ASSESSING STATISTICAL LITERACY AND STATISTICAL REASONING

Anelise Sabbag, Joan Garfield and Andrew Zieffler
California Polytechnic State University (San Luis Obispo), United States
University of Minnesota, United States
asabbag@calpoly.edu

Statistical literacy and statistical reasoning have been considered by the statistics education community as important learning goals to be developed in statistics courses (Garfield & Ben-Zvi, 2008). Many statistics educators and scholars have tried to define these learning goals (e.g., Gal, 2002; Watson & Callingham, 2003; Garfield, 2002; Garfield & Chance, 2000). However, there is a lack of agreement regarding the relationship between them. This study will report on the development process of the REAsoning and LIteracy (REALI) assessment, including the several types of validity evidence that were gathered to support the intended inferences and uses of the instrument's scores. REALI was developed to concurrently assess statistical literacy and reasoning and to be used as a tool to investigate the relationship between these learning goals.

INTRODUCTION

Statistical literacy and statistical reasoning are important learning goals to be developed in introductory statistics courses and many statistics educators and scholar have tried to define these terms (e.g., Gal, 2002; Budgett & Pfannkuch, 2007; Rumsey, 2002; Watson & Callingham, 2003; Garfield, 2002; Garfield & Chance, 2000; Jones et al., 2004; Chance, 2002). Often times, there is a great overlap in the definitions of these terms. In addition, assumptions of a hierarchy between and within these learning goals has been posed by some researchers (delMas, 2002; Chance, 2002; Garfield & Ben-Zvi, 2007 and 2008).

Despite the evidence of a possible overlap between the outcomes of statistical literacy and reasoning, there is no assessment measuring these learning goals concurrently and no empirical study has been done to examine the relationship between statistical literacy and reasoning. The assessments of statistical reasoning and statistical literacy that are currently available were developed independently without considering the possible overlap between these learning goals. To investigate this overlap, it is necessary to develop one instrument composed of statistical literacy items and statistical reasoning items. In this way, it will be possible to obtain one subscore for each learning goal and use measurement analysis to explore if statistical literacy and statistical reasoning could be measured reliably and distinctly or if these two learning goals are actually so similar that they cannot be distinguished. Such an assessment may help to clarify the structure of the relationship between statistical literacy and statistical reasoning.

ASSESSMENT DEVELOPMENT

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: the REAsoning and LIteracy (REALI) instrument. As suggested by the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999), careful attention was given to how the scores from REALI would be interpreted. To support the intended inferences and uses of the REALI's scores several types of validity evidence were gathered throughout the development process: working definitions, test blueprint, expert reviews, response process interviews with students, a pilot test, a field test, and psychometric analysis.

According to Standard 1.2 from the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999), the construct being measured by the test should be clearly defined. Therefore, working definitions of statistical literacy and statistical reasoning items were established based on the definitions from Ziegler (2014), Garfield and Ben-Zvi (2008), and delMas (2002, 2004):

- *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).

- *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.

The development process continued with the elaboration of a test blueprint. The REALI assessment was initially based on items from the GOALS instrument (Sabbag & Zieffler, 2015), which was designed to measure statistical reasoning, and items from BLIS (Ziegler, 2014), which was designed to measure statistical literacy. Items were grouped into areas of learning based on their content. Eight areas of learning were identified: (1) representations of data, (2) measures of center, (3) measures of variability, (4) study design, (5) hypothesis testing and *p*-values, (6) confidence intervals, (7) bivariate data, and (8) basic probability.

Further investigations were done to verify if the items aligned with the working definitions of statistical literacy and statistical reasoning used in this study. The behaviors needed to answer each item correctly were identified (see Figure 1) and used to classify each item as a statistical literacy item or a statistical reasoning item. This categorization of items was also performed by four experts in the field of statistics education as part of the expert review. In addition, these experts were also invited to critique the items. As specified by Standard 3.5 (AERA, APA, & NCME, 1999), feedback from the experts was used as validity evidence to assure items were categorized consistently and to improve quality of items.

ITEM: 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

Figure 1. Statistical literacy item and behaviors.

Additional validity evidence was gathered from think-aloud interviews with students. These interviews were conducted to better understand how students would respond to each of the items in REALI. Four students who had taken an introductory statistics course in the previous semester participated in the think-aloud interviews: three students from the Educational Psychology Department and one student from the Statistics Department at the University of Minnesota. Their responses were used to verify if items were behaving as intended and to clarify the categorization of some items which could not be categorized with certainty as a statistical literacy or a statistical reasoning item. Based on the students' responses, unclear/confusing items and items that were misinterpreted by the students were identified and modified.

A pilot test was the next step in the development process. A total of 237 students from two introductory statistics courses at the University of Minnesota and an introductory statistics course at Augsburg College completed the REALI assessment as part of the pilot study. A psychometric analysis of the instrument was performed to assess how the items were behaving. Because there were more items than needed in this version of the instrument, item difficulty, item discrimination, and distractor analysis were used to decide which items to delete. Additional changes were made to reduce unnecessary item complexity, to ensure that distractor choices were plausible, and to ensure that items had independent/non-overlapping choices. These changes resulted in the final version of the assessment which was used in the field test. The final version of REALI was composed of 20 items measuring statistical literacy and 20 items measuring statistical reasoning.

ASSESSMENT IMPLEMENTATION

The field test was performed at the end of the spring semester during the months of April and May of 2016. A total of 671 students from introductory statistics courses took the REALI assessment and consented to participate in the study. Those students represented 16 universities and colleges around the United States and Canada.

A histogram of the distribution of the REALI total scores for the 671 students in the sample is presented in Figure 2. The mean score was 24.16 and the standard deviation was 7.48. The estimate of the internal consistency, coefficient alpha, for these scores was 0.87.

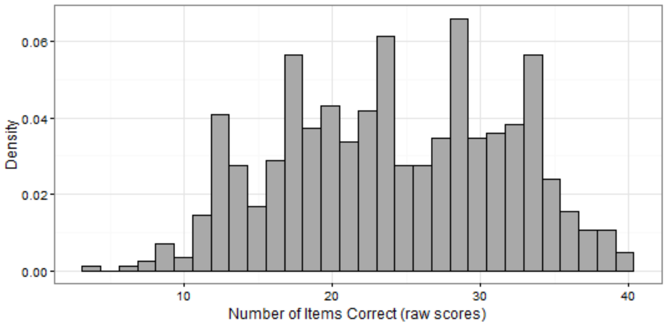

Figure 2. Distribution of total scores.

The statistical literacy and statistical reasoning subscores were investigated to better understand how students were performing under each construct. Histograms of the distributions of the two subscores for the students in the sample are presented in Figure 3. The mean statistical literacy subscore was 13.15 with standard deviation of 3.82, and the mean statistical reasoning subscore was 11.01 with standard deviation of 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. The statistical literacy and statistical reasoning subscores had a correlation of 0.76. Figure 4 shows the scatterplots of the two subscores.

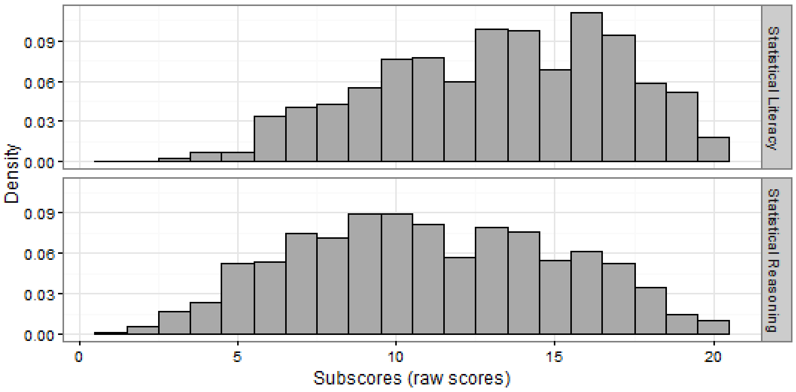

Figure 3. Distribution of the statistical literacy and statistical reasoning subscores.

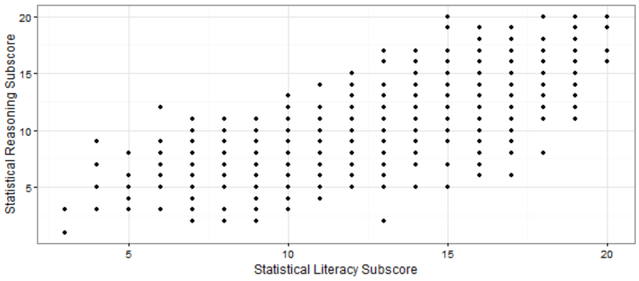

Figure 4. Scatterplots of the statistical literacy and statistical reasoning subscores.

Item Response Theory

Five IRT models were fitted to the data to better understand what measurement model best represents the construct of statistical literacy and the construct of statistical reasoning: a unidimensional model, three bi-dimensional models (uncorrelated model, correlated model, cross-loading model), and a bi-factor model. Model comparisons, reported in Sabbag (2016), suggested that the most useful model to represent the constructs of statistical literacy and statistical reasoning were the cross-loading model (see Figure 5). The cross-loading model was composed of two uncorrelated dimensions: a statistical literacy dimension and a statistical reasoning dimension. In addition, this model used direct effects from the statistical literacy dimension to the statistical reasoning items. The cross-loading model assumed that these direct effects would be the same for all items and they would be smaller than all the effects of the statistical reasoning dimension on the statistical reasoning items. In other words, the statistical reasoning dimension would have the highest effect on the statistical reasoning items.

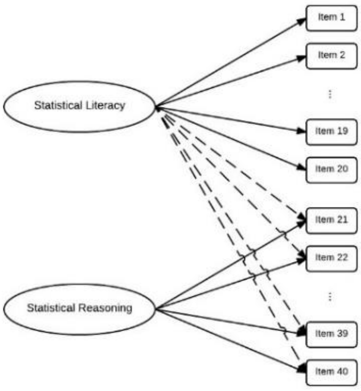

Figure 5. Bi-dimensional IRT models.

The assumptions for the cross-loading model support 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.

Figure 6 shows the scatterplot of the statistical literacy subscore and statistical reasoning subscore produced by the Cross-loading model. The correlation between these subscores was 0.640. The estimate of empirical reliability (Zimowski, Muraki, Mislevy, and Bock, 2003) for the statistical literacy subscore was 0.75 and for the statistical reasoning subscore was 0.70.

The Haberman analysis (Haberman, 2008) was used to evaluate if reporting two subscores, one for statistical literacy and one for statistical reasoning, had added value over reporting the total score from a unidimensional model. Among other things, this analysis considers the reliability of each subscore and how correlated the subscores are with each other and with the total score. The cross-loading model presented evidence of distinction and evidence that the statistical literacy and statistical reasoning scores can be measured reliably. In addition, this model presented evidence that the subscores provide information that is over and above the information provided by the total unidimensional score (see Sabbag, 2016 for more information). However, the usefulness of subscores can only be applied in the IRT subscore and not for raw scores.

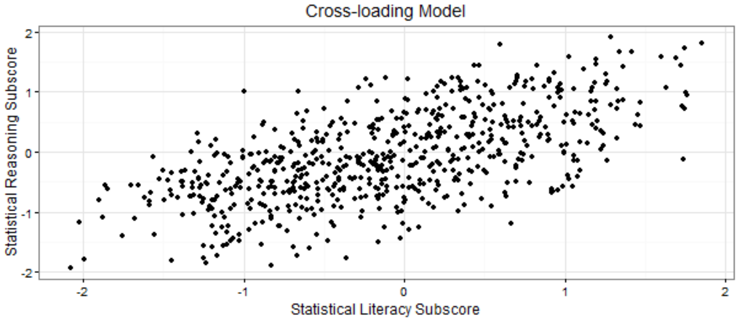

Figure 6. Scatterplot of the statistical literacy and statistical reasoning subscores for the Cross-loading model.

CONCLUSION AND IMPLICATION FOR TEACHING

This study built a validity argument supporting the intended inferences and uses of the subscores from the REALI instrument and its ability to measure students' statistical literacy and statistical reasoning in an introductory statistics course.

Recommendations in the field of statistics education have emphasized the importance of developing statistical literacy and statistical reasoning rather than computations and procedures (e.g., GAISE-ASA, 2005; Garfield & Ben-Zvi, 2008). The results from this study provided preliminary evidence that statistical literacy and statistical reasoning can be differentiated. Therefore, instructors need to set clear learning goals of statistical literacy and also learning goals of statistical reasoning in their classes. Instructors also need to show evidence, through the use of well-developed assessments, that students are indeed achieving these goals. In addition, because of the hierarchy between the statistical literacy and reasoning, it is important for instructors to note that developing students' statistical reasoning is not a straightforward step from statistical literacy. Firstly, students need to develop a certain level of statistical literacy to then be guided on how to make connections and relate the different statistical concepts they learned. Thus, instructors need to provide opportunities for students to learn how to reason with statistical concepts.

The REALI instrument can be a tool for identifying students' misconceptions and guiding changes and improvements in statistics courses. This instrument can be used to provide valuable information about students' performance on important statistical literacy and statistical reasoning topics. For instance, the content of two of the hardest items in the REALI instrument (with less than 30% of correct answers) involved graphs and the normal distribution. Further attention can be given to understand how these topics have been taught in the curriculum and why students are incorrectly answering these items. Looking at each of the distractors will give insight to possible students' misconceptions or need of curriculum improvement. For example, the curriculum might be over-emphasizing the normal distribution leading students to choose always a graph that looks more normal, without giving careful thought to the other information available in the problem.

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. For instance, there have been efforts to change introductory courses based on the recommendations by Cobb (2005, 2007) and GAISE (ASA, 2005). New statistics curricula have been developed (e.g., Garfield, delMas, Zieffler, 2012 and Tintle et al., 2011) using a modelling and simulation approach to teaching inference. However, these new curricula differ. Some curricula still teach traditional content such as t-tests, on the other hand, curricula such as the CATALST course (Garfield et al., 2012) do not teach traditional content; instead it focuses on randomization tests. It is important to examine how well students are performing on these new curricula and evaluate if there is a curriculum that is leading to better student's performance. A possible way to compare students from these different curricula is to use the REALI instrument at the end of the course and investigate if students are answering the questions in the same way or if the curricula they are in affects how students reason about statistical concepts leading to different answers.

REFERENCES

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

American Statistical Association. (2005). Guidelines for assessment and instruction in statistics education. *Published on the World Wide Web at http://www.Amstat.org/education/gaise.*

Budgett, S., & Pfannkuch, M. (2007). Assessing students' statistical literacy. *Assessment Methods in Statistical Education: An International Perspective*, *19*, 103.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10(3), 1-17.

Cobb, G. W. (2005). The introductory statistics course: A saber tooth curriculum. In talk presented at the *United States Conference on Teaching of Statistics, Columbus, OH*.

Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? *Technology Innovations in Statistics Education*, 1(1). Retrieved from www.escholarship.org/uc/item/6hb3k0nz.

delMas, R. C. (2002). Statistical literacy, reasoning and learning: A commentary. *Journal of Statistics Education*, *10*(3).

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.

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

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, 10(3), http://www.amstat.org/publications/jse/v10n3/garfield.html

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., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary- level statistics course. *ZDM, 44*(7), 883-898. doi:10.1007/s11858-012-0447-5

Haberman, S. J. (2008). When can subscores have value. *Journal of Educational and Behavioral Statistics*, *33*(2), 204-229.

Jones, G. A., Langrall, C. W., Mooney, E. S., & Thornton, C. A. (2004). Models of development in statistical reasoning. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 97-117). Springer Netherlands.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3), 6-13.

Sabbag, A. (2016). *Examining the relationship between statistical literacy and statistical reasoning* (Unpublished doctoral dissertation). Retrieved from the University of Minnesota Digital Conservancy.

Sabbag, A., & Zieffler A. (2015). Assessing Learning Outcomes: An analysis of the GOALS-2 instrument. *Statistics Education Research Journal*, *14*(2), 93–116.

Tintle, N., VanderStoep, J., Holmes, V. L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, *19*(1), n1.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3-46.

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.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (2003). BILOG-MG. Scientific Software lnternational.