ASSESSING STATISTICAL LITERACY AND STATISTICAL REASONING: THE REALI INSTRUMENT

ANELISE SABBAG
California Polytechnic State University
asabbag@calpoly.edu

JOAN GARFIELD
University of Minnesota
jbg@umn.edu

ANDREW ZIEFFLER
University of Minnesota
zief0002@umn.edu

ABSTRACT

Statistical literacy and statistical reasoning are important learning goals that instructors aim to develop in statistics students. However, there is a lack of clarity regarding the relationship among these learning goals and to what extent they overlap. The REasoning and Literacy Instrument (REALI) was designed to concurrently measure statistical literacy and reasoning. This paper reports the development process of the REALI assessment, which included test blueprint, expert review, item categorization, pilot and field testing, and data analysis to identify what measurement model best represents the constructs of statistical literacy and reasoning given the criteria of fit and parsimony. The results suggested that statistical literacy and reasoning can be measured effectively by the REALI assessment with high score precision.

Keywords: *Statistics education research; Assessment; Statistical learning goals; Evaluation of statistical knowledge*

1. INTRODUCTION

Assessments are used in research for many different purposes: to facilitate student learning, to provide feedback for students, to inform instructors regarding students' achievement, and to evaluate courses. National organizations such as the American Statistical Association (ASA, 2007) and a joint publication by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, NCME, 1999) have outlined several suggestions for developing and improving instruments.

The role of assessment in the field of statistics education is intrinsic not only in the learning but also the teaching of statistics. The literature includes arguments against the use of final exam scores or course grades as indicators of statistical reasoning (e.g., Chance & Garfield, 2002; Konold, 1995). Despite this information, Zieffler, Garfield, Alt, Dupuis, Holleque, and Chang (2008) indicate that many studies still use these measures.

Assessments can provide important information related to students' learning, but it is important to use quality instruments to capture this information. In a report published by the American Statistical Association (ASA, 2007) the authors suggested that every assessment should develop and report (1) information about the construct that is measured by the assessment, how the construct is aligned with the desired learning goals, and the limitations of the instrument; (2) information regarding the population of interest to which the assessment will be administered, the circumstances of administration or implementation of the assessment, and ways in which these are similar to or different from the setting in which published validity, reliability, and fairness evidence (if any) were obtained; and (3) evidence

Statistics Education Research Journal, 17(2), 141–160, http://www.stat.auckland.ac.nz/serj

of validity, reliability, and fairness that is specific to the setting in which the assessment is administered, the particular population to which it is administered, the way it is scored, and the use to which the scores are put.

Despite the challenges present in the assessment area in the field of statistics education, instruments assessing the current learning goals for introductory statistics courses have been developed. Statistical literacy and statistical reasoning (along with statistical thinking) have been considered by the statistics education community as important learning goals to be developed in introductory statistics courses (Garfield & Ben-Zvi, 2008). Many statistics educators and scholars have tried to define these learning goals. However, there is a lack of agreement regarding these definitions and the relationship between statistical literacy and statistical reasoning.

The purpose of this study was to develop an assessment tool to examine and clarify the relationship between statistical literacy and statistical reasoning. Determining whether these constructs are unique is important for many reasons. If statistical literacy and reasoning are not distinct constructs, then the idea of having two separate learning goals and, subsequently, different assessments for these constructs should be re-evaluated. A new assessment is needed that concurrently measures both outcomes in order to help clarify the structure of the relationship between statistical literacy and reasoning. The focus of this study is to describe the development of such an instrument. The next section provides a brief review of the literature regarding definitions and assessment of statistical literacy and reasoning. Section 3 describes the development phases of a new instrument designed to concurrently measure statistical literacy and reasoning. Section 4 reports the results of psychometric analyses of the instrument, and Section 5 provides a discussion regarding item categorization, what can be learned about the psychometric properties of the assessment, limitations, and implications for future research.

2. DEFINING AND ASSESSING STATISTICAL LITERACY AND STATISTICAL REASONING

Many statistics educators and scholars have tried to define and describe statistical literacy (e.g., Budgett & Pfannkuch, 2007; Gal, 2002; Rumsey, 2002; Watson & Callingham, 2003) and statistical reasoning (Garfield, 2002; Garfield & Ben-Zvi, 2008; Garfield & Chance, 2000; Jones, Langrall, Mooney, & Thornton, 2004). These topics are of such importance that the *Statistics Education Research Journal* (*SERJ*) published a special issue (vol. 16, 2017) with eight research papers focused on statistical literacy. However, no consensus has been reached regarding the definitions of these terms. For more information about the definitions of statistical literacy and reasoning see Sabbag (2016). The lack of consistency in the different definitions supports the idea that these concepts are still evolving. In addition, there seems to be a great overlap in the definitions of these terms and assumptions of a hierarchy between and within these learning goals has been posed by some researchers (Chance, 2002; delMas, 2002; Garfield & Ben-Zvi, 2007, 2008; Jones et al., 2004). However, no empirical study has been carried out to examine the relationship between these learning goals.

A review of the literature on assessing statistical literacy and reasoning suggests a lack of clarity regarding whether statistical literacy and statistical reasoning are two distinct constructs or whether literacy is wholly or partially subsumed in reasoning. There is a need, therefore, for a new assessment instrument that can concurrently measure both outcomes to help examine their relationship.

2.1. INSTRUMENTS TO ASSESS STATISTICAL LITERACY AND REASONING

Garfield and Ben-Zvi (2008) argued for changes in how students are assessed in light of the broader adoption of statistical literacy and reasoning as learning outcomes for introductory statistics students. To date, four instruments are described in the literature that have been developed and used to assess statistical literacy and statistical reasoning. These are described below.

1. The *Statistical Reasoning Assessment* (SRA) was developed as part of the ChancePlus Project (Garfield, 1991) and funded by the National Science Foundation (NSF Grant MDR-8954626) to evaluate a computer-based statistics curriculum. The SRA is composed of 20 forced-choice items that cover specific types of reasoning and misconceptions related to data, representations of data, statistical measures, uncertainty, sampling, association, and probability (Garfield, 1998, 2003).

2. The *Comprehensive Assessment of Outcomes in a First Statistics Course* (CAOS; delMas, Garfield, Ooms, & Chance, 2007) was designed to assess students' statistical reasoning after taking an introductory statistics course. This instrument is composed of 40 forced-choice items that assess students' conceptual understanding of data collection and design, descriptive statistics, graphical representations, boxplots, normal distribution, bivariate data, probability, sampling variability, confidence intervals, and tests of significance. Although the CAOS items were written to assess students' reasoning involving topics typically covered in an introductory course, primarily the items focused on assessing variability (Garfield, delMas, & Chance, n.d.).

3. The *Goals and Outcomes Associated with Learning Statistics* (GOALS; Sabbag & Zieffler, 2015) instrument was developed to assess important statistical reasoning outcomes in a first course of statistics. GOALS is composed of 20 forced-choice items that address the topics of study design, bivariate relationships, variability, sampling and sampling variability, interpreting confidence intervals and *p*-values, statistical inference, and modeling and simulation.

4. The *Basic Literacy in Statistics* (BLIS; Ziegler, 2014) assessment was created to measure students' statistical literacy, which is defined by Ziegler as the "ability to read, understand, and communicate statistical information" (p. 18). Her definition of statistical literacy was based on the idea that statistical literacy represents one of three cognitive levels, followed by *statistical reasoning* and *statistical thinking* as supported by Garfield and Ben-Zvi (2007), Garfield and delMas (2010), Garfield, delMas, & Chance (2003), and Garfield and Franklin (2011). Ziegler's definition was also based on the more general definition of literacy as the ability to read and write. The BLIS instrument is composed of 37 forced-choice items assessing data production, graphs, descriptive statistics, empirical sampling distributions, confidence intervals, randomization distributions, hypothesis tests, scope of conclusions, and regression and correlation (see also Zieler, 2018).

Research on these four instruments presented some evidence of content validity, response process validity, and internal structure validity. However, the SRA, CAOS and GOALS assessments were not written using a clear working definition of statistical reasoning. Therefore, there is a lack of clarity regarding the relationship between these assessments' content and the construct being measured (statistical reasoning). Although these tests are regarded as measuring important statistical concepts in introductory statistics courses, it is not clear how these concepts are related to the constructs of statistical reasoning and literacy. The next section of this paper describes the development process of a new instrument to help examine the relationship between these two learning goals.

3. DEVELOPING A NEW INSTRUMENT TO ASSESS STATISTICAL LITERACY AND REASONING

To investigate the degree of distinction between statistical literacy and statistical reasoning, a new instrument was created, composed of items measuring statistical literacy and items measuring statistical reasoning. This assessment was christened REALI (*REasoning and Literacy Instrument*). The development process for REALI comprised six steps: establishing working definitions of statistical literacy and statistical reasoning, developing a blueprint of the new assessment, gathering expert review, conducting think-aloud interviews with students, conducting a pilot test and a field test, and evaluating the characteristics of the instrument. Each of these is described below.

3.1. WORKING DEFINITIONS

In order to differentiate statistical literacy from statistical reasoning, working definitions for both constructs were developed. The definitions were focused around items because, ultimately, they were going to be used to categorize items. Ziegler (2014) extensively explored the learning goal of statistical literacy and this study builds on her research. The working definition for statistical literacy was based on the definition from Ziegler (2014): *Statistical literacy items assess students' ability to recall a definition, describe or interpret basic statistical information. Items at this level usually address a single statistical concept. If multiple statistical concepts are addressed, the item will not require that students make connections between them (recall information will be sufficient).* The working definition for statistical reasoning was based on definitions from Garfield and Ben-Zvi (2008) and delMas (2002, 2004): *Statistical reasoning items assess students' ability to make connections among statistical* *concepts, create mental representations of statistical problems, and explain relationships between statistical concepts. Items at this level usually address more than one statistical concept and require making connections between them. Because of the number of concepts addressed, statistical reasoning items require higher order thinking and higher cognitive load than statistical literacy items.*

It is important to note that, statistical literacy and reasoning items as defined in this paper also contemplate students' ability to be critical of statistical information. Graphs and descriptive statistics are often reported in the media and it is expected that students will be able to interpret and critically evaluate them. The main difference is that statistical literacy items might involve critiquing information that addresses only one statistical concept. On the other hand, if statistical information relates two or more statistical concepts, a student would first need to be able to interpret each statistical concept and then make connections between them to critically evaluate this statistical information and make data-related arguments. This would be an example of a statistical reasoning item that requires students to be critical of statistical information. Disagreement in terms of the definitions is expected and will most likely lead to healthy discussions in the field of statistics education.

3.2. BLUEPRINT

Because REALI was intended for use with introductory statistics students, content focused on topics related to learning goals for introductory students. Overall, REALI encompasses eight content areas: (1) representations of data, (2) measures of center, (3) measures of variability, (4) study design, (5) confidence intervals, (6) hypothesis testing & *p*-values, (7) probability, and (8) bivariate data.

After selecting the content to be covered in REALI, the next goal was to identify the number of literacy and reasoning items to be allocated to each content area. This was done with the constraints that (1) the overall number of items needed to be small enough for students to complete in a typical class period and (2) the number of literacy and reasoning items needed to be balanced. It was ultimately decided that REALI would contain a total of 40 items. This decision was based on the advice of Sinharay (2010), who suggested that for subscores to have added value beyond the total scores, they need to be composed of at least 20 items.

Item selection and categorization. Items for REALI were initially selected by identifying items from existing instruments, namely BLIS and GOALS, that were related to content in REALI's blueprint. Table 1 shows the number of items from BLIS and GOALS for each of the eight content areas. From this, it was apparent that the content areas of *hypothesis testing & p-values* (33%) and *study design* (23%) had the highest proportion of items. The list of topics assessed by the items from both BLIS and GOALS takes into account the learning goals of introductory statistics courses and topics that were emphasized in introductory statistics books (Sabbag & Zieffler, 2015; Ziegler, 2014). Therefore, it can be argued that these content areas might be considered more important than others. Consequently, we decided to include more items from these two content areas on REALI, and fewer items for the remaining topics. The last three columns of Table 1 show the targeted number of REALI items by content area.

Table 1. Number of BLIS and GOALS items, percentage for each area of learning and target for number of items in REALI

				Target		
Content area	BLIS items	GOALS items	Total	Statistical Literacy	Statistical Reasoning	Total
Representations of data	4	1	5 (9%)	2	2	4
Measures of center	2	1	3 (5%)	2	2	4
Measures of variability	3	2	5 (9%)	2	2	4
Study design	10	3	13 (23%)	3	3	6
Confidence intervals	4	3	7 (12%)	2	2	4
Hypothesis testing & *p*-values	10	9	19 (33%)	5	5	10
Probability	2	0	2 (4%)	2	2	4
Bivariate Data	2	1	3 (5%)	2	2	4
Total	37	20	57	20	20	40

Items from BLIS and GOALS were then categorized as literacy or reasoning items according to the working definitions. To make this classification, it was necessary to identify the behaviors, abilities, and student understanding needed to answer each item correctly. Two example items (statistical literacy item and statistical reasoning item) and the behaviors needed to answer them are given in Figure 1.

ITEM 1: The Pew Research Center surveyed 2,076 American adult cell phone users chosen at random in 2013. The sample percent of adult cell phone users who access the internet on their phones was 60%. The 95% confidence interval was 58% to 62%. What is this interval attempting to estimate?

a) The average number of American adult cell phone users who access the internet on their phones in 2013.
b) The percent of the 2,076 American adult cell phone users who access the internet on their phones in 2013.
c) The percent of all American adult cell phone users who access the internet on their phones in 2013.
d) For American adults who access the internet on their cell phones, only 58% to 62% were confident in using the internet on their phones.

BEHAVIORS: To answer the item above correctly, students need to

1. Understand what a confidence interval represents.
2. Recognize which parameter is being estimated.
3. Recognize the population of interest.
4. Understand what the level of confidence represents

ITEM 2: In 2011, it was reported that the mean home price in the Hamptons (New York) increased by 20% within a single year, while the median home price decreased by 2% during that same year. Which of the following is the best explanation for this occurrence?

a) The price of most homes in the Hamptons decreased and more homes were sold in the Hamptons that year.
b) The reporters made an error in presenting the results; if the mean home price increases, the median home price must also increase.
c) Most of the homes in the Hamptons decreased in price and a small number of homes had large increases in price.

BEHAVIORS: To answer the item above correctly, students need to

1. Identify factors that affect mean and median
2. Recognize that the median is more resistant to extreme values than the mean
3. Infer about the relationship between mean and median with no visual representation.

Figure 1. Statistical literacy item, statistical reasoning item, and behaviors

Based on the behaviors identified for each item and on the working definitions, each item was categorized as a statistical literacy item or a statistical reasoning item. At the end of the classification, two items from BLIS were classified as statistical reasoning items and four items from GOALS were classified as statistical literacy items. Therefore, these items were classified by the first author as different learning goals that they were initially designed to measure. None of these six items were included in the REALI assessment.

To meet the item targets, some items—those in content areas that had more existing items than were targeted—were deleted. Decisions about which items to delete were based on psychometric information (item discrimination and item difficulty) available from previous analyses and the alignment of the content addressed by the item and the content of the other items included in the instrument. For instance, some items measured very similar content so items with the worst item characteristics were deleted. Additionally, one item's content was modified to be more conducive to assessment, regardless of the specific instruction received.

After the process of categorization and verification of items, only the *measures of variability* and *confidence intervals* content areas had the targeted number of items. Therefore, additional items were written to round out the remaining six content areas. These items were modified from other instruments (e.g., CAOS, AIRS; Park, 2012), from the ARTIST Topic Scale item bank, and from materials used in introductory statistics courses at the University of Minnesota. Ultimately, 12 more items than necessary were written so that the item pool could be culled based on psychometric information gained from the pilot testing process. This first draft of the REALI assessment—27 statistical literacy items and 25 statistical reasoning items—was used in the expert review which we explain in the next section.

3.3. EXPERT REVIEW AND CATEGORIZATION

To help validate the categorization of items, an expert review was conducted in two phases. The intent of the first phase was to observe how reviewers (four statistics faculty from the University of Minnesota and one from Cleveland State University) would categorize the items and also to verify potential issues in the categorization process. The second phase of the expert review was conducted with four experts (Maxine Pfannkuch, Dani Ben-Zvi, Jane Watson, and Robert Gould) in the field of statistics education, all who had worked in or were interested in the domains of statistical literacy, statistical reasoning, or statistical thinking. The experts were sent a review form that included the working definitions, referred to as "Group 1" and "Group 2" and asked to categorize the 52 REALI items into one group or the other. (Groups 1 and 2 were used rather than "statistical literacy" and "statistical reasoning" to reduce some of the personal biases thought to be related to the latter terms.)

After categorization, the percentage of experts whose categorization was the same as that determined by us was calculated for each item. Out of the 52 items, 35 items had more than 50% of the experts agreeing with our categorization. For seven items, exactly half of the experts agreed with our categorization, and nine items had less than 50% of the experts agreeing with our categorization. In general, more statistical literacy items received 100% or 75% of agreement in the categorization by the experts, and fewer statistical reasoning items had as much agreement in their categorization. The reviewers were also asked to critique and provide feedback about the items. Changes were made to some of the items based on the expert feedback and this led to the second draft of the REALI instrument. See Sabbag (2016) for more detailed explanation on the expert feedback.

3.4. THINK-ALOUD INTERVIEWS

To provide evidence for how students were responding to the items (response process validity evidence), think-aloud interviews were conducted with students from the University of Minnesota. A total of four statistics students agreed to participate in the think-aloud interviews (one undergraduate and three MA students). Students were given the items from the REALI assessment and asked to read each item aloud. They were also asked to comment on their thinking-process as they answered the items. These think-aloud interviews were tape recorded. As items from BLIS and GOALS had already been through this type of validation process, only new and modified items were used in the think-aloud interview. Data from these interviews was also used to validate the categorization of items that could not definitively be categorized as statistical literacy or statistical reasoning. In total, 26 REALI items were used in the think-aloud interviews.

Based on students' responses, five items were identified as needing additional modification. For instance, some items were not interpreted by the students in the way that the items were designed. Other items were clearly confusing or not clear to students as they would read the item many times during the think-aloud interview. These items were modified based on students' responses. These changes led to the third draft of the REALI instrument which was used in the pilot study.

3.5. PILOT TEST

The next step in the development process was a pilot study using the third draft of the REALI assessment. A total of five instructors from the University of Minnesota and one from Augsburg College agreed to participate in the pilot study. Three instructors from the University of Minnesota were teaching different sections of the same introductory statistics undergraduate level course, the other two instructors from the same university were teaching two different introductory graduate level courses, and the instructor from Augsburg was teaching an introductory statistics undergraduate level course. To increase student participation, all instructors agreed to administer the REALI assessment as an extra credit opportunity. REALI was administered online through Qualtrics. The initial page of the assessment contained a consent form asking whether students were willing to participate in the study.

Information contained in the consent form clearly explained to the students that they would still receive extra credit even if they did not want to give consent.

Ultimately, 237 students participated and gave consent for their data to be used in this research. One hundred and twenty-nine students were from undergraduate-level statistics courses, 69 students were from a social science introductory graduate-level statistics course, and 39 students were from a biostatistics graduate-level course. Students' responses were used to compute item difficulty, item discrimination, and percentage of students responding to each alternative for each of the 52 proposed items. This information was used to select the final 40 items to be used on the REALI assessment. Additionally, these psychometric analyses also informed minor modifications to some items.

3.6. FIELD TEST

The 40-item REALI assessment was then used in a large-scale field test. A recruitment email was sent out via (1) the *Consortium for the Advancement of Undergraduate Statistics Education (CAUSE)* website (http://www.causeweb.org); (2) the Statistical Education section of the *American Statistical Association*; and (3) the *Isolated Statisticians* listserv (http://ww2.amstat.org/committees/isostat/isostat.html). This email contained information about the purpose of the study and about the REALI instrument. Instructors who were currently teaching an introductory statistics course (undergraduate and graduate level) from colleges (2 and 4 years) and universities in the United States were invited to administer the REALI assessment to their students. To encourage participation, any participating instructor was promised a report with information about their students' performance and a comparison to students from other institutions.

Instructors who were interested were sent a second email with additional information about the assessment. These instructors were also asked for the following information: (1) institution name, (2) course name, (3) number of sections, (4) number of students in each section, and (5) short description of the curriculum. The initial page of the assessment contained a consent form asking if students were willing to participate in the study. Information about the field test sample is reported in Section 4.

3.7. DATA ANALYSIS

To investigate the psychometric properties of the REALI instrument, Item Response Theory (IRT), a framework for relating student responses to an underlying latent ability trait, was used to analyze the data. The intent of this data analysis is to identify the measurement model that best represents the constructs of statistical literacy and statistical reasoning given the criteria of fit and parsimony. Three theoretical IRT models were fitted to students' responses from the field test: a unidimensional IRT model, a bi-dimensional IRT model with correlated dimensions, and a bi-dimensional IRT model with uncorrelated dimensions. Each of these models represents a different potential structure for how the constructs of statistical literacy and statistical reasoning may be related.

The first model, the unidimensional IRT model, represents a structure in which statistical literacy and statistical reasoning are indistinguishable from one another. In this model, all 40 REALI items would load on a single dimension, which will be hence referred to as "Statistical Knowledge." (See Figure 2.)

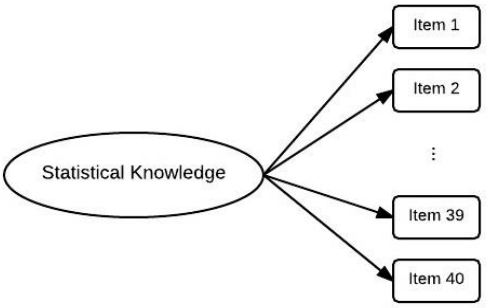


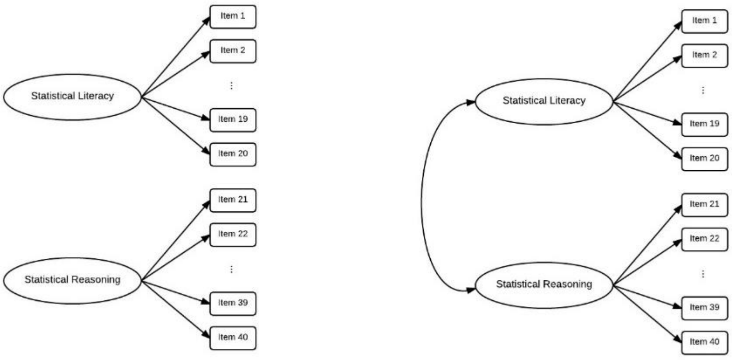
Figure 2. Unidimensional IRT model

This unidimensional model specifies the probability of a correct response to an item as a logistic distribution in which items are allowed to vary in terms of their difficulty and discrimination. In the unidimensional model, the probability of a correct response on any particular item is a function of a respondent's ability level (θ), that item's potential to discriminate between respondents of varying ability levels (α), and the item's difficulty level (δ). Mathematically, this can be expressed as

$$p\left(x_j = 1|\theta, \alpha_j, \delta_j\right) = \frac{\exp\left(\alpha_j(\theta - \delta_j)\right)}{1 + \exp\left(\alpha_j(\theta - \delta_j)\right)},$$

where θ is the latent trait (or person location parameter), α_j is the discrimination parameter for item j, and δ_j is the difficulty for item j.

The remaining two models consider a multidimensional structure between statistical literacy and statistical reasoning. The Uncorrelated Model (Figure 3a) utilizes a structure in which the two dimensions, statistical literacy and statistical reasoning, are uncorrelated with one another. In this model, the 20 literacy items on REALI would be expected to load on the literacy dimension, and the 20 reasoning items would be expected to load on the reasoning dimension. The Correlated Model (Figure 3b) is very similar to the Uncorrelated Model, except that the dimensions of statistical literacy and statistical reasoning are now allowed to correlate.



(a) (b)

Figure 3. Bi-dimensional IRT models

The two parameter-logistic IRT model (2PL) and multidimensional extension of this model (McKinley & Reckase, 1983; Reckase, 1985) were used to fit the unidimensional and bi-dimensional models, respectively. These models estimate the probability of a correct response to an item using a logistic function that takes into account the item's difficulty and how well that item discriminates between individuals of different ability levels. See Sabbag (2016) for technical details on these models. When fitting these models, the origin (mean of ability values) was fixed to zero and the variance of ability values was fixed to one. Model- and item-level fit was evaluated for all three IRT models.

Fit measures and model comparisons Several fit indices were used to evaluate the fit of the three IRT models to the REALI data. At the item-level, the $S\text{-}X^2$ statistic (Orlando & Thissen, 2000, 2003) was employed to assess whether each item fits the IRT model. This statistic is based on the observed and expected frequencies correct and incorrect for each summed score. Under the hypothesis that the model fits the data and the sample size is large, the $S\text{-}X^2$ statistic is approximately distributed as a Pearson chi-squared statistic. Significant values indicate lack of fit.

The Root Mean Square Error of Approximation (RMSEA) was used to evaluate model-level fit. Guidelines for evaluation suggest that RMSEA values between 0.00 and 0.05 indicate close fit, values between 0.05 and 0.08 indicate fair fit, values ranging from 0.08 to 0.10 indicate mediocre fit, and values above 0.10 indicate unacceptable fit (Browne & Cudeck, 1993).

The IRT models were further compared using the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978). The AIC and BIC statistics allow for comparison of both nested and unnested models, as long as the same outcome and data are used to estimate those models. Smaller AIC and BIC values indicate better data–model fit. After using the fit measures to select the best-fitting measurement model, empirical reliability (Zimowski, Muraki, Mislevy, & Bock, 2003) of this model's scores were also computed.

4. RESULTS

This section reports the results from the psychometric analyses of the REALI field test data. A total of 23 instructors from 16 colleges and universities around the United States and Canada administered the REALI assessment online through Qualtrics. A total of 671 students consented to participate and also completed the assessment. All students were enrolled in introductory level statistics courses at the undergraduate and graduate level. The method of administration (in-class or outside of class) was decided by the instructors. The only requirement was for students to work independently when completing the assessment. To increase student participation and effort, it was suggested instructors use the assessment to provide credit or extra credit to the students. All analyses were conducted using R version 3.3.0 (*R* Development Core Team, 2016).

4.1. DESCRIPTIVE ANALYSIS

A histogram of the distribution of the REALI total raw scores for the 671 students in the sample is presented in Figure 4. The mean of these scores was 24.16 (*SD* = 7.48) and the median was 24. The minimum and maximum values observed were 4 and 40 respectively. The estimate of the internal consistency, coefficient alpha, for these scores was 0.87.

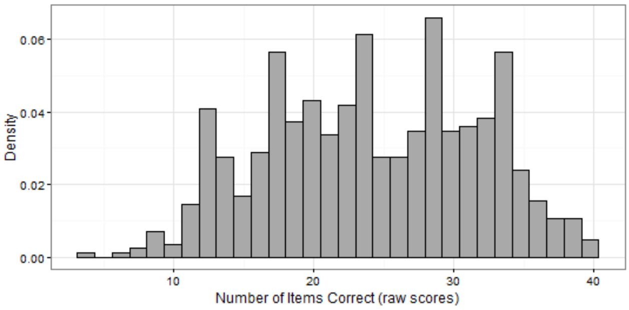


Figure 4. Distribution of total scores

The statistical literacy and statistical reasoning raw subscores were also investigated to better understand how students were performing in statistical literacy and statistical reasoning items. Histograms of the distributions of the two raw subscores for the 671 students in the sample are presented in Figure 5. The mean statistical literacy subscore was 13.15 (*SD* = 3.82) and the mean statistical reasoning subscore was 11.01 (*SD* = 4.15). The estimate of the internal consistency, coefficient alpha, for the statistical literacy subscore was 0.76 and for the statistical reasoning subscore was 0.78.

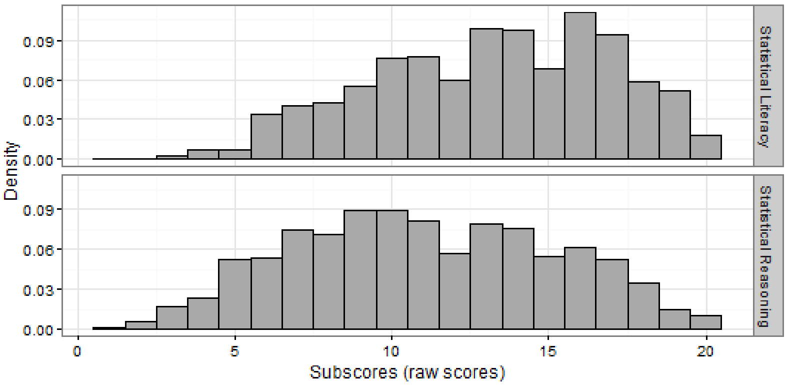


Figure 5. Distribution of the statistical literacy and statistical reasoning subscores

Item characteristics (difficulty and discrimination) were computed for all 40 items. Item difficulty values ranged from 0.27 (very difficult items) to 0.97 (very easy items). Item discrimination values ranged from 0.05 to 0.52. Only four out of 40 items presented low discrimination (values smaller than 0.2).

4.2. IRT ANALYSIS

This section presents the analytic results from fitting the three IRT models: the unidimensional model, the bi-dimensional uncorrelated model, and the bi-dimensional correlated model. Each of these models was fitted using the MIRT package in R (Chalmers, 2012). At the model-level, all three models indicated good overall residual fit, having a RMSEA of 0.00 (Browne & Cudeck, 1993). These values, along with the AIC and BIC values for each model, are presented in Table 2.

Table 2. Model-level fit measures for the three IRT models

	Model		
Fit Measures	Unidimensional	Uncorrelated	Correlated
RMSEA	0.00	0.00	0.00
AIC	29648.41	30294.51	29656.09
BIC	30009.12	30655.21	30021.30

Item-level fit was also examined for each model. Table 3 provides the estimated item parameters (intercept, item discrimination, and item difficulty) and standard errors obtained from fitting the three IRT models to the data. It is important to note that the high standard errors displayed in Table 3 might lead to poorly estimated parameters. In this table, items having poor discrimination (discrimination values lower than 0.8; De Ayala, 2009) are flagged for each of the three analyses. Items 1, 3, 5, 6, 28, 30, 33, and 36 presented with low discrimination in all three IRT models. The unidimensional model also flagged Item 11, whereas the bi-dimensional uncorrelated model flagged Items 17 and 24. Of note, the unidimensional model also flagged Item 5 as an extremely easy item; 97% of the students answered it correctly.

To further examine the item-level fit, item-level diagnostics statistics were computed (presented in Table 4). Fourteen items were flagged in all three models as having statistically significant misfit: Items 2, 4, 18, 19, 25, 27, 28, 30, 34, 36, 38, and 40. The unidimensional and bi-dimensional uncorrelated models also identified Item 35 as showing misfit. The bi-dimensional correlated model also flagged Items 9 and 31.

Lastly, reliability and correlation estimates were calculated for each model. These are presented in Table 5. The reliability of the estimated ability scores was higher for the unidimensional and correlated bi-dimensional models. The correlation between the statistical literacy dimension and statistical literacy dimension was set to zero for the uncorrelated bi-dimensional model. The estimate of this correlation given by the correlated bi-dimensional model was 0.96. This refers to the model estimated correlation between the latent traits of statistical literacy and statistical reasoning.

Table 3. Estimates and standard errors for the item parameters based on fitting the three IRT models. items with low discrimination are in italic

	Unidimensional			Uncorrelated			Correlated		
				Discrimination			Discrimination		
Item	Discrimination	Intercept	Difficulty	Literacy	Reasoning	Intercept	Literacy	Reasoning	Intercept
1	*0.70 (0.12)*	1.20 (0.10)	-1.72 (0.27)	*0.65 (0.12)*	-	1.19 (0.10)	*0.70 (0.11)*	-	1.21 (0.10)
2	1.07 (0.13)	1.20 (0.11)	-1.12 (0.13)	1.06 (0.14)	-	1.19 (0.12)	1.08 (0.13)	-	1.20 (0.11)
3	*0.72 (0.10)*	-0.98 (0.10)	1.37 (0.21)	-	*0.7 (0.11)*	-0.98 (0.10)	-	*0.73 (0.10)*	-0.98 (0.10)
4	1.66 (0.19)	2.04 (0.18)	-1.23 (0.11)	-	1.52 (0.20)	1.95 (0.17)	-	1.67 (0.20)	2.06 (0.19)
5	*0.33 (0.25)*	3.48 (0.24)	-10.5 (7.82)	*0.28 (0.26)*	-	3.47 (0.23)	*0.35 (0.25)*	-	3.49 (0.24)
6	*0.35 (0.09)*	0.12 (0.08)	-0.35 (0.24)	*0.32 (0.09)*	-	0.12 (0.08)	*0.35 (0.09)*	-	0.12 (0.08)
7	0.99 (0.12)	0.82 (0.10)	-0.83 (0.12)	-	1.03 (0.13)	0.83 (0.10)	-	1.02 (0.12)	0.83 (0.10)
8	0.84 (0.11)	-1.14 (0.10)	1.35 (0.18)	-	0.87 (0.12)	-1.15 (0.11)	-	0.86 (0.11)	-1.14 (0.11)
9	1.42 (0.17)	1.87 (0.16)	-1.32 (0.13)	1.56 (0.20)	-	1.96 (0.17)	1.50 (0.17)	-	1.93 (0.16)
10	1.24 (0.13)	-0.26 (0.10)	0.21 (0.08)	1.02 (0.12)	-	-0.26 (0.10)	1.22 (0.12)	-	-0.25 (0.10)
11	*0.78 (0.12)*	1.41 (0.11)	-1.80 (0.26)	-	0.84 (0.14)	1.43 (0.12)	-	0.81 (0.13)	1.42 (0.11)
12	0.87 (0.11)	-0.14 (0.09)	0.16 (0.11)	-	0.94 (0.12)	-0.14 (0.09)	-	0.89 (0.11)	-0.13 (0.09)
13	1.09 (0.13)	1.04 (0.11)	-0.95 (0.12)	1.07 (0.14)	-	1.03 (0.11)	1.12 (0.13)	-	1.05 (0.11)
14	1.21 (0.15)	1.55 (0.13)	-1.28 (0.14)	1.12 (0.15)	-	1.51 (0.13)	1.22 (0.15)	-	1.56 (0.13)
15	0.91 (0.11)	-0.42 (0.09)	0.46 (0.11)	0.90 (0.12)	-	-0.42 (0.09)	0.92 (0.11)	-	-0.41 (0.09)
16	1.72 (0.2)	1.97 (0.18)	-1.15 (0.09)	-	1.62 (0.21)	1.90 (0.17)	-	1.71 (0.20)	1.97 (0.18)
17	0.89 (0.11)	-0.45 (0.09)	0.51 (0.11)	-	*0.78 (0.11)*	-0.44 (0.09)	-	0.89 (0.10)	-0.45 (0.09)
18	1.54 (0.17)	1.53 (0.15)	-1.00 (0.10)	-	1.56 (0.19)	1.54 (0.15)	-	1.60 (0.17)	1.57 (0.15)
19	1.71 (0.18)	1.41 (0.15)	-0.83 (0.08)	1.66 (0.19)	-	1.37 (0.15)	1.75 (0.18)	-	1.44 (0.15)
20	1.21 (0.13)	0.44 (0.10)	-0.37 (0.10)	-	1.26 (0.14)	0.45 (0.11)	-	1.23 (0.13)	0.45 (0.10)
21	1.79 (0.18)	1.16 (0.14)	-0.65 (0.07)	1.78 (0.20)	-	1.13 (0.14)	1.80 (0.17)	-	1.17 (0.14)
22	0.94 (0.13)	1.44 (0.12)	-1.53 (0.19)	1.02 (0.14)	-	1.47 (0.12)	0.99 (0.13)	-	1.46 (0.12)

	Unidimensional			Uncorrelated			Correlated		
				Discrimination			Discrimination		
Item	Discrimination	Intercept	Difficulty	Literacy	Reasoning	Intercept	Literacy	Reasoning	Intercept
23	1.65 (0.18)	1.48 (0.15)	-0.90 (0.09)	-	1.49 (0.18)	1.40 (0.14)	-	1.66 (0.18)	1.49 (0.15)
24	0.83 (0.11)	0.45 (0.09)	-0.54 (0.12)	-	*0.77 (0.11)*	0.43 (0.09)	-	0.82 (0.11)	0.45 (0.09)
25	1.24 (0.13)	0.32 (0.10)	-0.26 (0.08)	1.38 (0.15)	-	0.33 (0.11)	1.30 (0.13)	-	0.34 (0.11)
26	1.13 (0.12)	0.06 (0.10)	-0.05 (0.09)	1.14 (0.13)	-	0.05 (0.10)	1.16 (0.12)	-	0.07 (0.10)
27	0.80 (0.10)	-0.01 (0.09)	0.02 (0.11)	-	0.80 (0.11)	-0.02 (0.09)	-	0.80 (0.10)	-0.01 (0.09)
28	*0.64 (0.10)*	0.62 (0.09)	-0.96 (0.18)	-	*0.57 (0.10)*	0.60 (0.09)	-	*0.63 (0.10)*	0.62 (0.09)
29	1.00 (0.11)	-0.40 (0.10)	0.40 (0.10)	0.98 (0.12)	-	-0.41 (0.10)	1.00 (0.11)	-	-0.40 (0.10)
30	*0.32 (0.09)*	0.30 (0.08)	-0.94 (0.35)	*0.25 (0.09)*	-	0.30 (0.08)	*0.31 (0.09)*	-	0.30 (0.08)
31	1.44 (0.14)	0.10 (0.11)	-0.07 (0.07)	-	1.45 (0.16)	0.10 (0.11)	-	1.47 (0.14)	0.11 (0.11)
32	1.19 (0.12)	-0.19 (0.10)	0.16 (0.08)	-	1.16 (0.13)	-0.19 (0.10)	-	1.2 (0.12)	-0.18 (0.10)
33	*0.50 (0.09)*	0.42 (0.08)	-0.83 (0.22)	-	*0.61 (0.10)*	0.43 (0.09)	-	*0.53 (0.10)*	0.42 (0.08)
34	1.35 (0.14)	0.24 (0.11)	-0.18 (0.08)	1.37 (0.15)	-	0.23 (0.11)	1.36 (0.13)	-	0.25 (0.10)
35	0.87 (0.11)	-0.18 (0.09)	0.20 (0.11)	-	1.04 (0.13)	-0.19 (0.10)	-	0.91 (0.11)	-0.18 (0.09)
36	*0.44 (0.10)*	1.06 (0.09)	-2.42 (0.56)	*0.50 (0.11)*	-	1.08 (0.09)	*0.44 (0.10)*	-	1.07 (0.09)
37	1.46 (0.26)	3.40 (0.29)	-2.34 (0.29)	1.21 (0.24)	-	3.21 (0.26)	1.42 (0.24)	-	3.38 (0.28)
38	1.25 (0.14)	0.82 (0.11)	-0.65 (0.09)	1.21 (0.14)	-	0.79 (0.11)	1.26 (0.14)	-	0.82 (0.11)
39	0.87 (0.11)	-0.44 (0.09)	0.51 (0.12)	-	0.90 (0.12)	-0.45 (0.09)	-	0.89 (0.11)	-0.44 (0.09)
40	0.96 (0.11)	-0.79 (0.10)	0.82 (0.12)	-	1.03 (0.13)	-0.81 (0.10)	-	0.99 (0.11)	-0.79 (0.10)

Note. When fitting the bi-dimensional uncorrelated model, the covariance between the statistical literacy and the statistical reasoning constructs was set to 0. When fitting the bi-dimensional correlated model, the covariance between the statistical literacy and the statistical reasoning constructs was freely estimated as 0.959.

Table 4. Item-level diagnostic statistics for the three IRT models

	Unidimensional		Uncorrelated		Correlated			Unidimensional		Uncorrelated		Correlated	
Item	X^2 (d.f.)	*p*	X^2 (d.f.)	*p*	X^2 (d.f.)	*p*	Item	X^2 (d.f.)	*p*	X^2 (d.f.)	*p*	X^2 (d.f.)	*p*
1	25.67 (27)	0.54	24.06 (27)	0.63	25.57 (26)	0.49	21	25.21 (22)	0.29	31.43 (24)	0.14	25.10 (21)	0.24
2	46.84 (24)	*0.00*	44.43 (24)	*0.01*	46.55 (23)	*0.00*	22	25.85 (26)	0.47	25.90 (25)	0.41	25.96 (25)	0.41
3	31.62 (26)	0.21	33.72 (26)	0.14	31.69 (25)	0.17	23	27.86 (22)	0.18	31.32 (22)	0.09	27.88 (21)	0.14
4	41.75 (20)	*0.00*	44.45 (23)	*0.01*	41.53 (19)	*0.00*	24	30.12 (25)	0.22	30.04 (25)	0.22	29.73 (24)	0.19
5	7.99 (12)	0.79	7.78 (12)	0.80	7.98 (11)	0.72	25	48.65 (25)	*0.00*	55.75 (25)	*0.00*	50.71 (24)	*0.00*
6	29.29 (29)	0.45	29.10 (27)	0.36	29.30 (28)	0.40	26	34.61 (25)	0.10	38.27 (26)	0.06	34.89 (24)	0.07
7	30.37 (25)	0.21	31.36 (25)	0.18	30.74 (24)	0.16	27	47.19 (26)	*0.01*	49.52 (26)	*0.00*	47.02 (25)	*0.01*
8	30.41 (26)	0.25	33.89 (26)	0.14	30.63 (25)	0.20	28	46.67 (27)	*0.01*	47.72 (27)	*0.01*	46.30 (26)	*0.01*
9	38.20 (23)	*0.02*	34.10 (22)	0.05	37.99 (21)	*0.01*	29	18.48 (25)	0.82	19.40 (25)	0.78	18.26 (24)	0.79
10	27.56 (24)	0.28	33.45 (26)	0.15	27.23 (23)	0.25	30	45.79 (29)	*0.03*	45.67 (29)	*0.03*	45.59 (28)	*0.02*
11	25.44 (26)	0.49	28.63 (26)	0.33	25.62 (25)	0.43	31	37.62 (24)	*0.04*	35.63 (24)	0.06	37.53 (23)	*0.03*
12	34.88 (26)	0.11	37.94 (26)	0.06	35.21 (25)	0.08	32	24.48 (24)	0.43	23.91 (24)	0.47	24.39 (23)	0.38
13	24.69 (24)	0.42	25.73 (25)	0.42	24.78 (23)	0.36	33	27.82 (28)	0.47	33.09 (26)	0.16	28.03 (26)	0.36
14	21.83 (23)	0.53	23.54 (25)	0.55	21.83 (22)	0.47	34	41.47 (24)	*0.02*	43.49 (25)	*0.01*	40.97 (23)	*0.01*
15	31.40 (26)	0.21	31.60 (26)	0.21	31.51 (25)	0.17	35	37.11 (26)	0.07	42.80 (26)	*0.02*	37.51 (25)	0.05
16	13.66 (20)	0.85	19.86 (22)	0.59	14.97 (20)	0.78	36	52.53 (28)	*0.00*	45.95 (27)	*0.01*	52.13 (27)	*0.00*
17	23.28 (26)	0.62	23.15 (26)	0.62	23.14 (25)	0.57	37	9.20 (16)	0.91	11.44 (18)	0.88	9.35 (15)	0.86
18	42.13 (22)	*0.01*	42.60 (22)	*0.01*	42.18 (21)	*0.00*	38	42.34 (24)	*0.01*	39.44 (25)	*0.03*	41.86 (23)	*0.01*
19	36.89 (22)	*0.02*	37.78 (24)	*0.04*	36.93 (21)	*0.02*	39	37.52 (26)	0.07	39.71 (26)	*0.04*	37.65 (25)	0.05
20	23.63 (25)	0.54	23.82 (25)	0.53	23.57 (24)	0.49	40	43.50 (25)	*0.01*	47.55 (25)	*0.00*	43.83 (24)	*0.01*

Note. Italics indicates a *p*-value < 0.05.

Table 5. Reliability and correlation estimates for the IRT models

	Model				
	Unidimensional	Uncorrelated		Correlated	
		Literacy	Reasoning	Literacy	Reasoning
Reliability	0.88	0.78	0.79	0.86	0.87
Correlation	-	-		0.96	

5. DISCUSSION

This section provides a critique of the expert reviews of the initial set of items, as well as a discussion of what was learned about the psychometric properties of the REALI assessment. The section concludes with limitations, future research, and conclusion.

5.1. EXPERT REVIEW—CATEGORIZATION OF ITEMS

During the expert review process, the experts were asked to categorize each item into two groups: Group 1 (statistical literacy) and Group 2 (statistical reasoning). In general, for most items, the experts' classifications were the same as the classification of the author. However, there was more agreement between the experts and the author for the statistical literacy items than for the statistical reasoning items. For instance, 27 out of the 28 (96%) statistical literacy items produced a moderate to high level of agreement between the experts and the author (half or more of the experts agree with the categorization of the author). On the other hand, only 16 out of the 24 (66%) statistical reasoning items had a moderate to high level of agreement. A possible reason why this happened could be related to the definitions of statistical literacy and statistical reasoning items used in this study. For example, to categorize an item as a statistical reasoning item, the reviewers had to carefully examine each item to verify how many statistical concepts were being addressed and then examine if these concepts needed to be connected to answer the question correctly. Therefore, recognizing statistical reasoning items demanded more steps than recognizing a statistical literacy item.

An additional problem in the categorization happened with items that displayed the relationship between concepts in the alternative options and not on the stem of the problem. This was the case for all four items with no agreement between the author and the experts. These items were categorized as statistical reasoning items by the author because the alternative options for each item addressed more than one statistical concept. Therefore, when students go through the alternative options to answer the item, they are forced to make connections between more than one statistical concept, thus exhibiting statistical reasoning. None of the expert reviewers classified these items as statistical reasoning items. A possible reason for this disagreement could be because the experts did not consider that statistical reasoning could happen whereas students were reading the alternative options. Thus, experts might have focused only on the stem of the items.

5.2. MEASUREMENT MODELS

Three IRT models were fitted to students' responses. The first model was a unidimensional model with only one overall dimension (statistical knowledge). The other two models (uncorrelated model, correlated model) were bi-dimensional models, each composed of two dimensions: a statistical literacy dimension and a statistical reasoning dimension.

Table 6 summarizes the evidence used in the model comparison and the model that indicated the most favorable fit based on this evidence. At the model-level, the RMSEA values suggest that all three models have similar fit. The AIC and BIC measures support the unidimensional model. At the item-level, the $S\text{-}X^2$ statistic flagged similar numbers of misfitting items for the three models, although there were differences in which items were identified as having misfit. The reliability estimates supported both the unidimensional and correlated models. In terms of correlation between the statistical literacy and statistical reasoning constructs, it is clear that these two constructs are highly correlated. In fact, even setting the correlation between constructs to zero, for the uncorrelated bi-dimensional model, the model still provided highly correlated estimated ability scores for statistical literacy and reasoning (Pearson correlation coefficient of 0.99). This supports the idea that not allowing these constructs to correlate is not a proper assumption given how intertwined they are. Taking into account the information provided above, the unidimensional model seems to be the best model to represent the construct of statistical literacy and the construct of statistical reasoning given the criteria of *fit* and *parsimony*.

Table 6. Summary of evidence and the supported models

Model	RMSEA	AIC	BIC	S-X^2	Reliability
Unidimensional	✓	✓	✓	✓	✓
Uncorrelated	✓			✓	
Correlated	✓			✓	✓

5.3. INSTRUMENT’S CHARACTERISTICS

The REALI assessment was developed to concurrently measure statistical literacy and statistical reasoning. This instrument comprised 40 items, with 20 items measuring statistical literacy and 20 item measuring statistical reasoning.

As suggested by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), careful attention was given to how the scores from the REALI instrument would be interpreted. To support the intended inferences and uses of the scores, different types of validity evidence were gathered throughout the development process: expert reviews, response process interviews with students, a pilot test, a field test, and psychometric analyses. Regarding the preciseness of scores from REALI, there was evidence of high score reliability. This gives good evidence for the extent to which the scores from REALI are precise and supports the interpretation and use of the REALI scores.

From the psychometric analysis, it was observed that the unidimensional model presented nine items with low item discrimination. Five of these items were statistical literacy items and four were statistical reasoning items. These nine items addressed the following statistical concepts: interpreting a distribution in terms of shape, center, and variation; interpretation of a sample mean and how it is affected by outliers; measures of variability; relationship between statistical significance and sample size; understanding that a confidence interval for a proportion is centered at the sample statistic; probability; and randomness. The items with the lowest discrimination values were three statistical literacy items: Item 5, Item 6, and Item 30. The next paragraphs will explore the students’ responses to these items.

Item 30 (see Figure 6) was the worst discriminating item in the REALI instrument. This item was designed to assess students’ ability to understand that a confidence interval for a population proportion is centered at the sample statistic. A total of 57% of the students got this item correct, but the low item discrimination gives evidence that these students were not necessarily the ones with the highest abilities. Around one fourth of the students with the highest abilities chose alternative B which stated that “37% of veterans in the *population* have been divorced at least once.” A possible reason why this happened could be that students might be thinking of 37% as a plausible value for the population parameter because 37% is included in the confidence interval. In addition, alternative B does not state any level of confidence when making an inference about the population and this does not appear to concern students. This item also presented concerning results on the pilot test and even though its performance on the field test improved with modification of the item, it seems that this item is still not performing well enough. This item will most likely be deleted and replaced by another statistical literacy item addressing confidence interval concepts.

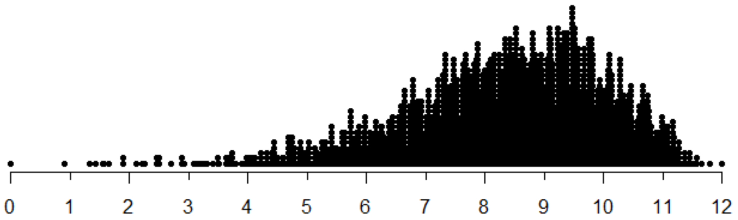
- In a recent study of Vietnam veterans, researchers randomly selected a sample of veterans and asked them if they had been divorced at least once. They calculated a 95% confidence interval for the percent of veterans that had been divorced at least once (35% to 39%). Which of the following statements is true about the center of the interval (37%)?

a) We can say that 37% of veterans in the *sample* have been divorced at least once.
b) We can say that 37% of veterans in the *population* have been divorced at least once.
c) We can say that 95% of veterans in the *sample* have been divorced at least once.
d) We can say that 95% of veterans in the *population* have been divorced at least once.

Figure 6. Item 30

Item 5 (Figure 7) was another poorly discriminating item which was designed to assess students' ability to understand how the mean is affected by skewness. The main reason this item produced such a low discrimination was because almost all students (97%) correctly answered this item. This was the easiest item in the whole instrument and does not seem to differentiate students with low and high ability. This item will most likely be re-written so that the level of difficulty is increased and thus higher discrimination is achieved.

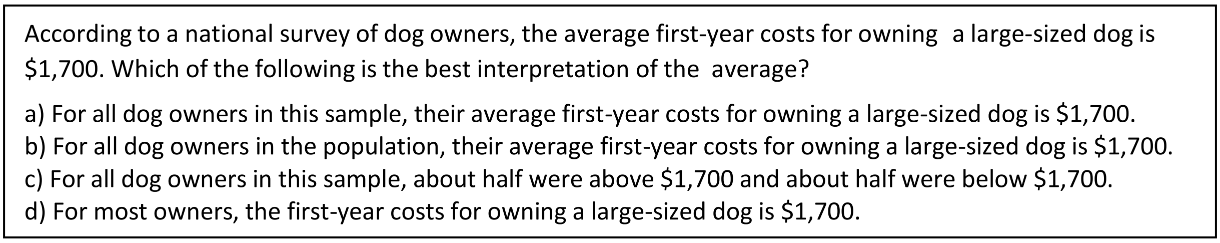
Which of the following intervals is *most* likely to include the mean of the distribution below?



a) 4 to 6
b) 7 to 9
c) 10 to 12

Figure 7. Item 5

Similar to Item 5, Item 6 also belongs to the category of "Measures of Center" (see Figure 8). This item was designed to measure students' ability to interpret the mean in the context of the data. About half of the students correctly answered this item recognizing that the average was a summary measure representing the dogs in the sample. Alternative D behaved properly having negative discriminations of -0.20. However, alternatives B and C presented discrimination values of -0.12 and -0.03. This means that some high ability students were likely to choose alternative B or alternative C as the correct answer. Alternative C interprets the median instead of the mean, which is a common misconception. Alternative B has a similar interpretation of the average as alternative A (the correct answer), but it refers to the *population* instead of the *sample*. Students who chose alternative B might be ignoring the role of study design and making a generalization to a population even without any information about how the survey's responses were obtained. In addition, the word "national" in the stem of the item might be misleading students. Maybe, students are interpreting the word "national" as equivalent to a *representative* sample. It seems that there is no problem with the item itself (e.g., bad item writing) and the reason for the bad discrimination could be because of students' misconceptions leading them to choose the wrong alternatives. However, more think-aloud interviews are needed to understand the reason why students are choosing incorrect alternatives.



According to a national survey of dog owners, the average first-year costs for owning a large-sized dog is $1,700. Which of the following is the best interpretation of the average?

a) For all dog owners in this sample, their average first-year costs for owning a large-sized dog is $1,700.
b) For all dog owners in the population, their average first-year costs for owning a large-sized dog is $1,700.
c) For all dog owners in this sample, about half were above $1,700 and about half were below $1,700.
d) For most owners, the first-year costs for owning a large-sized dog is $1,700.

Figure 8. Item 6

5.4. LIMITATIONS

Much has been learned about the relationship between statistical literacy and statistical reasoning in this study. However, the limitations of the study are important to consider in interpreting the results. Firstly, instructors and students participated in this study on a voluntary basis, and the administration of the REALI instrument was not uniform among all institutions. Some instructors used REALI as a required part of the course, and others as an extra credit opportunity or a review for the exam. Therefore, students' effort and response rate varied greatly among institutions. In addition, most, if not all of, the students completed the REALI assessment outside of class. This could add additional variation in students' scores due to environmental issues such as distractions. These differences in test administration and the small sample size might be the cause for the high standard errors and low discriminating items in Table 3.

Another point to consider is that the content covered in introductory statistics courses and the time spent on that content varies widely, and it is likely that not all test takers had the same opportunity to learn the content covered in REALI. Lack of opportunity to learn can introduce guessing and consequently measurement error in students' responses. This adds to uncertainty regarding students' responses and therefore decreases the reliability of scores. Item order effects and test fatigue are also problems not measured in this study that could potentially influence the results.

5.5. IMPLICATIONS FOR FUTURE RESEARCH

In this study, the choice for a best measurement was based on the criteria of *fit* and *parsimony*. However, questions remain regarding the possible hierarchy and overlap between these two learning goals. Therefore, further research will explore what measurement model best represents the construct of statistical literacy and statistical reasoning given the criteria of *reliability* and *distinction*, with the final goal of finding the most useful model for understanding the relationship between statistical literacy and statistical reasoning. Additional models that were not considered in this study will be added in the analysis: a bi-factor model and a model with cross-loading from the statistical literacy dimension to the statistical reasoning dimension.

It is also important to explore the REALI items, across the three IRT models, which presented misfit and low discrimination. Removing items due to item-misfit or low discrimination is not desirable because this could lead to a lack of representation of the learning goals being measured in the instrument. In terms of item discrimination, additional items could be written to better differentiate between students with high and low ability levels. However, research about the effect of misfit items has more currently focused on evaluating what are the practical consequences of item misfit, instead of focusing on the statistical item fit analyses (Köhler & Hartig, 2017). Therefore, more explorations of the data are necessary to understand the practical impact of these misfitting items. Further research is also needed to understand why low discriminating items are behaving as they are and how these items can be improved. In addition, as mentioned in Section 5.4, IRT models' parameters are not being estimated accurately; thus, additional research is needed to improve parameter estimation for each of the IRT models. As an anonymous reviewer of this manuscript pointed out, usually multidimensional IRT models present a better model fit than unidimensional models. However, this was not observed in this study as the unidimensional model provided better fit than the bi-dimensional correlated model. This could be due to the high correlation between the constructs of statistical literacy and reasoning but further research is needed to explore if the preference for a unidimensional model can be replicated.

Item changes could affect the empirical results, so it would be important to see whether the results of this study replicate in such a study. Researchers could also try to replicate the results of this study with different populations of students. For example, would the same results present when REALI is administered to students in upper-level statistics courses?

5.6. CONCLUSION

This research study reported on the development process of the REALI instrument and provided a validity argument supporting the uses of the scores from the REALI instrument. The results from expert reviews and think-aloud interviews support the instrument’s ability to measure students’ statistical literacy and statistical reasoning. Data analysis of the pilot and field tests suggest high evidence of score precision and good psychometric properties. In addition, this study also provides solid and research-based definitions of statistical literacy and statistical reasoning that can be used to bring unity to the research in statistics education. Therefore, this study provides valuable and significant information for the statistics education community.

The REALI assessment can be used at the end of an introductory statistics course to provide information about important statistical literacy and statistical reasoning topics to evaluate students’ learning outcomes. In addition, REALI can also be used in the evaluation of curricula or to assess the effect of curriculum changes, as long as the learning goals assessed by this instrument are closely aligned with the intended learning goals of the curricula being used in class. Thus, the REALI instrument can be a tool for identifying students’ misconceptions and guiding changes and improvements in statistics courses.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

American Statistical Association (ASA, 2007), *Using Statistics Effectively in Mathematics Education Research.* Alexandria, VA: Author
[Online: www.amstat.org/asa/files/pdfs/EDU-UsingStatisticsEffectivelyinMathEdResearch.pdf]

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

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: SAGE.

Budgett, S., & Pfannkuch, M. (2007). Assessing students’ statistical literacy. In P. Bidgood, N. Hunt & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 103–121). Chichester, UK: John Wiley & Sons Ltd.

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

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3), 1–17.
[Online: https://doi.org/10.1080/10691898.2002.11910677]

Chance, B. L., & Garfield, J. B. (2002). New approaches to gathering data on student learning for research in statistics education. *Statistics Education Research Journal*, *1*(2), 38–41.
[Online: http://iase-web.org/documents/SERJ/SERJ1(2).pdf]

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

delMas, R. C. (2002). Statistical literacy, reasoning and learning: A commentary. *Journal of Statistics Education*, *10*(3).
[Online: https://doi.org/10.1080/10691898.2002.11910679]

delMas, R. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi and J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 79–95). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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]

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

Garfield, J. (1991). Evaluating students' understanding of statistics: Development of the statistical reasoning assessment. In R. Underhill (Ed.) *Proceedings of the Thirteenth Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, 2*, 1–7. Blacksburg, VA.

Garfield, J. (1998). The statistical reasoning assessment: Development and validation of a research tool. In L. Pereira-Mendoza, L. Kea, T. Kee, & W. Wong (Eds.) *Proceedings of the Fifth International Conference on Teaching Statistics (ICOTS-5),* Singapore (pp. 781–786). Voorburg, The Netherlands, International Statistical Institute.
[Online: https://www.stat.auckland.ac.nz/~iase/publications/2/Topic6u.pdf]

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10*(3).
[Online: https://doi.org/10.1080/10691898.2002.11910676]

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., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review, 75*(3), 372–396.

Garfield, J., & Ben-Zvi, D. (2008). Developing students' statistical reasoning. *Connecting Research and Teaching Practice*. Dordrecht, The Netherlands: Springer.

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning, 2*(1-2), 99–125.

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. (n.d.). *Assessment Resource Tools for Improving Statistical Thinking*.
[Online: https://apps3.cehd.umn.edu/artist/]

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.
[Online: https://apps3.cehd.umn.edu/artist/articles/AERA_2003.pdf]

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.

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

Köhler, C., & Hartig, J. (2017). Practical significance of item misfit in educational assessments. *Applied Psychological Measurement, 41*(5), 388–400.

Konold, C. (1995), Issues in assessing conceptual understanding in probability and statistics, *Journal of Statistics Education, 3*(1).
[Online: https://doi.org/10.1080/10691898.1995.11910479]

McKinley, R. L., & Reckase, M. D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Report No. ONR83-2). Iowa City, IA: American College Testing Program.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289–298.

Park, J. (2012). *Developing and validating an instrument to measure college students' inferential reasoning in statistics: an argument-based approach to validation* (Unpublished doctoral dissertation). Retrieved from the University of Minnesota Digital Conservancy.
[Online: https://conservancy.umn.edu/handle/11299/165057]

R Development Core Team (2016). *R*: A language and environment for statistical computing [Computer software.] R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401–412.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3), 6–13.
[Online: https://doi.org/10.1080/10691898.2002.11910678]

Sabbag, A. (2016). *Examining the relationship between statistical literacy and statistical reasoning* (Unpublished doctoral dissertation). Retrieved from the University of Minnesota Digital Conservancy.
[Online: http://hdl.handle.net/11299/182193]

Sabbag, A., & Zieffler A. (2015). Assessing learning outcomes: An analysis of the GOALS-2 instrument. *Statistics Education Research Journal*, *14*(2), 93–116.
[Online: http://iase-web.org/documents/SERJ/SERJ14(2)_Sabbag.pdf]

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Sinharay, S. (2010). When can subscores be expected to have added value? Results from operational and simulated data. *ETS Research Report Series*, 2010(2), i–28.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3–46.
[Online: http://iase-web.org/documents/SERJ/SERJ2(2).pdf]

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature. *Journal of Statistics Education*, *16*(2).
[Online: https://doi.org/10.1080/10691898.2008.11889566]

Ziegler, L. (2014). *Reconceptualizing statistical literacy: Developing an assessment for the modern introductory statistics course* (Unpublished doctoral dissertation). Retrieved from the University of Minnesota Digital Conservancy.
[Online: http://hdl.handle.net/11299/165153]

Ziegler, L. (2018). Developing a statistical literacy assessment for the modern introductory statistics course. *Statistics Education Research Journal*, *17*(2), 161–178.
[Online: http://iase-web.org/documents/SERJ/SERJ17(2)_Ziegler.pdf]

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG (Version 3) [Computer Program]. Mooresville, IN: Scientific Software.

ANELISE SABBAG
Department of Statistics
1 Grand Avenue
San Luis Obispo, CA 93407
USA