DEVELOPMENT OF AN INSTRUMENT TO ASSESS STATISTICAL THINKING

Andrew Zieffler, Joan Garfield, Robert delMas and Auðbjörg Björnsdóttir
University of Minnesota, United States of America
zief0002@umn.edu

The Comprehensive Assessment of Outcomes in a First Statistics course (CAOS) test consists of 40 multiple-choice items that were judged by a group of Statistics Education experts in 2004 to cover important learning outcomes for a first course in statistics (delMas, Garfield, Ooms & Chance, 2007). More currently, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) college report suggested several important learning goals for students enrolled in an introductory statistics course which have been endorsed by the American Statistical Association. This paper describes the development of a new instrument to assess the desired student learning outcomes presented in the GAISE college report. This paper discusses the process used to select and add items, which was based not only on content analysis but also on psychometric methods (e.g., item response theory, differential item functioning).

INTRODUCTION

In 1997 Gal and Garfield issued a challenge to the statistics education community to produce and use better assessments of student outcomes in teaching and research (Gal & Garfield, 1997). One project that attempted to respond to this challenge was ARTIST (Assessment Resource Tools for Improving Statistical Thinking) funded by the National Science Foundation (DUE-0206571). Over a five-year period a team of researchers and advisers developed assessments that could be used in classroom or research settings to evaluate students' statistical literacy, reasoning, and thinking. One of the assessments developed as a part of the ARTIST project was the Comprehensive Assessment of Outcomes in Statistics (CAOS).

DEVELOPMENT OF THE CAOS ASSESSMENT

All of the items on CAOS were intended to measure concepts and learning outcomes that any student completing an introductory statistics course would be expected to understand. The current version of CAOS, CAOS-4, includes 40 multiple-choice items, which have gone through a systematic and extensive process of testing and revision over a three-year period. Psychometric evidence was collected continually to inform the decisions made throughout the process. A summary of this process is described below. The focus of the CAOS test was on students' basic literacy and reasoning, particularly regarding important concepts such as distribution, center, and variability. Multiple-choice items were initially either selected from a database containing over 1000 items (another resource developed by the ARTIST project) or created to assess a particular concept. These items were revised to adhere to item writing guidelines (e.g., Haladyna, Downing, & Rodriguez, 2002) and then further revised based on feedback from the advisory board and from class testing. For a description of the process of developing, validating, and testing this instrument, see delMas, Garfield, Ooms and Chance (2007).

ITEM ANALYSIS OF THE CAOS-4 ASSESSMENT

Item analysis is an essential part of the test development process (Livingston, 2006). It is a statistical analysis of the responses of individual test takers for each individual item, with the explicit purpose of gaining information about the items rather than the people taking the test. This type of analysis provides important information on the *difficulty*, *discrimination*, and potential *differential functioning* for each item.

Item difficulty indicates how difficult an item is. Difficulty is important since including too many items at either end of the difficulty spectrum provides poor information regarding the inferences that a researcher or teacher is trying to make regarding her/his students' understanding. Discrimination refers to the predisposition of a test taker to answer an item correctly when s/he possesses the knowledge or ability in the content area that the item is designed to measure and also the tendency for students who do not possess that knowledge to answer incorrectly. Items that do not discriminate well need to be examined for possible ambiguity or problems with the distractors.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

Lastly, differential item functioning (DIF) refers to the potential for an item(s) to function differently for different subgroups of test takers. For example, an item functions differentially when a test item is more difficult for a particular subgroup than is expected given the general difficulty of the item.

Research on test development and item analysis have suggested that methods using Item Response Theory (IRT) provide an alternative theoretical framework to Classical Test Theory (CTT) in estimating the characteristics of assessment items (Barnard, 1999; Lord, 1980; Yen & Fitzpatrick, 2006). Using IRT methods complementary with CTT methods–which were used on all previous versions of the CAOS assessment–provides the maximum information to test developers about the items and their contributions to the assessment in question.

Using the responses from 6,111 students who completed the CAOS-4, item analyses were undertaken to study the items that compose the CAOS-4. Initially an investigation into the underlying number of dimensions was conducted. Based on this investigation, different IRT models were examined for fit. The three-parameter IRT model was settled on and fitted to the data using a bi-factor structure. The parameters for each item are provided in Table 1.

Table 1. Item parameters for each item in the CAOS-4

Item	Discrimination	Difficulty	Guessing	Item	Discrimination	Difficulty	Guessing
1	0.474	-1.194	0.276	22	2.077	1.005	0.411
2	1.142	1.541	0.445	23	1.209	1.034	0.500
3	2.226	-0.464	0.172	24	0.502	0.115	0.174
4	1.803	-0.025	0.196	25	5.875	1.098	0.466
5	2.700	-0.298	0.176	26	1.518	0.991	0.444
6	2.735	1.172	0.103	27	4.189	1.046	0.429
7	2.693	1.924	0.068	28	2.644	1.079	0.360
8	1.216	0.798	0.430	29	1.518	0.734	0.422
9	2.286	1.326	0.103	30	6.370	1.939	0.398
10	2.746	1.089	0.138	31	0.813	-0.068	0.500
11	1.684	-1.454	0.142	32	6.671	1.916	0.158
13	1.073	-1.668	0.179	33	1.900	1.499	0.296
14	1.797	-0.158	0.372	34	1.132	0.592	0.420
15	1.823	0.276	0.155	35	1.942	1.146	0.318
16	2.430	1.359	0.423	36	1.705	0.763	0.327
17	2.291	0.838	0.115	37	2.521	1.637	0.147
18	1.447	0.604	0.181	38	1.763	1.160	0.187
19	0.803	-1.431	0.188	39	3.903	1.624	0.157
20	2.188	0.637	0.447	40	1.599	0.623	0.261
21	0.688	-1.750	0.242				

Using IRT and applying a full information factor analysis, the responses of the CAOS-4 were analyzed for DIF based on sex. Eight items were identified as exhibiting DIF, with the male subgroup having a higher probability of getting an item correct for all ability levels.

GUIDELINES FOR ASSESSMENT AND INSTRUCTION IN STATISTICS EDUCATION (GAISE)

In 2005, the American Statistical Association (ASA) endorsed a set of proposed guidelines for assessment and instruction of the introductory undergraduate level statistics course. These guidelines were developed based on discussions and reviews of "existing standards and guidelines, relevant research results from the studies of teaching and learning statistics, and recent discussions and recommendations regarding the need to focus instruction and assessment on the important concepts that underlie statistical reasoning" (ASA, 2005).

The guidelines, which do not focus on content, represent knowledge and understanding gained through the discipline of thinking statistically which were deemed important outcomes for any introductory tertiary statistics course. Examples of the 22 outcomes listed in the GAISE report are that students should believe and understand: why data beat anecdotes; association is not

causation; how to determine when a cause and effect inference can be drawn from an association, based on how the data were collected; and how to make appropriate use of statistical inference. The full report and list of learning goals along with specific recommendations for helping achieve those goals is available at http://www.amstat.org/education/gaise/.

USING GAISE TO STRUCTURE A NEW CAOS TEST

Given the need to reexamine the CAOS test based on the item-level analysis and also the outcomes enumerated in the GAISE report, the decision was made to revise the CAOS test to meet both needs. The first step in this process is to align the items from CAOS-4 with the learning outcomes documented in the GAISE report. Alignment is a process to judge whether or not there is a match between the assessment being used and the outcomes that are desired. To create this mapping, each item of CAOS-4 that was retained through the IRT analysis was examined and matched to one or more of the learning outcomes. This alignment was discussed and consensus was reached between eight raters who were familiar with both the GAISE report and the CAOS-4 assessment. Table 2 shows the result of these mappings.

Table 2. The mapping of the CAOS-4 items to the learning outcomes documented in the GAISE report for a statistically educated student

CAOS-4 item(s)	GAISE learning outcome
	Students should believe and understand why:
---	Data beat anecdotes.
#15(?)	Variability is natural and is also predictable and quantifiable.
#38	Random *sampling* allows results of surveys and experiments to be extended to the population from which the sample was taken.
#7, #24	Random *assignment* in comparative experiments allows cause and effect conclusions to be drawn.
#22	Association is not causation.
---	Statistical significance does not necessarily imply practical importance, especially for studies with large sample sizes.
#23	Finding no statistically significant difference or relationship does not necessarily mean there is no difference or no relationship in the population, especially for studies with small sample sizes.
	Students should recognize:
#38(?)	Common sources of bias in surveys and experiments.
#38(?)	How to determine the population to which the results of statistical inference can be extended, if any, based on how the data were collected.
#22	How to determine when a cause and effect inference can be drawn from an association, based on how the data were collected (e.g., the design of the study)
---	That words such as “normal”, “random” and “correlation” have specific meanings in statistics that may differ from common usage.
	Students should understand the parts of the process through which statistics works to answer questions, namely:
#37; #38(?)	How to obtain or generate data.
---	How to graph the data as a first step in analyzing data, and how to know when that’s enough to answer the question of interest.
#6; #7; #9; #10; #15; #32; #33; #36	How to interpret numerical summaries and graphical displays of data - both to answer questions and to check conditions (in order to use statistical procedures correctly).
#19; #25; #26; #27; #28; #29; #30; #31	How to make appropriate use of statistical inference.
---	How to communicate the results of a statistical analysis.

CAOS-4 item(s)	GAISE learning outcome
	Students should understand the basic ideas of statistical inference:
#16; #34; #35	The concept of a sampling distribution and how it applies to making statistical inferences based on samples of data (including the idea of standard error)
#19; #25; #26; #27	The concept of statistical significance including significance levels and *p*-values.
#28; #29; #30; #31	The concept of confidence interval, including the interpretation of confidence level and margin of error.
	Finally, students should know:
#40(?)	How to interpret statistical results in context.
#39(?)	How to critique news stories and journal articles which include statistical information, including identifying what's missing in the presentation and the flaws in the studies or methods used to generate the information.
---	When to call for help from a statistician.

Note. --- indicates a learning outcome that has not been matched to a CAOS-4 item.
? indicates that an item is tangentially related to the outcome.

CONCLUSION

Six of the 22 learning outcomes have not been mapped to a CAOS-4 item and five others have been mapped to items that are only tangentially related to that outcome. Since writing assessment items requires a systematic, well thought-out approach to ensure "sufficient validity evidence to support the proposed inferences from test scores" (Downing, 2006, p. 3), this part of the project is currently ongoing. The presentation at ICOTS-8 will further describe the IRT analysis as well as the mapping and selection of CAOS-4 items to the GAISE learning outcomes. Furthermore, the process for writing items that align with the missing outcomes will be described. Initial data for evidence of validity and reliability will also be reported at the conference.

REFERENCES

American Statistical Association. (2005). GAISE college report. Online: www.amstat.org/education/gaise/GAISECollege.htm.

Barnard, J. J. (1999). Item analysis in test construction. In G. N. Masteres & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 195-206). Oxford: Elsevier.

delMas, R. C., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28-58.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gal, I., & Garfield, J. (Eds.) (1997). *The assessment challenge in statistics education*. Amsterdam: IOS Press and the International Statistical Institute.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 421-441). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational Measurement, 4th Ed.* (pp. 623-646). Westport, CT: Praeger Publishers.