CAN WE DISTINGUISH STATISTICAL LITERACY AND STATISTICAL REASONING?

ANELISE SABBAG
California Polytechnic State University
asabbag@calpoly.edu

ANDREW ZIEFFLER
University of Minnesota
zief0002@umn.edu

CASEY NG
California Polytechnic State University
cng27@calpoly.edu

ABSTRACT

One of the most important goals in a statistics class is to develop students who are statistically literate and can reason with statistical concepts. The REALI instrument was designed to concurrently assess statistical literacy and reasoning in introductory statistics students. This paper reports a measurement analysis of the statistical literacy and reasoning subscores from the REALI assessment and the extent to which they are reliable and distinct. Investigation of these subscores is used clarify the relationship between the constructs of statistical literacy and statistical reasoning and to what extent they overlap. The results of this analysis, under a Multidimensional Item Response Theory framework, show that the statistical literacy and reasoning subscores provide no added value over a single general statistical knowledge score. This indicates the two constructs might be indistinguishable from one another.

Keywords: Statistics education research; Assessment; Statistical Literacy; Statistical Reasoning; Subscores

1. INTRODUCTION

There are a growing number of introductory statistics courses that have less emphasis on calculations and procedures, and instead focus on developing students' understanding and reasoning (e.g., Adams et al., 2021; Cummiskey et al., 2020; Garfield et al., 2012; Hudiburgh et al., 2020; Tintle et al., 2012). As tertiary statistics educators look to update their courses, one resource guiding their pedagogical decisions is often the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report* (GAISE College Report ASA Revision Committee, 2016).

The *GAISE College Report* outlines a set of recommendations for teaching introductory statistics in two- and four-year tertiary-level institutions. It also lists several learning objectives for introductory statistics students with the intent of producing statistically educated students. The first recommendation is to "[t]each statistical thinking" (p. 12) a recommendation initially made in 1992 when a focus group led by George Cobb was initiated to suggest improvements in the teaching of introductory statistics (Cobb, 1992).

The *GAISE College Report* defines statistical thinking as, "the type of thinking that statisticians use when approaching or solving statistical problems" (p. 12). This definition is derived largely from the work of Wild and Pfannkuch (1999) who interviewed applied statisticians about their reasoning and thinking as they engaged in an empirical enquiry. This focus on the understandings and elements underlying the process of "doing statistics" is at the root of many early attempts at defining statistical thinking (e.g., Mallows, 1998; Moore, 1990; Snee, 1990; Sylwester, 1993).

Statistics Education Research Journal, 24(1), Article 2. https://doi.org/10.52041/serj.v24i1.587

While statistical thinking is an important outcome for introductory statistics students, courses that emphasize conceptual understanding are often more focused on developing students' statistical literacy and reasoning. Within the *GAISE College Report*, these two outcomes are subsumed in the recommendation to "teach statistical thinking." (p. 3) For example, in further describing this recommendation, the authors indicated that "[e]ffective statistical thinking requires *seeing connections among statistical ideas* …" (p. 12) (statistical reasoning) and an end goal is to have "… students to *become statistically literate*" (p. 12) (statistical literacy).

Despite the documented importance of developing students' statistical literacy and reasoning in both the educational and research communities, there is little consensus about the conceptualization of these constructs. Also, while researchers have opined about the relationship between these constructs (e.g., Callingham & Watson, 2017; Chance, 2002; delMas, 2002; Garfield & Ben-Zvi, 2007; 2008), there has been no empirical work, to date, that has confirmed any of these hypotheses. This study seeks to better understand the structure of the relationship between statistical literacy and reasoning through a psychometric analysis of students' responses on the REALI assessment (Sabbag et al., 2018), an instrument developed to assess statistical literacy and reasoning.

2. DEFINING AND DISTINGUISING STATISTICAL LITERACY AND STATISTICAL REASONING

This section summarizes definitions and descriptions of statistical literacy and statistical reasoning available in the statistics education literature. Information is also reported about the relationship between these learning goals.

2.1. STATISTICAL LITERACY

Much work has been done to describe and understand statistical literacy better in the field of statistics education (e.g., Cobb, 1992; Gal, 2002; Rumsey, 2002; Utts, 2003). One of the most cited definitions of statistical literacy comes from Gal's 2002 paper about the nature of adults' statistical literacy and its components. Gal proposed a model of statistical literacy with two components: *knowledge* and *dispositions*. The knowledge component is comprised of cognitive elements such as literacy skills, knowledge of statistics, mathematics and context, and critical questions. The dispositions component is comprised of critical stance and composed of beliefs and attitudes. In addition to proposing this model of statistical literacy, Gal also defined statistical literacy as the ability to comprehend, interpret, communicate, and evaluate statistical information critically. Budgett and Pfannkuch (2007) built on Gal's (2002) work but added a reasoning piece to his definition. This piece was comprised of statistical argumentation knowledge and everyday events knowledge (viewing daily events from a statistical perspective). Kaplan and Thorpe (2010) defined statistical literacy as the skills and knowledge adults need to be consumers of statistics, which is consistent with Gould's (2017) observation that, "[t]he set of knowledge and understanding required to be statistically literate is often defined by differentiating the needs of consumers of statistics from those of producers of statistics, a dichotomy that goes back at least as far as Hotelling (1940)" (p. 22).

These definitions give us a broad view of what statistical literacy is, but are not adequate to describe, nor measure the construct. In a pioneering article, reviewing the scholarship on statistical literacy, Rumsey (2002) tried to present a more nuanced description of statistical literacy by considering the related student goals encompassed by statistical literacy in an introductory statistics course. She categorized the goals into two sub-components of statistical literacy that she termed *statistical competence* and *statistical citizenship.* Goals categorized into the first component were related to the knowledge students need to acquire before being able to reason and think statistically (e.g., data awareness, data collection, basic statistical concepts, basic interpretation, and communication skills). The citizenship component included goals related to students' ability to operate in a data-driven society (e.g., critiquing results, making decisions based on statistical information). Other frameworks used to describe statistical literacy have been proposed by Watson and Callingham (2003), Gal (2004), Kaplan and Thorpe (2010), and Sharma et al. (2011).

Despite the amount of published work undertaken to describe and better understand statistical literacy in the field of statistics education, there is little consistency observed among how this construct

is defined within this literature. Some of this variation seems to be related to the population being studied—statistical literacy looks different for school-level students than for tertiary students. Another source of variation may be that the construct of statistical literacy has changed over time. As the world changes, so does the knowledge and skills people need to engage as informed citizens. This in turn, changes what is encompassed in the construct of statistical literacy. For example, MacGillivray (2021) proposed students should be able to tackle modern data in different levels such as recording, wrangling, dissecting, handling, exploring, processing, presenting, and analyzing. Engel (2017) also stated that cleaning, transforming, and structuring data are necessary skills for students to have in today's world. These types of changes portend that the construct of statistical literacy adapts to include some degree of data literacy and elements of computational thinking (e.g., Gould, 2021; MacGillivray, 2021; Prodromou & Dunne, 2017; Ridgway et al., 2013).

2.2. STATISTICAL REASONING

Like statistical literacy, statistical reasoning has been defined by many statistics educators and researchers. Unlike statistical literacy, most of the definitions of statistical reasoning are primarily focused on college and school students. In addition, most of the work in this area has focused on reasoning about specific statistical concepts.

Garfield (2002) reviewed definitions of statistical reasoning from six researchers (Chervaney et al., 1977; Chervaney et al., 1980; Hawkins et al., 1992; Nisbett, 1993; Sedlmeier, 1999; and Lovett, 2001) and concluded that no clear agreement has been reached regarding these definitions. In addition, Garfield stated that more studies are needed to better understand students' statistical reasoning and how it can be developed in statistics courses.

To clarify the learning goal of statistical reasoning, Garfield and Chance (2000) and Garfield (2002) defined statistical reasoning as reasoning with statistical concepts and understanding statistical information. According to the authors, statistical reasoning included interpretations, representations and summarizing data. It also included connecting statistical concepts from which further inferences can be drawn. Garfield and Ben-Zvi (2008) expanded the previous definition, stating that statistical reasoning is "mental representations and connections that students have regarding statistical concepts" (p. 34). Further efforts were made by Jones et al. (2004) who reviewed three papers about models of development in statistical reasoning (Jones et al., 2000; Mooney, 2002; Watson et al., 1995) and concluded that this construct was composed of hierarchical stages and cycles.

2.3. RELATIONSHIP BETWEEN STATISTICAL LITERACY AND REASONING

Despite the prevalence in the literature of references to statistical literacy and reasoning, understanding of these constructs is still evolving. For example, the literature reviewed in the previous sections reveals a lack of consensus regarding how the two constructs are defined, described, and related.

To understand the possible connections between these learning goals, delMas (2002) reviewed definitions from Garfield (2002), Rumsey (2002), and Chance (2002). From this review, he proposed two models of the relationship between these constructs. In the first model, statistical literacy has content that is independent from statistical reasoning; there is, however, an overlap between them. In the second model proposed, statistical literacy is an all-encompassing goal of instruction and statistical reasoning does not have content independent from literacy.

This first model from delMas (2002) aligned with what was observed in the literature as some of the statistical literacy definitions had statistical reasoning components which might be due to this overlap between constructs. For instance, Rumsey (2002) defined "statistical citizenship" (which is considered a part of statistical literacy) and mentioned that students take actions that may require statistical reasoning, such as the judgment and evaluation of statistical information. Rumsey (2002), however, also stated that "statistical competence" (which is also considered as a part of statistical literacy) is a requirement for statistical reasoning. Additionally, in their investigations of the construct of statistical literacy, Callingham and Watson (2017) called attention to statistical reasoning aspects in Gal's (2004) definition of statistical literacy, and they also used statistical reasoning as part of their definition of upper levels of statistical literacy. Another example of this overlap is found in Budgett and

Pfannkuch's (2007) definition of adults' statistical literacy, which also contained a statistical reasoning component. Based on these ideas, it seems that statistical literacy and reasoning have their own attributes, being independent of each other, but may also overlap. The extent to which these learning goals might overlap, however, is not clear in the literature. The terms statistical literacy and reasoning have been used interchangeably in the literature (Chance, 2002; delMas, 2004; Garfield, 2002), and this might point to the idea that these concepts might overlap so much, that they might be indistinguishable.

Another issue identified in the literature concerns a possible hierarchy between statistical literacy and reasoning. Garfield and Ben-Zvi (2007, 2008) argued for a hierarchy between these learning goals with statistical literacy as the basis for statistical reasoning. To better understand this possible hierarchy, Garfield and Ben-Zvi (2008) compared each of these learning goals with the categories in Bloom's taxonomy (Bloom et al., 1956); no empirical evidence was reported, however, regarding this possible alignment with Bloom's taxonomy and the hierarchy between these learning goals.

In summary, there are unanswered questions regarding the relationship between statistical literacy and reasoning. It might be that these are distinct but related constructs, or it is possible that they might overlap (and the extent of the overlap is not clear in the literature). Finally, a possible hierarchy between these constructs (with literacy being a requirement for reasoning) has also been proposed in the literature. These possible relationships between statistical literacy and reasoning will be translated to three different statistical models to be examined later in the paper.

2.4. SUMMARY AND RESEARCH QUESTION

Although research in statistics education has posed different models for how the constructs of statistical literacy and statistical reasoning might be related, much of what is reported about the relationship between statistical literacy and reasoning is not based on empirical evidence. This study seeks to understand the relationship between these two learning goals using student response data from the Reasoning and Literacy Instrument (REALI) instrument, an assessment designed to concurrently measure statistical literacy and statistical reasoning. For more information about other assessments of statistical literacy and reasoning see Sabbag et al. (2018).

The construction of REALI and item development process is laid out in Sabbag et al. (2018), which also provides initial validity evidence. Some of that evidence, based on comparing the fit indices of three potential models of how the constructs of statistical literacy and reasoning may be related, indicated that the unidimensional model had the best model–data fit. These results suggest that the constructs of statistical literacy and reasoning may overlap so much that they cannot be distinguished (at least when measured by the REALI instrument).

The analysis undertaken, however, had several limitations including the criteria used to select the "best" model (fit and parsimony) and a homogenous sample of students (all from a single university). Moreover, the study employed a limited set of simple models to describe the structure and relationship of these constructs. While these limitations are understandable given the early stage of the assessment's development cycle, the collection and presentation of validity evidence is a continual process (Messick, 1989), and is important for further informing the structure of these constructs and the use of test scores from REALI.

Despite the careful construction and development of the REALI assessment, there is an open question about whether separate statistical literacy and reasoning scores should be reported in addition to a total score. Depending on how statistical literacy and reasoning are related, scores for these two constructs (subscores) might be highly correlated and, therefore, might not provide distinct information from the total score. This could mean that the constructs of statistical literacy and reasoning are indistinguishable from one another (at least as when measured by the REALI instrument). If this is true, then reporting subscores is not useful or appropriate, as they would not provide independent diagnostic information about students' statistical literacy and reasoning (Haberman, 2008; American Educational Research Association et al., 2014; Tate 2004).

To build on earlier work related to REALI, the current study seeks to investigate the nature of the relationship between statistical literacy and reasoning by evaluating a set of measurement models hypothesized in the literature. Relatedly, we will examine whether statistical literacy and reasoning can be measured *reliably* and *distinctly* and the extent to which it is worth reporting subscores for these constructs in addition to an overall score when using the REALI assessment.

3. METHODS

The REALI instrument is composed of a total of 40 items, which are intended to measure the constructs of statistical literacy and statistical reasoning. Half of the items were written to focus on statistical literacy, and the other half were written to focus on statistical reasoning. Sabbag et al. (2018) contains detailed information about the development of REALI instrument together with the validity evidence that was gathered to support the intended inferences as uses of the instruments' scores. To illustrate, we provide one example of a statistical literacy item and a statistical reasoning item. Item 9 (Figure 1) is an example of a statistical literacy item and to answer this question correctly, students need to know how to interpret a standard deviation given a context. In the REALI instrument, statistical literacy items assess students' ability to recall, 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 students to make connections between them (recall information will be sufficient). Item 18 (Figure 2) is an example of a statistical reasoning item that measures students' ability to recognize biased and unbiased sampling methods; students also must recognize the sampling method used in the stem of the question is not a random sample. Finally, students also had to understand that accuracy was related to study design and not sample size. It is important to note that the reasoning part of this item is included in the four alternative options. Each alternative addresses a different statistical concept; therefore, forcing students to reason with these concepts while answering the question. In the REALI assessment, 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. Please see Sabbag et al. (2018) for additional information about REALI.

9. Thirty introductory statistics students took a quiz worth 30 points. The standard deviation of the quiz scores was 1 point. Which of the following gives the most suitable interpretation of this standard deviation?
   a. All of the individual scores are one point apart.
   b. The difference between the highest and lowest score is 1 point.
   c. The difference between the upper and lower quartile is 1 point.
   d. A typical distance of a score from the mean is 1 point.

Figure 1. Example of statistical literacy item from REALI

18. A sportswriter wants to know the extent to which football fans in a large city support building a new football stadium. She stands outside the current football stadium before a game and interviews the first 250 people who enter the stadium. The newspaper reports the results from the sample as an estimate of the percentage of football fans in the city who support building a new stadium. Which statement is correct in terms of the sampling method?
   a. This is a simple random sample. It will give an accurate estimate.
   b. Because the sample is so small, it will *not* give an accurate estimate.
   c. Because all fans had a chance to be asked, it will give an accurate estimate.
   d. The sampling method is biased. It will *not* give an accurate estimate.

Figure 2. Example of statistical reasoning item from REALI

To investigate the relationship between statistical literacy and statistical reasoning, the response data from REALI will be used to fit a set of measurement models hypothesized in the statistics education literature. The results from fitting these models will be used to provide additional validity evidence and inform how student scores from the REALI instrument should be reported and interpreted. The data

collected, candidate models fitted, and analyses undertaken are described in detail in the following subsections.

3.1. DATA COLLECTION

The 40-item REALI assessment was administered in two phases. In the first phase, 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 statistics education section of the *American Statistical Association*; and (3) the *Isolated Statisticians* listserv (http://ww2.amstat.org/committees/isostat/isostat.html). See Sabbag et al. (2018) for more information on this first phase of data collection. One of the limitations reported by Sabbag et al. (2018) was the small sample size, which might have been the cause of high standard errors and low discriminating items. For this reason, a second round of data collection took place mostly at California Polytechnic State University. The lead author administered the assessment to students in her introductory statistics courses during the following quarters: Spring 2018, Fall 2018, Winter 2019, Spring 2019, and Winter 2020. Additional data in this second phase were also collected from instructors who contacted the author and showed interest in using the REALI assessment. These instructors were from the University of Minnesota, University of Colorado Boulder, University of Northern Colorado, Wesleyan College, and Washington and Lee University and data was collected during the 2017–2020 academic years. The Institutional Review Board from California Polytechnic State University reviewed and approved this research project (IRB2018263).

3.2. CANDIDATE MODELS

To determine the nature of the relationship between statistical literacy and statistical reasoning, we identified three candidate models based on descriptions of the hypothesized relationship between statistical literacy and reasoning posited in the statistics education literature (Models A–C). In addition, we consider an additional candidate model in which the two constructs are indistinguishable from one another (Model D.) Path diagrams of the four candidate models are presented in Figure 3a–d and described more thoroughly below.

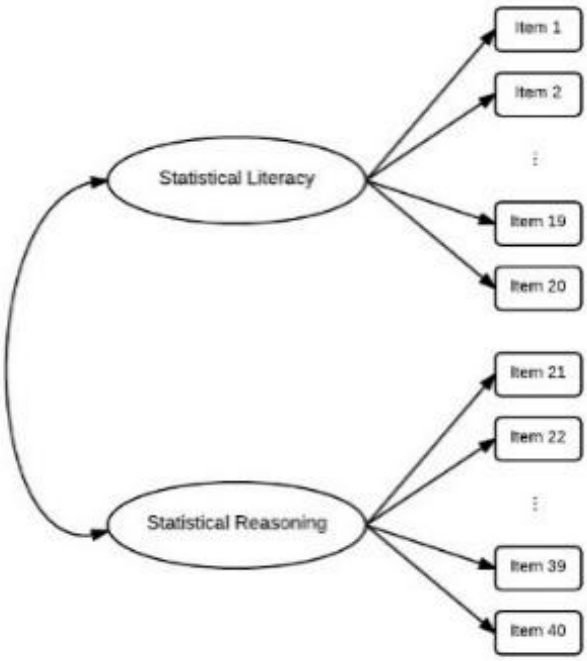

Figure 3a. Model A: Candidate model fitted to examine distinctiveness in subscores for students' statistical literacy and statistical reasoning

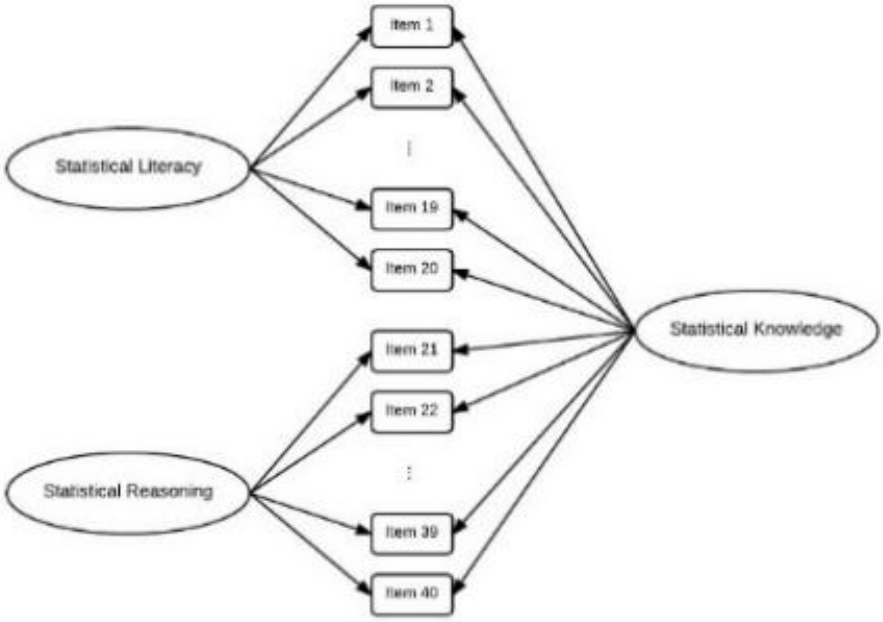

Figure 3b. Model B: Candidate model fitted to examine distinctiveness in subscores for students' statistical literacy and statistical reasoning

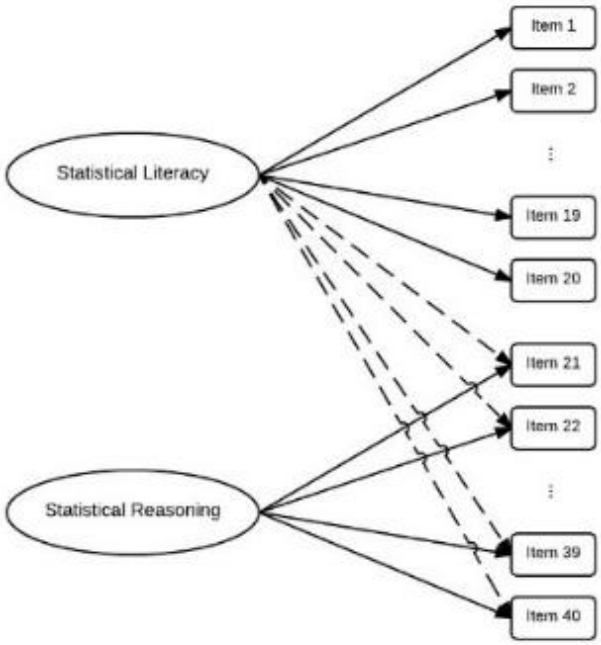

Figure 3c. Model C: Candidate model fitted to examine distinctiveness in subscores for students' statistical literacy and statistical reasoning

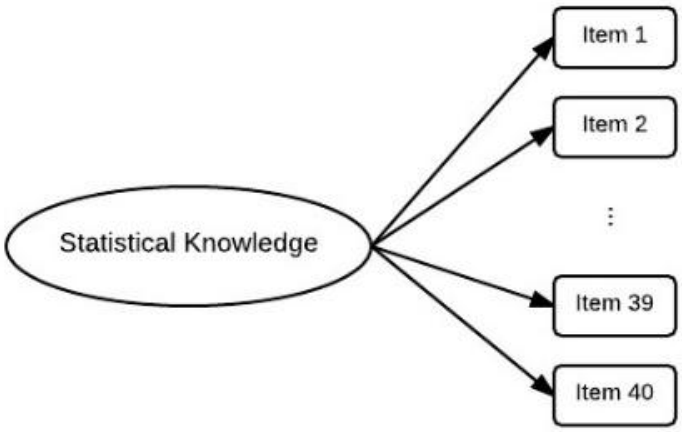

Figure 3d. Model D: Candidate model fitted to examine distinctiveness in subscores for students' statistical literacy and statistical reasoning

The measurement model underlying the first candidate model, Model A, presumes that the covariation in the response data is due to two correlated latent dimensions, namely statistical literacy and statistical reasoning. Within this model, statistical literacy and reasoning are distinct yet related constructs, without an assumed hierarchy.

A second measurement model used to investigate the structure between statistical literacy and statistical reasoning (Model B) is the bi-factor model (Gibbons & Hedeker, 1992). In this model, the covariation in the response data is again due to the direct effects of two independent latent constructs, statistical literacy and reasoning. In this model, the dimension of statistical literacy has a direct effect solely on statistical literacy items and the dimension of statistical reasoning has a direct effect solely on statistical reasoning items. There is also a general construct (which we refer to as *Statistical Knowledge*) that also has direct effects on all the literacy and reasoning items. These effects account for the shared variability in students' responses.

Model C, similar to Model A, suggests that the relationship between statistical literacy and reasoning is distinct yet related. In this model, however, there is a hierarchical structure with statistical literacy forming the basis of statistical reasoning. Here, the covariation in the response data is again due to the two latent constructs, but the dimension of statistical literacy now has a direct effect on all the items (including the items assessing statistical reasoning). This can be seen by the dotted lines in the model which represent the cross-loadings. In fitting this model, the effect of statistical literacy on all items will be fixed to a constant value, and the cross-loadings will be constrained so that literacy's effect on the reasoning items is smaller than the direct effects from statistical reasoning. This reflects the idea that statistical literacy plays the same role in responding to all items and that statistical reasoning plays a larger role than literacy in responding to the reasoning items.

Lastly, Model D 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 we will again refer to as *Statistical Knowledge*.

3.3. ITEM RESPONSE THEORY ANALYSIS

Because all items in the REALI instrument were scored dichotomously (correct or incorrect), and to make the analysis parallel to the one performed in Sabbag et al. (2018), all four candidate models were fitted utilizing an Item Response Theory (IRT) framework. Model D has a single ability dimension (statistical knowledge), and it was fitted using a unidimensional 2-Parameter Logistic (2PL) IRT model that specifies the probability of a correct response in which items are allowed to vary in terms of their difficulty and discrimination. In the 2PL model, the probability of a correct response is given by

$$p(x_{ij} = 1|\theta_j, \alpha_i, \delta_i) = \frac{e^{\alpha_i(\theta_j - \delta_i)}}{1+e^{\alpha_i(\theta_j - \delta_i)}} \qquad (3.1)$$

where θ_j is the statistical knowledge ability for person *j*, α_i is the discrimination parameter for item *i*, and δ_i is the difficulty for item *i*. The remaining candidate models were fitted using Multidimensional Item Response Theory (MIRT), an extension of the unidimensional IRT model that seeks to explain a student's response to an item based on the student's abilities across multiple latent dimensions (Reckase, 2009). The MIRT models estimate the probability of a correct response to an item using a logistic function that considers the item's difficulty and how well that item discriminates between individuals of different ability levels. Models A, B, and C are based on a multidimensional extension of the 2PL Model (McKinley & Reckase, 1983; Reckase, 1985), where the probability of a response (x_{ij} = 1 as correct and x_{ij} = 0 as incorrect) on item *i* by person *j* is given by

$$p(x_{ij} = 1|\boldsymbol{\theta_j}, \boldsymbol{\alpha_i}, \gamma_i) = \frac{e^{\boldsymbol{\alpha_i}\boldsymbol{\theta_j}' + \gamma_i}}{1+e^{\boldsymbol{\alpha_i}\boldsymbol{\theta_j}' + \gamma_i}}, \qquad (3.2)$$

where $\boldsymbol{\theta_j} = [\theta_{j1}, \theta_{j2}, \theta_{j3}, \ldots, \theta_{jm}]$ is the vector of abilities for each person on each of the *m* dimensions, the vector $\boldsymbol{\alpha}_i = [\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \ldots, \alpha_{im}]$ contains the discrimination parameter of item *i* for each of the *m* dimension, and the intercept parameter (or overall multidimensional item difficulty) of item *i* is γ_i.

Since Models A and C have two dimensions (statistical literacy and statistical reasoning), then $\boldsymbol{\theta_j} = [\theta_{j1}, \theta_{j2}]$ with θ_{j1} representing the statistical literacy ability of Person *j* and θ_{j2} the statistical reasoning ability for Person *j*. In addition, $\boldsymbol{\alpha}_i = [\alpha_{i1}, \alpha_{i2}]$ with α_{i1} representing the discrimination parameter of Item *i* on the statistical literacy dimension and α_{i2} representing the discrimination parameter of Item *i* on the statistical reasoning dimension. It is important to note that for Model A, each item is directly influenced only by the dimension it belongs to, therefore one of the elements in vector $\boldsymbol{\alpha}_i$ that is related to the dimension the items do *not* belong to will be set to 0. For instance, Item 1 is a statistical *literacy* item, and its discrimination vector $\boldsymbol{\alpha}_1$ will have α_{12} (discrimination parameter of item 1 on the statistical reasoning dimension) set 0 zero: $\boldsymbol{\alpha}_1 = [\alpha_{11}, 0]$. Item 3 is a statistical *reasoning* item, and its discrimination vector $\boldsymbol{\alpha}_3$ will have α_{31} (discrimination parameter of item 3 on the statistical literacy dimension) set 0 zero: $\boldsymbol{\alpha}_3 = [0, \alpha_{32}]$.

For Model C, the description of Model A regarding the discrimination values will remain the same for statistical literacy items. However, for statistical reasoning items, their discrimination parameters on the statistical literacy dimension will *not* be set to 0. Instead, the discrimination parameters on the statistical literacy dimension will be set to a fixed value that is the same across all the reasoning items. In addition, that fixed value for the discrimination parameters of reasoning items on the statistical literacy dimension will be smaller than the discrimination parameters of reasoning items on the statistical reasoning dimension. In other words, each of the statistical reasoning items will have its highest discrimination value from the statistical reasoning dimension to represent the need of statistical reasoning knowledge to answer a reasoning item beyond the need of statistical literacy knowledge.

Model B has three dimensions (statistical knowledge, statistical literacy, and statistical reasoning). For this model, the ability vector is $\boldsymbol{\theta_j} = [\theta_{j1}, \theta_{j2}, \theta_{j3}]$ with θ_{j1} and θ_{j2} having the same interpretation as for Models A and C, but θ_{j3} representing the statistical knowledge ability for person *j*. In addition, the discrimination vector is $\boldsymbol{\alpha}_i = [\alpha_{i1}, \alpha_{i2}, \alpha_{i3}]$ with α_{i1} and α_{i2} having the same interpretation as for Models A and C, but with α_{i3} representing the discrimination parameter of item *i* on the statistical knowledge dimension. The discrimination parameters on each dimension will be set in the same way as in Model A; however the discrimination vector $\boldsymbol{\alpha}_i$ for all items will also contain a non-zero α_{i3} value for each item.

When fitting these models, the ability estimates are on a continuous scale ranging from -3 to 3, the origin (mean of ability values) was fixed to zero and the variance of ability values was fixed to one.

3.4. ANALYSIS

Several fit indices were used to evaluate the fitted candidate models. These indices help evaluate fitness to the response data at the item- and model-level. 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 of correct and incorrect for each summed score. Under the hypothesis that the model fits the data, and the sample size is large, the S-X^2 statistic is approximately distributed as a Pearson chi-squared statistic. Large values of the statistic, corresponding to small p-values values indicate lack of fit.

The Root Mean Square Error of Approximation (RMSEA), Bentler's Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI) was also used to evaluate model-level fit. Guidelines for evaluation suggest that RMSEA values between 0.00 and 0.05 and TLI and CFI values greater than 0.95 indicate close fit to the data (Browne & Cudeck, 1993; Hu & Bentler, 1999).

The IRT models were further compared using the Akaike Information Criterion (AIC; Akaike, 1974), the corrected AIC (AICc; Sugiura, 1978), the Bayesian Information Criterion (BIC; Schwarz, 1978), and the sample-size-adjusted BIC (SABIC; Sclove, 1987). These statistics allow for comparison of both nested and unnested models, if the same outcome and data are used to estimate those models with smaller values indicate better data–model fit.

To evaluate the quality of subscores for statistical literacy and statistical reasoning produced from the three MIRT models, we examined measures of distinctness and reliability for these scores. To investigate the distinctiveness of subscores, correlations between the subscores produced from each model were compared. Models that produced more distinctive scores should have lower correlations between the subscores. To evaluate the reliability of the subscores, we computed and compared a measure of empirical reliability proposed by Zimowski et al. (2003) for the subscores from each MIRT model. This measure of reliability:

$$\rho_S^2 = \frac{Var(\theta_S)}{Var(\theta_S) + MSE} \ , \qquad (3.3)$$

where θ_s are the person-ability estimates for each latent dimension and MSE is the mean of the conditional error variance for the θ_s estimates for all students. Models producing subscores with higher reliability measures will be preferred.

4. RESULTS

In the first phase of data collection, a total of 23 instructors from 16 colleges and universities around the United States and Canada administered the REALI assessment online using *Qualtrics.* A total of 671 students consented to participate and completed the assessment. During the second phase of data collection, eight instructors from six colleges and universities in the United States administered the REALI assessment online using Survey Monkey. A total of 818 students consented to participate and completed the assessment. The final sample size was 1,489 students 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 4.1.0 (R Development Core Team, 2020).

4.1. TOTAL SCORES

The distribution of the total scores (number out of the 40 items that were answered correctly) for all students is displayed in Figure 4 (left). The mean and median scores were 26.79 and 28 with a standard deviation of 7.76. The scores varied from 4 to 40.

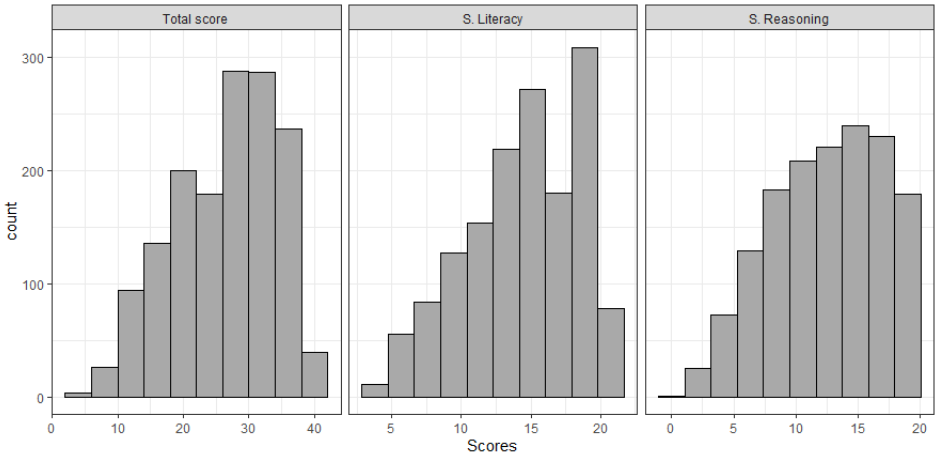

Figure 4. Distribution of REALI total score (left), statistical literacy subscores (center), and statistical reasoning subscores (right).

The distributions of the statistical literacy and statistical reasoning subscores (number out of 20 subscale items that were answered correctly) are presented in Figure 4 (center and right respectively). The mean and median statistical literacy subscore were 14.43 and 15 with a standard deviation of 3.95. The mean and median statistical reasoning subscore were 12.36 and 13 with a standard deviation of 4.26. There was a high correlation between the statistical literacy and reasoning subscores (r = 0.79), and both sets of subscores were also highly correlated with the total scores (r = 0.94 for literacy and r = 0.95 for reasoning).

4.2. ITEM RESPONSE THEORY ANALYSIS

Each of the four candidate models was fitted using the MIRT package in *R* (Chalmers, 2012). Fit indices and information criteria are reported in Table 1 for each model.

Table 1. Model-level fit indices and information criteria for the correlated (Model A), cross-loadings (Model B), bi-factor (Model C), and unidimensional (Model D) models

	Fit Indices			Information Criteria			
Model	RMSEA	TLI	CFI	AIC	AICc	SABIC	BIC
Model A	**0.03**	**0.98**	**0.98**	60524	60534	60697	**60954**
Model B	**0.02**	**0.98**	**0.98**	**60384**	**60405**	**60639**	61021
Model C	**0.04**	**0.95**	**0.95**	61734	61743	61904	62158
Model D	**0.03**	**0.98**	**0.98**	60513	60523	60684	60938

Note. Fit indices in bold indicate good fit to the data. Information criteria in bold indicate the model with the most empirical evidence given the data and candidate models.

All four candidate models presented good fit to the data with TLI and CFI values greater than 0.95 and RMSEA values smaller than 0.05. Based on the different information criteria, there was no consensus as to which model has the most empirical evidence. The bi-factor model (Model B) had the smallest AIC, AICc, and SABIC values. The correlated model (Model A) which had the smallest BIC value. Of note, the correlated model also had the second smallest AIC, AICc, and SABIC values.

Items were also examined for potential misfit using the $S\text{-}X^2$ statistic (Orlando & Thissen, 2000). This chi-squared based statistic measures the discrepancy between the observed and expected proportion of examinees who answer the item correctly for each potential test score (i.e., number of items correct). Table 2 includes the $S\text{-}X^2$ statistics for the 40 items in each of the four candidate models. Ten of the 40 items were flagged in all four models (Items 3, 8, 12, 19, 21, 25, 27, 30, 31, and 35).

Table 2. Item-level diagnostic statistics for the correlated (Model A), bi-factor (Model B), cross-loadings (Model C), and unidimensional (Model D)

	Model A		Model B		Model C		Model D	
Item	X^2 (d.f.)	*p*	X^2 (d.f.)	*p*	X^2 (d.f.)	*p*	X^2 (d.f.)	*p*
1	42.00 (29)	0.06	42.66 (29)	**0.05**	41.72 (30)	0.08	42.07 (30)	0.07
2	38.62 (27)	0.07	38.08 (27)	0.08	36.82 (28)	0.12	39.02 (28)	0.08
3	58.29 (27)	**0.00**	61.69 (27)	**0.00**	59.04 (28)	**0.00**	58.01 (28)	**0.00**
4	36.66 (26)	0.08	37.07 (26)	0.07	44.17 (28)	**0.03**	36.63 (27)	0.10
5	23.09 24()	0.51	22.94 (24)	0.52	20.57 (23)	0.61	21.23 (24)	0.63
6	34.17 (29)	0.23	34.23 (29)	0.23	38.52 (30)	0.14	34.08 (30)	0.28
7	34.94 (27)	0.14	34.18 (27)	0.16	33.91 (28)	0.20	35.02 (28)	0.17
8	43.88 (27)	**0.02**	44.88 (27)	**0.02**	45.86 (28)	**0.02**	43.94 (28)	**0.03**
9	36.54 (26)	0.08	35.53 (26)	0.10	36.78 (28)	0.12	36.75 (27)	0.10
10	30.66 (26)	0.24	30.48 (27)	0.29	32.10 (28)	0.27	30.94 (27)	0.27
11	38.20 (29)	0.12	38.57 (29)	0.11	44.79 (30)	**0.04**	37.61 (30)	0.16
12	45.53 (27)	**0.01**	44.27 (27)	**0.02**	48.57 (27)	**0.01**	46.97 (29)	**0.02**
13	30.83 (26)	0.24	31.27 (26)	0.22	34.22 (27)	0.16	30.87 (26)	0.23
14	33.52 (26)	0.15	33.16 (26)	0.16	35.38 (27)	0.13	33.79 (27)	0.17
15	28.63 (26)	0.33	28.24 (26)	0.35	31.48 (27)	0.25	28.78 (27)	0.37
16	12.22 (23)	0.97	13.27 (23)	0.95	22.71 (26)	0.65	12.25 (24)	0.98
17	34.48 (26)	0.12	35.80 (27)	0.12	37.36 (29)	0.14	35.24 (27)	0.13
18	34.75 (26)	0.12	33.86 (26)	0.14	43.36 (27)	**0.02**	34.66 (27)	0.15
19	52.84 (25)	**0.00**	51.42 (25)	**0.00**	47.93 (26)	**0.01**	53.53 (26)	**0.00**
20	17.51 (27)	0.92	17.51 (27)	0.92	19.76 (28)	0.87	17.42 (28)	0.94
21	45.60 (24)	**0.01**	46.84 (24)	**0.00**	44.49 (25)	**0.01**	46.47 (25)	**0.01**
22	39.20 (27)	0.06	37.88 (27)	0.08	38.82 (28)	0.08	39.23 (29)	0.10
23	22.00 (23)	0.52	22.57 (23)	0.49	56.63 (27)	**0.00**	21.86 (24)	0.59
24	32.38 (27)	0.22	32.79 (27)	0.20	32.67 (29)	0.29	32.27 (28)	0.26
25	52.80 (25)	**0.00**	50.85 (25)	**0.00**	55.72 (26)	**0.00**	52.25 (26)	**0.00**
26	29.16 (26)	0.30	28.61 (26)	0.33	35.47 (27)	0.13	28.92 (27)	0.37
27	54.58 (27)	**0.00**	50.48 (27)	**0.00**	57.40 (29)	**0.00**	55.02 (28)	**0.00**
28	31.68 (28)	0.29	31.89 (28)	0.28	33.74 (29)	0.25	31.71 (29)	0.33
29	20.47 (26)	0.77	20.13 (26)	0.79	25.71 (28)	0.59	20.44 (27)	0.81
30	47.78 (29)	**0.02**	47.14 (29)	**0.02**	51.06 (30)	**0.01**	47.83 (30)	**0.02**
31	47.50 (26)	**0.01**	47.11 (25)	**0.01**	48.96 (27)	**0.01**	47.98 (27)	**0.01**
32	25.88 (26)	0.47	25.34 (26)	0.50	28.90 (27)	0.37	25.77 (27)	0.53
33	26.57 (28)	0.54	27.60 (28)	0.49	30.18 (29)	0.41	26.57 (29)	0.60
34	33.45 (26)	0.15	33.70 (26)	0.14	33.84 (27)	0.17	33.74 (27)	0.17
35	44.64 (28)	**0.02**	45.58 (28)	**0.02**	53.73 (28)	**0.00**	44.38 (29)	**0.03**
36	34.72 (29)	0.21	33.85 (29)	0.25	36.92 (29)	0.15	34.97 (30)	0.24
37	13.20 (19)	0.83	13.41 (19)	0.82	19.29 (23)	0.68	13.23 (20)	0.87
38	37.96 (28)	0.10	38.38 (28)	0.09	33.87 (28)	0.21	38.56 (29)	0.11
39	31.87 (27)	0.24	30.18 (27)	0.31	33.68 (28)	0.21	31.79 (28)	0.28
40	34.14 (27)	0.16	33.66 (27)	0.18	35.63 (28)	0.15	33.95 (28)	0.20

Subscores. Subscores for statistical literacy and statistical reasoning were estimated from the three multidimensional models. These model-estimated subscores are distinct from the subscores displayed in Figure 4, as they are calculated based on the items that each student correctly answered and the difficulty of such items. A score for statistical knowledge was also estimated for Model B (the bi-factor model). The reliabilities for and correlations between these scores are presented in Table 3.

Based on these results, the subscores for Model A have high reliability but no evidence of distinction. The subscores from Model B seemed to be distinct but had very low reliability. The scores for the statistical knowledge dimension were highly reliable. The subscores from Model C have lower, albeit acceptable, reliability. In addition, they do show some evidence of distinction between the two sets of subscores.

The scores from the unidimensional model had a high reliability of 0.87. Almost all of the subscores from the multidimensional models (Models A–C) were highly correlated with the score from the

unidimensional model. The exceptions to this are the literacy and subscores from the bi-factor model (Model B).

Table 3. Reliability and correlation estimate for the multidimensional models

		Reliability	Correlation between Literacy and Reasoning subscores	Correlation between Literacy and Reasoning subscores and scores from the Unidimensional model
Model A	Literacy	0.87	0.99 (0.964*)	0.99
	Reasoning	0.87		0.99
Model B	Literacy	0.26	-0.11	0.04
	Reasoning	0.35		0.15
	Statistical knowledge	0.87	—	0.99
Model C	Literacy	0.79	0.76	0.95
	Reasoning	0.74		0.92

Note. *Model estimated correlation between the latent dimensions of statistical literacy and statistical reasoning.

All three multidimensional models (Models A–C) showed evidence of good fit to the data. Given the subscore analysis, however, Model C seems to have the most evidence of reliability and distinctiveness of the subscores. The estimated item parameters and standard errors obtained from fitting this model are presented in Table 4 (see Appendix) along with those from Model D.

5. DISCUSSION

This section provides a discussion of what was learned about the relationship between statistical literacy and reasoning constructs as measured by the REALI assessment. The section concludes with limitations and ideas for future research.

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) state that, if a test provides more than a single score, evidence of distinctiveness and reliability for these scores needs to be provided to support the use and interpretation of the subscores. Examining the total score and subscores from the REALI assessment and the extent to which they are reliable and distinct can provide this type of evidence.

All three multidimensional models (Models A–C) and the unidimensional model (Model D) showed evidence of good fit to the data. After evaluating the evidence of distinctiveness and the reliabilities of the subscores from the multidimensional models, however, the evidence suggests the subscores may not be that meaningful in terms of the information they provide.

5.1. RELATIONSHIP BETWEEN STATISTICAL LITERACY AND REASONING

The results of these analyses may also help clarify the relationship between the constructs of statistical literacy and reasoning. Under a MIRT approach, the cross-loading model (Model C) presented good fit to the data and evidence of reliability and distinction making this the most useful model to describe the possible relationship between statistical reasoning and literacy. This model supports the theory from Garfield and Ben-Zvi (2008) of a hierarchy between statistical literacy and statistical reasoning, with statistical literacy being the basis for statistical reasoning.

As stated by Sinharay et al. (2018), however, if subscores have added value over a total score, the MIRT model should provide a better fit of the data than a unidimensional IRT model. In this study, the unidimensional model (Model D) showed improved values of RMSEA, TLI and CIF as well as smaller values of AIC, AICc, SABIC, and BIC, indicating that this model provides a better fit to the data than the cross-loading model. In addition, the unidimensional model flagged fewer items as poorly discriminating or with misfit (see Appendix and Table 2). Finally, the statistical literacy and reasoning

subscores from the cross-loading model were highly correlated with the scores from the unidimensional model which also suggests that the literacy and reasoning subscores might not provide additional information than the scores from the unidimensional model.

These results suggest the subscores of statistical literacy and statistical reasoning provided by the REALI assessment do not have added value beyond a total score. This indicates that the constructs of literacy and reasoning might overlap so much that they cannot be distinguished. This supports the second theoretical model proposed by delMas (2002) in which statistical reasoning is presented as a subset of statistical literacy.

5.2. LIMITATIONS

This study aimed to clarify the relationship between statistical literacy and reasoning; however, the limitations of the study are important to consider in interpreting the results. Firstly, what this study revealed about the relationship between statistical literacy and reasoning is tied to how these constructs were assessed (REALI assessment) and the definitions of statistical literacy and reasoning used in this study. As stated earlier, there are disagreements in terms of the definitions of these terms and the constructs could also be measured using different assessments (thought at this point, REALI is the only assessment who was designed to concurrently measure these constructs). Therefore, it is important to note that different results could be achieved if different definitions or assessments were used.

Another point to consider is that even though all participants were students enrolled in introductory statistics courses at institutions of higher education, no information was gathered on these students except for the answers to each question in the REALI instrument. For this reason, additional information regarding students' characteristics is not available. In addition, it is important to note that the participants were not evenly divided among the 31 instructors who administered the instrument in their classes and this nested structure of the data was not accounted for by the IRT framework. Finally, the instructors and students participated in this study on a voluntary basis, and the administration of the REALI instrument was not uniform among all institutions. These differences in test administration might have increased errors in student responses. As mentioned in Sabbag et al. (2018), the issue of the lack of opportunity to learn the concepts covered in this assessment can also introduce guessing and consequently measurement error in students' responses. This adds to the uncertainty regarding students' responses, and therefore potentially decreases the reliability of scores.

5.3. IMPLICATIONS FOR FUTURE RESEARCH

As previously stated in Sabbag et al. (2018), and as can be seen in Table 4, there are some items in the REALI assessment that do not discriminate well between students with high and low ability (discrimination values lower than 0.8; De Ayala, 2009). This continues to be true in this study. Additional research is needed to understand the students' thought process when they are answering high difficulty level items and how these items can be improved. It is also important to note that the average item difficulty for REALI items is around –0.80 (Table 4) indicating that items are on average easier and some students with high ability might not be challenged by most of these items. Therefore, if items are re-written or new items are included in the instrument, it is desired that they have a higher level of difficulty to ensure that there will be enough information to estimate students' abilities throughout all ability ranges.

Research about statistical literacy and reasoning is still evolving. This study and other recent ones like Callingham and Watson (2017) provide empirical evidence to help researchers, scholars, and instructors to better understand these constructs. As suggested by Gould (2017), however, these constructs also continue to be updated based on how the role of data in citizens' lives has changed. If the new demands to learn from modern data are leading instructors to re-think what it means for students to be statistically literate and to reason with statistical concepts, then there comes a need to update the teaching and assessment of these constructs. Further studies can explore how these constructs relate to data literacy, which is now being introduced in some introductory statistics courses. Investigations can also examine what aspects of statistical literacy and reasoning are present in introductory data science courses and how they differ from introductory statistics courses.

REFERENCES

Adams, B., Baller, D., Jonas, B., Joseph, A. C., & Cummiskey, K. (2021). Computational skills for multivariable thinking in introductory statistics. *Journal of Statistics and Data Science Education*, *29*(1), 123–131. https://doi.org/10.1080/10691898.2020.1852139

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

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

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, Handbook I: Cognitive domain*. David McKay.

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). SAGE Publications. https://doi.org/10.1177%2F0049124192021002005

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). John Wiley & Sons. https://doi.org/10.1002/9780470710470.ch9

Callingham, R., & Watson, J. M. (2017). The development of statistical literacy at school. *Statistics Education Research Journal*, *16*(1), 181–201. https://doi.org/10.52041/serj.v16i1.223

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

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

Chervaney, N., Collier, R. Fienberg, S., Johnson, P, & Neter, J. (1977). A framework for the development of measurement instruments for evaluating the introductory statistics course, *The American Statistician*, *31*(1), 17–23. https://doi.org/10.1080/00031305.1977.10479186

Chervaney, N., Benson, P. G., & Iyer, R. (1980), The planning stage in statistical reasoning. *The American Statistician*, *34*(4), 222–226. https://doi.org/10.2307/2684064

Cobb, G. (1992). Teaching statistics. In L. A. Steen (Ed.), *Heeding the call for change: Suggestions for curricular action* (pp. 3–43). The Mathematical Association of America.

Cummiskey, K., Adams, B., Pleuss, J., Turner, D., Clark, N., & Watts, K. (2020). Causal inference in introductory statistics courses. *Journal of Statistics Education*, *28*(1), 2–8. https://doi.org/10.1080/10691898.2020.1713936

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

delMas, R. C. (2002). Statistical literacy, reasoning and learning: A commentary. *Journal of Statistics Education, 10*(3), Article 1. https://doi.org/10.1080/10691898.2002.11910679

delMas, R. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 79–95). Kluwer Academic Publishers. https://doi.org/10.1007/1-4020-2278-6_4

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

GAISE College Report ASA Revision Committee. (2016). *Guidelines for assessment and instruction in statistics education report 2016*. https://www.amstat.org/education/guidelines-for-assessment-and-instruction-in-statistics-education-(gaise)-reports

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1–25. https://doi.org/10.2307/1403713

Gal, I. (2004). Statistical literacy. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 47–78). Springer. https://doi.org/10.1007/1-4020-2278-6_3

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

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. http://dx.doi.org/10.1111/j.1751-5823.2007.00029.x

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice.* Springer. https://doi.org/10.1007/978-1-4020-8383-9

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning*, *2*(1–2), 99–125. https://doi.org/10.1207/S15327833MTL0202_5

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

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423-436. https://doi.org/10.1007/BF02295430

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22–25. https://doi.org/10.52041/serj.v16i1.209

Gould, R. (2021). Toward data-scientific thinking. *Teaching Statistics*, *43*, S11–S22. https://doi.org/10.1111/test.12267

Hawkins, A., Jolliffe, F., & Glickman, L. (1992), *Teaching statistical concepts*. Longman Publishers. https://doi.org/10.4324/9781315845517

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229. https://doi.org/10.3102/1076998607302636

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

Hudiburgh, L. M., & Garbinsky, D. (2020). Data visualization: Bringing data to life in an introductory statistics course. *Journal of Statistics Education*, *28*(3), 262–279. https://doi.org/10.1080/10691898.2020.1796399

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). https://doi.org/10.1007/1-4020-2278-6_5

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000). A framework for characterizing children's statistical thinking. *Mathematics Thinking and Learning*, *2*(4), 269–307. https://doi.org/10.1207/S15327833MTL0204_3

Kaplan, J. J., & Thorpe, J. (2010). Post secondary and adult statistical literacy: Assessing beyond the classroom. Paper presented at the *Data and context in statistics education: Towards an evidence-based society*. In C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics* (ICOTS8), Ljubljana, Slovenia. International Statistical Institute. *https://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_5E3_KAPLAN.pdf*

Lovett, M. (2001). A collaborative convergence on studying reasoning processes: A case study in statistics. In S. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress.* Lawrence Erlbaum.

MacGillivray, H. (2021). Statistics and data science must speak together. *Teaching Statistics*, *43*, S5–S10. https://doi.org/10.1111/test.12281

Mallows, C. (1998). The zeroth problem. *The American Statistician*, *52*(1), 1–9. https://doi.org/10.2307/2685557

McKinley, R. L., & Reckase, M. D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space*. American College Testing Program. https://apps.dtic.mil/sti/pdfs/ADA137769.pdf

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Macmillan Publishing; American Council on Education.

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning*, *4*(1), 23–63. https://doi.org/10.1207/S15327833MTL0401_2

Moore, D. S. (1990). Uncertainty. In L.A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–138). National Academy Press.

Nisbett, R. (1993), *Rules for Reasoning.* Lawrence Erlbaum.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177%2F01466216000241003

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. https://doi.org/10.1177%2F0146621603027004004

Prodromou, T., & Dunne, T. (2017). Statistical literacy in data revolution era: Building blocks and instructional dilemmas. *Statistics Education Research Journal*, *16*(1), 38–43. https://doi.org/10.52041/serj.v16i1.212

R Development Core Team (2016). *R*: A language and environment for statistical computing [Computer software.] *R Foundation for Statistical Computing.* 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. https://doi.org/10.1177%2F014662168500900409

Reckase, M. D. (2009). *Multidimensional item response theory* (pp. 79–112). Springer. https://doi.org/10.1007/978-0-387-89976-3

Ridgway, J., Nicholson, J., & McCusker, S. (2013). Open data and the semantic web require a rethink on statistics teaching. *Technology Innovations in Statistics Education*, *7*(2). https://doi.org/10.5070/T572013907

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

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. https://doi.org/10.1007/BF02294360

Sedlmeier, P. (1999), *Improving statistical reasoning: Theoretical models and practical implication.* Lawrence Erlbaum. https://doi.org/10.1002/acp.749

Sharma, S., Doyle, P., Shandil, V., & Talakia'atu, S. (2011). Developing statistical literacy with Year 9 students. *Set: Research Information for Teachers, 1*, 43–50. https://doi.org/10.18296/set.0398

Sinharay, S., Puhan, G., Haberman, S. J., & Hambleton, R. K. (2018). Subscores: When to communicate them, what are their alternatives, and some recommendations. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 35–49). https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9781351136501-5/subscores-sandip-sinharay-gautam-puhan-shelby-haberman-ronald-hambleton

Snee, R. D. (1990). Statistical thinking and its contribution to total quality. *The American Statistician*, *44*(2), 116–121. https://doi.org/10.2307/2684144

Sugiura N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics: Theory and Methods*, *7*(1), 13–26. https://doi.org/10.1080/03610927808827599

Sylwester, D. (February,1993). Statistical thinking. *AMSTAT News.*

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education, 17*, 89–112. https://doi.org/10.1207/s15324818ame1702_1

Tintle, N. L., Topliff, K., VanderStoep, J., Holmes, V. L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal*, *11*(1), 21–40. https://doi.org/10.52041/serj.v11i1.340

Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician, 57*(2), 74–79. https://doi.org/10.1198/0003130031630

Watson, J., & 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

Watson, J. M., Collis, K. F., Callingham, R. A., & Moritz, J. B. (1995). A model for assessing higher order thinking in statistics. *Educational Research and Evaluation*, *1*(3), 247–275. https://doi.org/10.1080/1380361950010303

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

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

ANELISE SABBAG
California Polytechnic State University
asabbag@calpoly.edu

APPENDIX

Table 4. Estimates and standard errors for the item parameters for the cross-loadings (Model C) and unidimensional (Model D) models.

	Model C		Model D			Model C		Model D	
Item	Disc.	Difficulty	Disc.	Difficulty	Item	Disc.	Difficulty	Disc.	Difficulty
Literacy					*Reasoning*				
1	**0.67 (0.08)**	1.46 (0.07)	**0.71 (0.08)**	1.47 (0.08)	3	**0.80 (0.08)**	-0.70 (0.07)	0.96 (0.08)	-0.70 (0.07)
2	1.17 (0.10)	1.62 (0.09)	1.23 (0.10)	1.62 (0.09)	4	1.27 (0.13)	2.28 (0.13)	1.55 (0.13)	2.40 (0.13)
5	**0.44 (0.17)**	3.64 (0.18)	**0.49 (0.18)**	3.65 (0.18)	7	1.11 (0.11)	1.34 (0.09)	1.21 (0.09)	1.33 (0.08)
6	**0.56 (0.06)**	0.53 (0.06)	**0.59 (0.07)**	0.54 (0.06)	8	**0.76 (0.08)**	-0.75 (0.06)	0.93 (0.08)	-0.75 (0.07)
9	1.34 (0.11)	2.18 (0.12)	1.42 (0.12)	2.19 (0.12)	11	**0.58 (0.09)**	1.65 (0.08)	**0.67 (0.08)**	1.63 (0.08)
10	0.96 (0.08)	0.03 (0.06)	1.11 (0.08)	0.04 (0.07)	12	0.85 (0.08)	0.07 (0.06)	0.91 (0.07)	0.07 (0.06)
13	1.44 (0.11)	1.66 (0.10)	1.55 (0.11)	1.70 (0.10)	16	1.76 (0.18)	2.62 (0.17)	2.05 (0.16)	2.78 (0.16)
14	1.55 (0.13)	2.30 (0.13)	1.70 (0.13)	2.35 (0.13)	17	*0.75 (0.08)*	0.08 (0.06)	1.11 (0.08)	0.10 (0.07)
15	1.33 (0.09)	0.30 (0.07)	1.36 (0.09)	0.30 (0.07)	18	1.35 (0.13)	1.98 (0.12)	1.60 (0.12)	2.07 (0.12)
19	1.65 (0.13)	2.00 (0.12)	1.73 (0.13)	2.01 (0.12)	20	1.11 (0.10)	0.73 (0.07)	1.26 (0.09)	0.74 (0.07)
21	1.87 (0.14)	1.86 (0.12)	1.94 (0.13)	1.86 (0.12)	23	1.69 (0.16)	2.12 (0.13)	2.09 (0.15)	2.33 (0.14)
22	1.21 (0.11)	2.03 (0.11)	1.18 (0.10)	1.97 (0.10)	24	0.91 (0.09)	0.89 (0.07)	1.15 (0.09)	0.93 (0.07)
25	1.56 (0.11)	1.12 (0.09)	1.56 (0.10)	1.11 (0.09)	27	0.88 (0.09)	0.61 (0.07)	1.13 (0.08)	0.64 (0.07)
26	1.38 (0.10)	0.73 (0.08)	1.40 (0.09)	0.73 (0.08)	28	**0.54 (0.08)**	0.78 (0.06)	**0.72 (0.07)**	0.80 (0.06)
29	1.14 (0.09)	0.05 (0.07)	1.24 (0.08)	0.06 (0.07)	31	1.41 (0.12)	0.54 (0.08)	1.61 (0.10)	0.57 (0.08)
30	**0.46 (0.06)**	0.42 (0.06)	**0.53 (0.06)**	0.43 (0.06)	32	0.98 (0.09)	0.21 (0.06)	1.13 (0.08)	0.21 (0.07)
34	1.18 (0.09)	0.38 (0.07)	1.28 (0.09)	0.40 (0.07)	33	**0.61 (0.08)**	0.86 (0.06)	**0.70 (0.07)**	0.85 (0.06)
36	0.81 (0.08)	1.45 (0.08)	0.81 (0.08)	1.44 (0.08)	35	**0.69 (0.08)**	-0.06 (0.06)	**0.71 (0.07)**	-0.06 (0.06)
37	1.69 (0.21)	4.19 (0.29)	1.93 (0.23)	4.32 (0.29)	39	0.83 (0.08)	-0.09 (0.06)	0.93 (0.07)	-0.08 (0.06)
38	1.10 (0.09)	1.06 (0.08)	1.18 (0.09)	1.07 (0.08)	40	0.88 (0.09)	-0.45 (0.06)	1.02 (0.08)	-0.44 (0.07)

Note: Items with discrimination estimates smaller than 0.80 were bolded. The "statistical reasoning" items in Model B were cross-loaded to the "statistical literacy" construct with a fixed discrimination of 0.2 so that literacy's effect on the reasoning items was smaller than the direct effects from statistical reasoning.