QUALITY ASSESSMENTS IN STATISTICS EDUCATION: A FOCUS ON THE GOALS INSTRUMENT

Anelise Sabbag, Joan Garfield, and Andrew Zieffler
University of Minnesota, United States
sabb0013@umn.edu

Research in statistics education and measurement suggest the use of quality instruments with good psychometric characteristics to measure students' learning outcomes. However, very few quality instruments have been developed, evaluated, and are available to researchers to measure students' statistical knowledge. This paper reports the development of the Goals and Outcomes Associated with Learning Statistics (GOALS) instrument, which was designed to measure statistical reasoning learning outcomes in a first course in statistics.

BACKGROUND

Statistics education has experienced many changes in the last decade, one of which is a shift in the focus of learning outcomes in introductory statistics courses. After calls for change by the American Statistical Association (see GAISE report—ASA, 2005) and by key visionaries such as Cobb (2005, 2007), introductory statistics courses are shifting more to an emphasis on thinking and reasoning than on calculation and procedures. Some of these courses (e.g., Garfield, delMas, Zieffler, 2012; Tintle, VanderStoep, Holmes, Quisenberry and Swanson, 2011) are also using a modelling and simulation approach to teaching statistics instead of the conventional parametric methods approach.

As new curriculum are developed and implemented, it is important to use high quality assessment instruments to evaluate the learning outcomes for students and to study the impact of the new courses on these desired outcomes. Garfield, delMas, and Zieffler (2010) suggest methods of designing and creating assessments to use for evaluating curriculum and stressed the importance of the alignment between the assessment and the learning outcomes of the course.

A variety of instruments have been developed to assess the learning outcomes of statistical literacy, reasoning, and thinking. Examples of these instruments include (1) the *Statistical Reasoning Assessment* (SRA; Garfield, 1998b, 2003), (2) the *Comprehensive Assessment of Outcomes in a First Statistics Course* (CAOS; delMas Garfield, Ooms & Chance, 2007), and (3) the *Basic Literacy In Statistics* (BLIS; Ziegler, 2014). Despite all this information available in the literature, Zieffler, Garfield, Alt, Dupuis, Holleque, and Chang (2008) point out that there are still many studies using inappropriate measures to assess student learning, such as final exam scores or course grade. Other measurement issues have been reported such as the "lack of enough diversity in available validated tests to allow good alignment between existing tests and intended outcomes in a particular research study" (Pearl, Garfield, delMas, Groth, Kaplan, McGowan, & Lee, 2012). This suggests that there is a need in the field of statistics education for psychometrically sound assessments that measure different learning outcomes for introductory statistics courses. This paper described the creation and analysis of one such instrument, the GOALS assessment.

THE IMPORTANCE OF QUALITY ASSESSMENTS

Assessments are of vital importance when used in educational research. In a report detailing standards for quality research in mathematics and statistics education (ASA, 2007), the authors state that every assessment should develop and report

- 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;
- Information regarding the population of interest that the assessment will be administered to, 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) was obtained; and

In: M.A. Sorto (Ed.), *Advances in statistics education: developments, experiences and assessments. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),* July 2015, Rio de Janeiro, Brazil. iase-web.org/Conference_Proceedings.php

- Evidence 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.

Clearly defining and distinguishing between the learning outcomes to be assessed can help in the development of quality assessments (Garfield and Ben-Zvi, 2008). One way of aligning an instrument to important learning outcomes is to create a test blueprint. A test blueprint lays out the content of the test and its relationship with the construct being measured by the test.

Thorndike (Thorndike & Thorndike-Christ, 2010) reported two other qualities that are desirable in an instrument: reliability and validity. Thorndike defined reliability as "the accuracy or precision of a measurement procedure" (p.118). Thus, reliability can be seen as the extent to which the test scores are precise. It is a measure of how much variance in the test scores is true variance and not measurement error. The second quality, mentioned by Thorndike was validity. The definition of validity given in the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) is "the degree to which evidence and theory support the interpretation of the test scores entailed by proposed uses of tests" (p. 9). Validity is considered an intrinsic part in the process of developing an assessment. According to the Standards for Educational and Psychological Testing there is one type of validity but many different types of validity evidence. Validity evidence can be based on test content, response processes, relationships with other variables, internal structure, and consequences of testing. Validation is the process of gathering validity evidence to support the intended inferences and uses of the test scores. The type of validity evidence collected depends on what the test is intended to measure and thus depends on the intended inferences and uses of test scores.

The following sections describe the creation and revision of the *Goals and Outcomes Associated with Learning Statistics* (GOALS) test to illustrate the use of these principles in designing and evaluating an assessment instrument. Some of the development information is provided in this paper but the final data analyses will be included in the actual presentation and will be available after the conference in a larger version of this paper.

METHODOLOGY

GOALS was originally envisioned as an updated version of the *Comprehensive Assessment of Outcomes in a First Statistics Course* (CAOS). The first version of GOALS was developed to evaluate the effectiveness of the *Change Agents for Teaching and Learning Statistics* (CATALST) course in helping students develop statistical literacy and reasoning (see Garfield, delMas, Zieffler, 2012). Analysis of CAOS data and examination of alignment with current learning goals for introductory statistics courses were used to create GOALS, which was administered to students taking either a CATALST course or a more conventional parametric course. The development process continued based on two rounds of administration and data analysis, revisions, and then feedback from experts. After two years of revision, data collection, and psychometric analyses, GOALS-4 was developed to focus primarily on assessing statistical reasoning. Table 1 shows the test blueprint with the reasoning learning goals for each item.

Table 1. GOALS-4 blueprint

Item	Measured Learning Goal
1	Able to reason about the purpose of random assignment
2	Able to reason about the factors that allow a sample of data to be representative of the population.
3	Able to reason that random assignment is needed to make a causal statement.
4	Able to reason about the effect of moving an influential point in a scatterplot to a new location on the correlation coefficient
5	Able to reason about factors that affect the mean and median
6	Able to reason that a large *p*-value does not provide significant evidence of an effect.
7	Able to reason about the meaning of variability in the context of repeated measurements

Item	Measured Learning Goal
	and in a context where small variability is desired.
8	Able to reason that given two distributions that have the same range, the one with less mass in the center has the larger standard deviation (tests for misconception that a distribution has less "variability" than a non-uniform distribution)uniform
9	Able to reason about how differences in variability affect strength of evidence against the null hypothesis of no difference
10	Able to reason about differences between distributions of sample proportions for large and small sample sizes
11	Able to reason about how the width of a confidence interval is related to sample size.
12	Able to reason about data as an aggregate that has characteristics of shape, center, and variability.
13	Able to reason about a misinterpretation of a confidence level (using it to make a prediction for a single case)
14	Able to reason that a smaller *p*-value provides stronger evidence against the null hypothesis than a larger *p*-value.
15	Able to reason about what model should be used for the null hypothesis when comparing two groups
16	Able to reason about what the null model represents in a research study
17	Able to reason about a conclusion based on a statistically significant *p*-value in the context of a research study that compares two groups
18	Able to reason about an incorrect interpretation of a *p*-value (probability of a treatment being more effective).
19	Able to reason about the visual depiction of confidence intervals if there is evidence of a difference between groups.
20	Able to reason about how increasing the sample size affects the *p*-value, all else being equal.

RESULTS

GOALS-4 was administered to students in a large filed test in the Fall 2014. The current form of GOALS consists of 20 forced-choice items given in an online format, either during or outside of class. Data were collected from 1,109 undergraduate students from 19 courses in 17 different higher education institutions.

A histogram of the distribution of the GOALS-4 total scores is presented in Figure 1. The mean of these scores was 9.35 (SD = 3.42). The minimum and maximum values observed were 2 and 20 respectively.

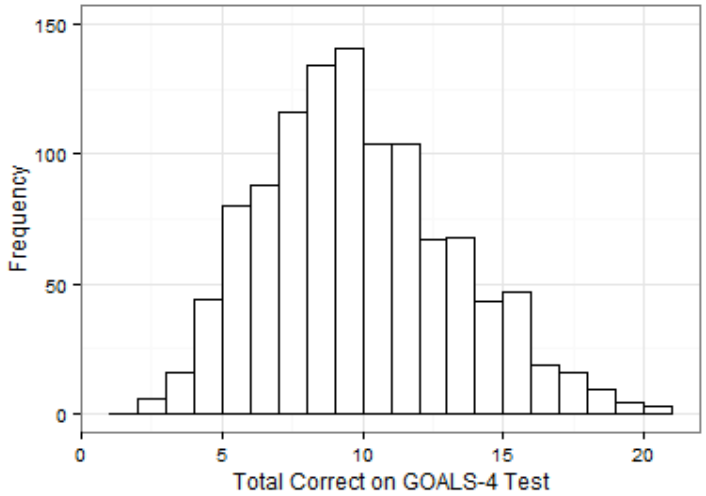

Figure 1. The distribution of GOALS-4 scores for 1,109 students

Data on which incorrect response where selected also provide valuable information on how students reason or are misunderstanding the concepts included in the assessment. Table 2 lists the percent correct for each item. The percent correct ranged from 15.6% (Item 3) to 93.2% (Item 7).

Table 2. Percent correct and item discrimination for all items in GOALS-4

Item	Percent correct	Item discrimination	Item	Percent correct	Item discrimination
1	23.6%	0.241	11	51.1%	0.385
2	63.0%	0.286	12	31.9%	0.272
3	15.6%	0.178	13	46.3%	0.178
4	45.5%	0.213	14	45.2%	0.301
5	73.1%	0.337	15	34.0%	0.185
6	68.3%	0.055	16	31.1%	0.355
7	93.2%	0.125	17	48.1%	0.196
8	36.4%	0.345	18	41.9%	0.243
9	65.5%	0.317	19	28.2%	0.119
10	48.1%	0.271	20	44.9%	0.205

Item discrimination (point biserial correlation corrected for spuriousness) was computed for each of the GOALS-4 items under the Classical Test Theory framework. These values are presented in Table 2. Item discriminations ranged from 0.06 to 0.39.

Results of full psychometric analyses will be presented at the time of the presentation. This information will include description of the structure and reliability of test-scores for GOALS-4, under the classical test theory framework. In addition, item responses will be analyzed to inform about students' statistical reasoning strengths and weakness.

DISCUSSION OF RESULTS AND IMPLICATIONS FOR FUTURE RESEARCH

Overall, the students found the test to be extremely challenging. Students answered, on average, only 9 of the 20 items correct. According to Table 2, the most difficult items in the GOALS test were Item 3 and Item 1, which were both related to the concept of random assignment. To correctly answer Item 3 students needed to reason about how correlation does not imply causation and that random assignment is needed to make a causal statement. Only 15.6% of the students correctly answered this item. Even though this was a very difficult item, its discrimination was moderate (0.178) indicating that this item was low/moderately discriminating among students with high and low ability. Item 1, was the second most difficult item with only 23.6% of the students answering it correctly. This item assesses if students are able to reason about the purpose of random assignment. Its discrimination value was high (0.241) which means that the few students who got Item 1 correct were the ones that presented a higher total score on the GOALS-4 test.

The easiest item on the GOALS-4 test (Item 7) measured if students were able to reason about the meaning of variability. The percent correct for this item was 93.2% and it had a poor discrimination (0.125). The item that was worst at discriminating students with high and low ability was Item 6, which presented the result of a *t*-test and assessed if students were able to reason that a large *p*-value does not provide significant evidence of an effect. Around 68% of the students correctly answered this item; however, it is most likely that those students were the one with lower ability.

GOALS-4 will be widely available for use in future research and evaluation studies, and may be translated into other languages as well. It may be used in connection with other assessments such as the measures of statistical literacy and statistical thinking. The development of assessments such as GOALS-4 allows instructors to use quality instruments to learn about

students' statistical knowledge and in turn to inform and modify their teaching. Additionally, researches are also able to use such instruments to perform high-level research.

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). College Report. *Guidelines for assessment and instruction in statistics education*. Washington DC: Author

American Statistical Association. (2007), "*Using Statistics Effectively in Mathematics Education Research,"* Retrieved Nov. 02, 2013, from ASA Web site: http://www.amstat.org/education/pdfs/UsingStatisticsEffectivelyinMathEdResearch.pdf

Cobb, G. W. (2005). The introductory statistics course: A saber tooth curriculum. In talk presented at *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), Online: www.escholarship.org/uc/item/6hb3k0nz.

delMas, R. C., Garfield, J.B., Ooms, A., & Chance, B. L. (2007). Assessing Students' Conceptual Understanding after a First Course in Statistics. *Statistics Educational Research Journal*, *6*(2), 28–58. Retrieved September 10, 2010, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ6(2)_delMas.pdf

Garfield, J. (1998). The statistical reasoning assessment: Development and validation of a research tool. Paper presented at the *Proceedings of the 5th International Conference on Teaching Statistics*.

Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, 2(1), 22-38.

Garfield, J. B., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer.

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

Pearl, D. K., Garfield, J. B., delMas, R., Groth, R. E., Kaplan, J. J., McGowan, H., & Lee, H. S. (2012). Connecting Research to Practice in a Culture of Assessment for Introductory College-level Statistics. Online: http://ww.causeweb.org/research/guidelines/ResearchReport_Dec_2012.pdf

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.

Thorndike, R.M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Upper Saddle River, NJ: Pearson Education.

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). http://www.amstat.org/publications/jse/v16n2/zieffler.pdf

Ziegler, L. (2014). Reconceptualizing statistical literacy: developing an assessment for the modern introductory statistics course. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/165153.