This work highlights a need for additional study to better understand how contexts may factor into all aspects of teaching and learning with and about data, not just for assessment. Future work may also consider the type or genre of context that would be best suited for a particular purpose.

4.1. LIMITATIONS

The challenge of open-access to raw datasets accompanying candidate studies is a significant hurdle to the goal of developing structurally isomorphic tasks that use real data from real (nearly always published) studies. In general, most challenging were those BLIS items that required students to directly engage with raw quantitative data, data from a randomized experiment, or nuanced interpretation of specific features of data. Noteworthy concessions were made for three items, in particular, along these lines. For one such item, the parameter of interest was switched from a mean to a proportion, and an observational study was discussed instead of a randomized experiment as a concession for a second item. As seen in the example in Table 2 reverse skewness was accepted for a third item.

The objective to retain the structural integrity of item phrasing and the statistical idea in the isomorph had to be loosened for two items, in consultation with the expert reviewers. One of those items (Table 4) was a subject of lengthy discussions some of which included the expert reviewers. The implicit assumption of a coin being unbiased and our intuition about 50% of them landing on heads benefitted the original item. Upon deliberation, it was agreed that it is extremely hard to find other phenomena with an unconditional 0.5 probability of occurrence that is understood intuitively, and therefore the substantial change in wording was included. The original item was an interesting case because students are assumed to be so familiar with fair coins that the frequency of their “encounters” with the context might actually outweigh the other dimensions of engagement/relevance we are seeking in this study.

The authors also acknowledge that even though we used anchor items to compare the two sets of respondents at baseline, we had to account for possible ordering effect. These identical items could function differently across BLIS and MBLIS, especially since they may appear out-of-context on an assessment based entirely on one unified theme for the wide majority of item contexts (e.g., COVID-19 pandemic).

Additionally, balancing the competing goals of maximizing engagement and minimizing emotional impact lead to the inclusion of some topics that may not be most relevant to the lives of our target population for the study—college students, in this case—and exclusion of some topics that may be directly related to them. For example, one of the modified items referred to performance of elementary school students on standardized tests both before, and during, the COVID-19 pandemic. This issue is confounded by the expectations of the “college student” audience, which is typical to an educational research study, though that may not need to be the case for the general purpose of the research. The choice of the test population may bias the choice of contexts.

In some cases, the expert review panel—including the author of the BLIS assessment—suggested that one of the original BLIS tasks might benefit from improvement. Since this project was not intended to develop a new assessment tool or refine the BLIS assessment, we decided that changes to BLIS were out of scope for this study. While this decision has the benefit of allowing the present study to achieve a closer comparison to the published version of the BLIS assessment, it forfeits the opportunity to improve the BLIS form prior to administration in the study. For example, the two items with highest difficulty (consistent across the two forms) are the two items for which respondents chose an incorrect option most frequently. These items, however, have negative correlations with the total score without accounting for the given item on both assessments. This reverse discrimination indicates a possible flaw in the original item design.

Finally, survey questions were asked at the end of the assessment. Therefore, we did not expect that students’ performance on the assessment would be affected by these. However, responses to the survey questions may have contained some cognitive bias based on whether the participants had just seen an assessment based on COVID-19 or not.

4.2. IMPLICATIONS FOR FUTURE WORK

Since the instruments are broadly observed to function comparably, we argue that isomorphic assessment can be created to assess statistical literacy in various pertinent contexts. Even though it may be quite tedious to create them, these instruments can be valuable tools in getting respondents to consider statistics through a contextual lens that is perhaps more personally relevant to them, and continue to measure how curricular strategies may affect statistical literacy levels. Therefore, future research can be directed towards two purposes. 1) measurement of statistical literacy in various disciplinary or societal contexts using isomorphs of BLIS, and 2) using these isomorphic assessment forms to support design-based research (e.g., Cobb et al., 2003). As well, additional work exploring the transfer and cognitive processes behind statistical problem solving will be essential for better understanding the role of context.

This work reports psychometric properties of MBLIS in comparison with BLIS to determine whether the BLIS and MBLIS are psychometrically similar, even when the item contexts are changed. To draw reliable conclusions, it was a high priority that we made a concerted effort to preserve the ability to compare results from our study to the field test conducted during the development of the BLIS assessment. To achieve this, it was important to ensure that the BLIS items remained identical to the version implemented and evaluated previously, and therefore, the MBLIS form was closely aligned to that version. At no point did we change any details in the original assessment as an effort to ensure comparability across the original work (Ziegler, 2014) and our study. The results from this paper are specific to one definition and assessment of statistical literacy. Future research should study the role of contexts using other assessment instruments.

Differential student performance on BLIS and MBLIS forms with a low p -value on inferential results indicates that the context in which a statistical question is posed is associated with an effect on student responses during an assessment. While the empirical results in this study suggested that students found tasks more challenging when they were revised to include real data from real contexts, it must be noted that the result may in fact be confounded with the choice of contexts present in this study. It would be quite possible that the unifying theme among the contexts (e.g., collateral impacts of the COVID-19 pandemic) is associated with the observed effect as opposed to a broad commentary on *any* use of real data from real studies during similar assessments.

In reference to the discussion in Section 2.3 regarding sensitive contexts, this finding also has implications for teaching practices. If additional research finds that the sensitivity of the topic may have

contributed to the lower scores on MBLIS, an argument can be made to favor inclusion of such topics on curricular materials instead of including them in grade-affecting assessments (Fallstrom et al., 2021).

ACKNOWLEDGEMENTS

Work supported, in part, by Penn State's Centre for Social Data Analytics Accelerator Award Program Seed Grant. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect those of the sponsor. We wish to acknowledge the three expert reviewers of MBLIS who provided valuable input. We wish to also thank Laura Ziegler for her support during the adaptation of her original work.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. <https://www.apa.org/science/programs/testing/standards>
- Barbieri, G. A., & Giacché, P. (2006). The worth of data: The tale of an experience for promoting and improving statistical literacy. http://iase-web.org/documents/papers/icots7/1A1_BARB.pdf
- Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Sheri Johnson, S., Perez, L., & Spangler, D. A. (2020). *Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II): A Framework for Statistics and Data Science Education*. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf
- Barniol, P., & Zavala, G. (2014). Force, velocity, and work: The effects of different contexts on students' understanding of vector concepts using isomorphic problems. *Physical Review Special Topics: Physics Education Research*, 10(2), Article 020115. <https://doi.org/10.1103/PhysRevSTPER.10.020115>
- Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(1), 153–166. <https://doi.org/10.1037/0278-7393.15.1.153>
- Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Ben-Zvi, D., & Garfield, J. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning, and thinking*. Kluwer Academic Publishers.
- Ben-Zvi, D., & Garfield, J. (2008). Introducing the emerging discipline of statistics education. *School Science and Mathematics*, 108(8), 355–361. <https://doi.org/10.1111/j.1949-8594.2008.tb17850.x>
- Ben-Zvi, D., Makar, K., & Garfield, J. (Eds.). (2018). *International handbook of research in statistics education*. Springer. <http://link.springer.com/10.1007/978-3-319-66195-7>
- Brown, M. (2016). Engaging students in quantitative methods: Real questions, real data. In *Promoting understanding of statistics about society. Proceedings of the roundtable conference of the International Association of Statistics Education (IASE)*. ISI/IASE.
- Carmichael, C. S. (2010). *The development of middle school children's interest in statistical literacy*. [Doctoral dissertation, University of Tasmania].
- Carotenuto, G., Di Martino, P., & Lemmi, M. (2021). Students' suspension of sense making in problem solving. *ZDM Mathematics Education*, 53, 817–830.
- Cobb, P., Confrey, J., DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32(1), 9–13.
- Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal*, 16(1), 44–49. <https://doi.org/10.52041/serj.v16i1.213>
- Fallstrom, S., Firouzian, S., Kubo, K., & Peck, R. (2021). *Increasing student engagement at two-year colleges using socially relevant contexts*. <https://www.causeweb.org/cause/uscots/uscots21/workshop/12-2>
- Fay, D. M., Levy, R., & Mehta, V. (2018). Investigating psychometric isomorphism for traditional and performance-based assessment. *Journal of Educational Measurement*, 55(1), 52–77.
- Ferligoj, A. (2015). How to improve statistical literacy? *Metodoloski Zvezki*, 12(1), 1–10.
- GAISE College Report ASA Revision Committee. (2016). *Guidelines for Assessment and Instruction*

- in *Statistics Education (GAISE) College Report 2016*. AMSTAT. <http://www.amstat.org/education/gaise>
- Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, 70(1), 1–25. <https://doi.org/10.1111/j.1751-5823.2002.tb00336.x>
- Gal, I. (2019). Understanding statistical literacy: About knowledge of contexts and models. *Actas Del Tercer Congreso Internacional Virtual de Educación Estadística*. <http://digibug.ugr.es/bitstream/handle/10481/55029/gal.pdf?sequence=1&isAllowed=y>
- Garfield, J., del Mas, 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: An International Perspective* (pp. 75–86). Wiley.
- Garfield, J., DelMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM Mathematics Education*, 44(7), 883–898. <https://doi.org/10.1007/s11858-012-0447-5>
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12(3), 306–355. [https://doi.org/10.1016/0010-0285\(80\)90013-4](https://doi.org/10.1016/0010-0285(80)90013-4)
- Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, 16(1), 22–25. <https://doi.org/10.52041/serj.v16i1.209>
- Kusairi, S., Hidayat, A., & Hidayat, N. (2017). Web-based diagnostic test: Introducing isomorphic items to assess students' misconceptions and error patterns. *Chemistry: Bulgarian Journal of Science Education*, 26(4).
- Kusairi, S., Puspita, D. A., Suryadi, A., & Suwono, H. (2020). Physics formative feedback game: Utilization of isomorphic multiple-choice items to help students learn kinematics. *TEM Journal*, 9(4), 1625–1632. <https://doi.org/10.18421/TEM94-39>
- Lee, H., & Tran, D. (2015). Statistical habits of mind. *Teaching Statistics through Data Investigations MOOC-Ed*. Friday Institute for Educational Innovation. <https://www-data.fi.ncsu.edu/wp-content/uploads/2020/12/12125035/Habitsofmind.pdf>
- Lehrer, R., & Schauble, L. (2007). A developmental approach for supporting the epistemology of modeling. In W. Blum, P. L. Galbraith, H.-W. Henn, & M. Niss (Eds.), *Modeling and applications in mathematics education* (14th ed., pp. 153–160). Springer.
- Lemon, J., (2006). Plotrix: A package in the red light district of R. *R-News*, 6(4), 8–12.
- Lin, S.-Y. Y., & Singh, C. (2011). Using isomorphic problems to learn introductory physics. *Physical Review Special Topics: Physics Education Research*, 7(2), Article 20104. <https://doi.org/10.1103/PhysRevSTPER.7.020104>
- Livingston, S. A. (2004). *Equating test scores (Without IRT)*. Educational Testing Service.
- Lovett, M. C., & Greenhouse, J. R. (2000). Applying Cognitive Theory to Statistics Instruction. *American Statistician*, 54(3), 196–206. <https://doi.org/10.1080/00031305.2000.10474545>
- Luger, G. F., & Bauer, M. A. (1978). Transfer effects in isomorphic problem situations. *Acta Psychologica*, 42(2), 121–131. [https://doi.org/10.1016/0001-6918\(78\)90011-2](https://doi.org/10.1016/0001-6918(78)90011-2)
- Millar, R., & Manoharan, S. (2021). Repeat individualized assessment using isomorphic questions: a novel approach to increase peer discussion and learning. *International Journal of Educational Technology in Higher Education*, 18(1). <https://doi.org/10.1186/s41239-021-00257-y>
- Parker, M. C., Guzdial, M., & Engleman, S. (2016). Replication, validation, and use of a language independent CS1 knowledge assessment. *ICER 2016. Proceedings of the 2016 ACM Conference on International Computing Education Research, Melbourne, Australia*, pp. 93–101. <https://doi.org/10.1145/2960310.2960316>
- Pearl, D. K., Garfield, J. B., DelMas, R. C., Groth, R. E., Kaplan, J. J., McGowan, H., & Lee, H. S. (2012). *Connecting research to practice in a culture of assessment for introductory college-level statistics*. https://www.amstat.org/docs/default-source/amstat-documents/researchreport_dec_2012.pdf
- R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Rao, C. R. (1975). Teaching of statistics at the secondary level An interdisciplinary approach. *International Journal of Mathematical Education in Science and Technology*, 6(2), 151–162.
- Ratnawati, O. A., Siswono, T. Y. E., & Wijayanti, P. (2020). Statistical literacy comprehension of students in the context of COVID-19 with collaborative problem solving (CPS). *Math Didactic*:

- Jurnal Pendidikan Matematika*, 6(3), 264–276.
- Sabbag, A., Garfield, J., & Zieffler, A. (2018). Assessing statistical literacy and statistical reasoning: The REALI instrument. *Statistics Education Research Journal*, 17(2), 141–160. <https://doi.org/10.52041/serj.v17i2.163>
- Sanchez, J. (2007). Building statistical literacy assessment tools with the IASE/ISLP. In B. Phillips & L. Weldon (Eds.), *Assessing student learning in statistics. IASE/ISI Satellite Conference*. <https://iase-web.org/documents/papers/sat2007/Sanchez.pdf>
- Schild, M. (2004). Statistical literacy curriculum design. In G. Burrill & M. Camden (Eds.), *Curricular development in statistics education. International Association for Statistical Education Roundtable* (pp. 54–74).
- Sharma, S. (2017). Definitions and models of statistical literacy: A literature review. *Open Review of Educational Research*, 4(1), 118–133. <https://doi.org/10.1080/23265507.2017.1354313>
- Suganda, T., Kusairi, S., Azizah, N., & Parno, P. (2020). The correlation of isomorphic, open-ended, and conventional score on the ability to solve kinematics graph questions. *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, 6(2), 173–180. <https://doi.org/10.21009/1.06204>
- Suhermi, & Widjajanti, D. B. (2020). What are the roles of technology in improving student statistical literacy? *Journal of Physics: Conference Series, the 3rd International Seminar on Innovation in Mathematics and Mathematics Education*, 1581, Article 012067. <https://doi.org/10.1088/1742-6596/1581/1/012067>
- Utts, J. (2021). Enhancing data science ethics through statistical education and practice. *International Statistical Review*, 89(1), 1–17. <https://doi.org/10.1111/insr.12446>
- Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, 88(421), 1–8. <https://doi.org/10.1080/01621459.1993.10594283>
- Watson, J. M. (1998). The role of statistical literacy in decisions about risk: Where to start. *For the Learning of Mathematics*, 18(3), 25–27.
- Watson, J. M. (2011). Foundations for improving statistical literacy. *Statistical Journal of the IAOS*, 27, 197–204. <https://doi.org/10.3233/SJI-2011-0728>
- Watson, J. M., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2(2), 3–46. <https://doi.org/10.52041/serj.v2i2.553>
- Weiland, T. (2017). Problematizing statistical literacy: An intersection of critical and statistical literacies. *Educational Studies in Mathematics*, 96(1), 33–47. <https://doi.org/10.1007/s10649-017-9764-5>
- Wickham, H., (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.
- Wilks, S. S. (1951). Presidential address. *Journal of the American Statistical Association*, 46(253), 1–18. <https://www.causeweb.org/cause/resources/library/r1266/>
- Williamson, D. M., Johnson, M. S., Sinharay, S., & Bejar, I. I. (2002). *Hierarchical IRT examination of isomorphic equivalence of complex structured response tasks*.
- Ziegler, L. (2014). *Reconceptualizing statistics literacy: Developing an assessment for the modern introductory statistics course* [University of Minnesota]. <http://hdl.handle.net/11299/165153>
- Ziegler, L., & Garfield, J. (2018). Developing a statistical literacy assessment for the modern introductory statistics course. *Statistics Education Research Journal*, 17(2), 161–178. <https://doi.org/10.52041/serj.v17i2.164>

SAYALI PHADKE
Pennsylvania State University
4655 Victoria Way,
Erie, PA 16509
USA

APPENDIX

ASSESSMENT RESPONSE SUMMARIES

Table 8 captures the percentage of respondents who answered each item correctly and the taxa of context on the BLIS item. The difference column is calculated as percentage of respondents who correctly answered the item on BLIS minus the percentage of respondents who correctly answered the corresponding item on MBLIS. The highlighted items with an * (e.g., 13*) identify anchor items critical in comparing the two groups of respondents at baseline.

Table 8. Difference in proportion of respondents correctly answering each item

Item	BLIS	M-BLIS	Difference	BLIS context - GAISE
1	74.6	73.2	1.4	Real from real study
2	44.0	50.7	-6.7	Realistic
3	53.4	52.5	0.9	Real
4	83.5	86.2	-2.7	Real
5	81.3	84.7	-3.4	Realistic
6	73.5	70.7	2.8	Realistic
7	35.6	41.1	-5.5	Real from real study
8	29.5	32.8	-3.3	Realistic
9	65.4	34.0	31.4	Realistic
10	56.3	39.2	17.1	Realistic
11	42.0	37.1	4.9	Real
12	58.3	48.8	9.5	Real from real study
13*	37.6	37.6	0.0	Real from real study
14	42.8	24.6	18.2	Naked
15	63.8	48.5	15.3	Realistic
16*	24.6	27.8	-3.2	Realistic
17*	46.1	46.8	-0.7	Realistic
18	45.9	45.4	0.5	Real from real study
19	40.8	38.4	2.4	Real from real study
20	37.9	34.3	3.6	Real from real study
21	16.5	16.3	0.2	Real from real study
22	58.5	61.0	-2.5	Realistic
23*	43.4	43.9	-0.5	Real from real study
24*	57.2	60.0	-2.8	Real from real study
25	55.5	61.6	-6.1	Real from real study
26	42.2	42.0	0.2	Real from real study
27	38.6	45.0	-6.4	Realistic
28	52.7	60.2	-7.5	Real from real study
29	52.0	48.9	3.1	Real from real study
30	48.3	45.5	2.8	Real from real study
31	86.4	83.9	2.5	Realistic
32*	48.0	43.6	4.4	Real from real study
33	64.4	62.0	2.4	Real from real study
34	70.4	65.2	5.2	Real
35	23.4	21.6	1.8	Real from real study
36	79.0	68.6	10.4	Real
37	57.8	54.3	3.5	Real from real study

Table 9 is the selected response table indicating the percentage of respondents who selected each possible distractor. Item numbers with an * in the highlighted rows indicate anchor items. Percentage values with an * next to them specify the correct response.

Table 9. Selected-response table

Item	BLIS				MBLIS			
	A	B	C	D	A	B	C	D
1	16.6	8.8	74.6*	NA	19.3	7.5	73.2*	NA
2	7.4	15.8	44*	32.8	8.6	18	50.7*	22.6
3	53.4*	20.5	7.2	18.8	52.5*	8.6	13	25.9
4	15.7	83.5*	0.8	NA	13	86.2*	0.8	NA
5	81.3*	16.5	2.2	NA	84.7*	13.8	1.5	NA
6	3.4	10.8	73.5*	12.2	4.7	15.8	70.7*	8.8
7	32.6	15.7	16.1	35.6*	21	17.1	20.8	41.1*
8	39.8	30.7	29.5*	NA	23.7	43.4	32.8*	NA
9	65.4*	30.7	3.9	NA	34*	56.4	9.6	NA
10	14.7	9.1	19.9	56.3*	21.6	21.5	17.7	39.2*
11	19.1	38.9	42*	NA	18.5	44.4	37.1*	NA
12	13.9	22.7	5	58.3*	11.9	31.1	8.3	48.8*
13*	37.6*	40.3	8.5	13.6	37.6*	35	14.1	13.3
14	5.8	42.8*	50.3	1.1	2.8	69.1	24.6*	3.6
15	5.6	8.9	21.6	63.8*	6	18.2	27.3	48.5*
16*	24.6*	19.9	36.1	19.4	27.8*	42.4	29.8	NA
17*	25.2	28.7	46.1*	NA	23.7	29.4	46.8*	NA
18	35.3	45.9*	18.8	NA	29.9	45.4*	24.7	NA
19	9.7	29.8	19.7	40.8*	11.1	27.6	22.9	38.4*
20	20.7	29	37.9*	12.4	17.9	34.8	34.3*	13
21	16.5*	8.8	39.3	35.4	16.3*	14.1	36.7	32.8
22	29.6	58.5*	11.9	NA	26.8	61*	12.2	NA
23*	12.2	29.8	43.4*	14.6	10.4	31.1	43.9*	14.6
24*	57.2*	25.4	17.4	NA	60*	22.4	17.6	NA
25	55.5*	25.9	18.7	NA	61.6*	28	10.4	NA
26	42.2*	42.5	15.4	NA	42*	42.3	15.8	NA
27	31.7	38.6*	18.8	11	27.5	45*	14.6	12.8
28	20.5	26.8	52.7*	NA	17.2	22.6	60.2*	NA
29	18	17.4	52*	12.5	13.3	20.2	48.9*	17.6
30	10	19.9	21.8	48.3*	9.6	22.4	22.4	45.5*
31	8.2	86.4*	5.5	NA	9.9	83.9*	6.2	NA
32*	19.7	20.8	48*	11.4	20	23.6	43.6*	12.8
33	64.4*	22.1	13.5	NA	62*	25.5	12.5	NA
34	25.4	70.4*	4.2	NA	31.2	65.2*	3.6	NA
35	23.4*	11.4	13	52.2	21.6*	16.9	15.4	46
36	79*	16.6	4.4	NA	68.6*	23.7	7.6	NA
37	16.6	20.7	57.8*	4.9	19.8	20.8	54.3*	5