TRENDS IN STUDENTS' CONCEPTUAL UNDERSTANDING OF STATISTICS

Robert delMas
University of Minnesota, USA
Delma001@umn.edu

The ARTIST project has been collecting data since 2005 on students' understanding of statistics through the administration of the Comprehensive Assessment of Outcomes in Statistics (CAOS) instrument. The CAOS test consists of 40 multiple-choice items that cover six topics: data collection and design, graphical representations, variability, sampling variability, tests of significance, and bivariate data. From 2005 through 2013, over 30,000 secondary and tertiary level students in the United States of America who were enrolled in a college-level first course in statistics completed the CAOS test at the end of their respective courses. A confirmatory factor analysis (CFA) was conducted to provide evidence for the dimensionality and reliability of the CAOS test. Students' responses were used to look at trends in students' understanding of the six statistical topics across the 8-year period.

BACKGROUND

The Comprehensive Assessment of Outcomes in Statistics (CAOS) test was developed to measure students' conceptual understanding of important statistical ideas at the end of an introductory course in statistics (delMas, Garfield, Ooms & Chance, 2007). The CAOS test consistes of 40 forced-choice items that are designed to measure students' statistical understanding in six conceptual areas: data collection and design, graphical representations, variability, sampling variability, tests of significance, and bivariate data. One of the primary purposes for the CAOS test was to support statistics education research, especially with respect to the effects of the statistics reform movement (see Cobb, 1992, 1993; Hogg, 1992). Information on the development of the CAOS test, as well as validity and reliability evidence are presented in delMas et al. (2007). The delMas et al. (2012) study was based on responses from 1470 students who completed the CAOS test during the first academic year that it was offered (2005-06). Since that time, 30,000 students who completed a college-level first course in statistics in the Unites States of America (USA) have taken the CAOS test. The purpose of the current study is to extend the study reported in 2007 by providing evidence that the CAOS test represents a single construct and to provide an estimate of the internal consistency of the CAOS test based on a much larger sample of respondents. In addition, performance trends on the CAOST test and sub-topics within the test are considered to see if there is evidence of an effect by the statistics reform movement on students' conceptual understanding.

METHODS

Respondents

The respondents consisted of secondary students enrolled in a college-level statistics course or college undergraduates enrolled in a first course in statistics who completed the CAOS test between 2005 and 2013. All of the secondary students were enrolled in an Advanced Placement (AP) Statistics course. Information on the AP program, courses and examinations in the USA can be found at https://apstudent.collegeboard.org/home. Each AP course is based on a standard college-level curriculum for the given subject area and is designed to provide secondary students experience with college-level coursework in the subject area. According to the AP website, more than 90 percent of colleges and universities in the USA offer college credit, advanced placement, or both, for qualifying AP Exam scores. Information on the AP Statistics course can be found at https://apstudent.collegeboard.org/apcourse/ap-statistics. Tertiary-level respondents were enrolled in a 2-year technical college, 2-year community college, 4-year college, or university within the USA.

To be included in the study, a student needed to complete the CAOS test during an in-class administration of the test, or if the test was administered outside of class time, the student must have completed the CAOS test in no less than 10 minutes and no more than 60 minutes. A total of

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2014 ISI/IASE]

23,645 respondents met the inclusion criteria. All respondents completed the CAOS during the final weeks of the statistics course in which they were enrolled.

Analyses

Each of the 40 items on the CAOS test was scored as 1 (a correct answer) or 0 (an incorrect answer). The CAOS test is hypothesized to measure a single construct of statistical understanding at the introductory level. A confirmatory factor analysis (CFA) was conducted to see if there was evidence that the CAOS test measures a single construct. Because of the dichotomous coding of responses, a robust method of estimation, mean-adjusted weighted least squares (WLSM), was used to fit the model (Kline, 2011). Both the observed and latent variables were standardized in the model, allowing a factor loading to be interpreted as the estimated correlation between an observed variable and the factor, and squared loadings as the proportion of variance in the observed variable explained by a factor (Kline, 2011). Items with near-zero factor loadings were identified, eliminated from the model, and the model was re-fit. A two-factor model was also estimated to see if it produced a better fit to the data than the one-factor model.

Once the best model was identified, factor scores were computed for all respondents by summing the products of the item responses by the respective item factor loadings and dividing the sum by the maximum possible factor score based on answering all items correctly. This produced a score with a possible range between 0 and 1. The distribution of the CAOS factor scores was explored using descriptive statistics. Sub-scores were also computed for each of the six conceptual areas. Descriptive statistics were used to explore trends across time for the overall CAOS factor score and each of the six conceptual areas.

RESULTS

Confirmatory Factor Analysis

A scree plot of eigenvalues indicated that a one-factor model was appropriate (see Figure 1). Confirmatory factor analysis (CFA) using structural equations modeling (SEM) was conducted using the `lavaan` package (Rosseel, 2012) in R. All of the CAOS items had positive, non-zero loadings on the single factor, with the exception of item 32 (see Table 1). The factor loading for item 32 is negative and close to zero. Item 32 was designed to measure students understanding of how sampling error is used to make an informal inference about a sample mean. One possible explanation for the low correlation between item 32 and the statistical understanding measured by the CAOS test is a lack of coverage of or practice with this type of informal inference in the introductory statistics courses taken by the respondents.

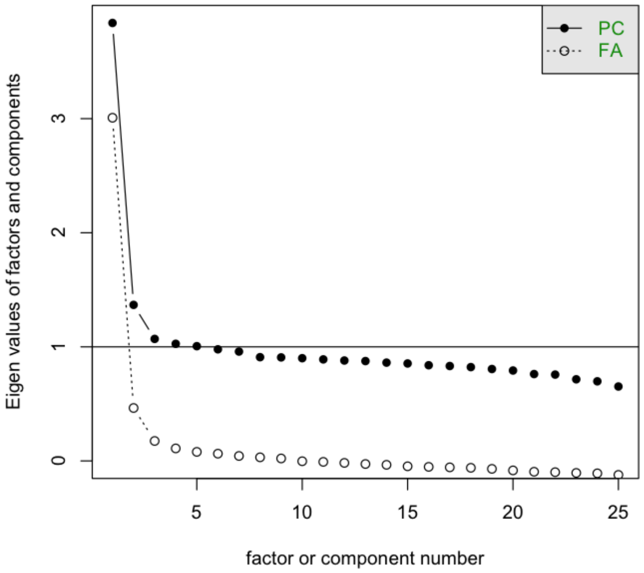

Figure 1. Scree plot of eigenvalues for factor analysis of CAOS items

Table 1. Factor loadings for one-factor CFA model of the 40 individual CAOS test items

Item	Loading	SE	P(>|Z|)
1	0.107	0.007	0.000
2	0.117	0.007	0.000
3	0.483	0.006	0.000
4	0.447	0.006	0.000
5	0.509	0.005	0.000
6	0.417	0.007	0.000
7	0.266	0.009	0.000
8	0.204	0.007	0.000
9	0.380	0.007	0.000
10	0.446	0.007	0.000
11	0.298	0.007	0.000
12	0.210	0.007	0.000
13	0.374	0.006	0.000
14	0.504	0.006	0.000
15	0.203	0.007	0.000
16	0.503	0.006	0.000
17	0.347	0.007	0.000
18	0.225	0.007	0.000
19	0.340	0.006	0.000
20	0.184	0.007	0.000

Item	Loading	SE	P(>|Z|)
21	0.168	0.007	0.000
22	0.285	0.006	0.000
23	0.194	0.007	0.000
24	0.140	0.007	0.000
25	0.259	0.006	0.000
26	0.243	0.007	0.000
27	0.304	0.006	0.000
28	0.313	0.007	0.000
29	0.293	0.006	0.000
30	0.046	0.007	0.000
31	0.175	0.007	0.000
32	-0.012	0.008	0.130
33	0.229	0.007	0.000
34	0.239	0.007	0.000
35	0.276	0.007	0.000
36	0.360	0.006	0.000
37	0.285	0.008	0.000
38	0.335	0.007	0.000
39	0.213	0.008	0.000
40	0.374	0.006	0.000

A second CFA model was fit to the CAOS items with item 32 excluded. All factor loadings were positive, non-zero and comparable to those of the first model presented in Table 1. The values of the fit indices (see Table 2) were comparable between the two CFA models. The 39-item model produced extremely large values for the model chi-square statistic, indicating that the model did not adequately reproduce the observed covariances. With respect to the other fit indices, recommended criteria that indicate a good fit (Browne & Cudeck, 1993; Hu & Bentler, 1999) are Comparative Fit Index (CFI) ≥ 0.9, Tucker-Lewis Index (TLI) ≥ 0.95, Standardize Root Mean Square Residual (SRMR) ≤ 0.08, and Root Mean Square Error of Approximation (RMSEA) ≤ 0.05. While the values for RMSEA and SRMR meet the criteria thresholds, the values of CFI and TLI were below the respective criteria thresholds.

Table 2. Confirmatory factor analysis model fit indices

MODEL	χ^2	CFI	TLI	RMSEA (conf. int.)	SRMR
40 items	28602.4	0.853	0.845	0.040 (0.039, 0.040)	0.035
39 items	28137.4	0.855	0.847	0.041 (0.040, 0.041)	0.036
One factor	4643.6	0.958	0.954	0.027 (0.026, 0.028)	0.023
Two factor	4636.5	0.958	0.954	0.027 (0.026, 0.028)	0.023

40 items: Model based on 40 individual CAOS items
39 items: Model based on all individual CAOS items except item 32
One factor: One-factor model based on testlets and individual items (24 variables)
Two factor: Two-factor model of Common and Uncommon topics
χ^2: Chi-square for test of model fit**SRMR**: Standardized Root Mean Square Residual
CFI: Comparative Fit Index **RMSEA**: Root Mean Square Error of Approximation
TLI: Tucker-Lewis Index

Kline (2011) suggests that correlation residuals greater than 0.10 may provide indications of model misfit. The 780 correlation residuals from the 39-item model were inspected to identify residuals with absolute values greater than 0.10. Fifteen item-pair correlation residuals were identified that met the criterion were identified (see Table 3). Thirteen of these extreme correlation residuals involved two items from the same testlet, with the other two extreme correlation residuals involving one item from a testlet. A testlet is a subset of items within a larger test that share the same context (Wainer, Sireci & Thissen, 1991). Correlation residuals for all item pairs consisting of two items from the same testlet were examined (see Table 3). Three of the nine additional correlation residuals were greater than 0.05.

Table 3. Correlation residuals from the 39-item CFA model for items in testlets

Item Pair	Residual	Item Pair	Residual	Item Pair	Residual	Item Pair	Residual
2, 1	0.090*	10, 9	0.186**	26, 19	0.135**	30, 28	-0.138**
4, 3	0.054*	12, 11	0.207**	23, 24	-0.062*	31, 28	-0.123**
5, 3	0.261**	13, 11	0.041	26, 25	0.081*	30, 29	0.164**
5, 4	0.179**	13, 12	0.317**	27, 25	0.258**	31, 29	0.163**
9, 8	0.033	15, 14	0.033	27, 26	-0.296**	31, 30	0.186**
10, 8	0.012	20, 11	0.106**	29, 28	-0.035	35, 34	0.254**

* Residual > 0.05
** Residual > 0.10

The results in Table 3 suggested that one source for the lack of model fit could be the local item dependency among items that shared the same context (Table 4 indicates the item groupings). A testlet score was computed for each set of items that shared the same context as the average of the item scores for all items in the set. This produced testlet scores between 0 and 1, with 0 indicating no items in a testlet were answered correctly and 1 indicating all items were answered correctly. The result was a set of 24 variables (9 testlet scores and 15 individual item scores).

A single-factor CFA based on the 24 variables was conducted. Factor loadings for the one-factor testlet model are presented in Table 4. All factor loadings are positive and non-zero. Fit indices (see Table 2) indicated a marked improvement in model fit compared to the 39-item model. While the model chi-square statistic was statistically significantly, the magnitude of difference between the 39-item and one-factor testlet model indicated a large improvement in model fit. This is also indicated by the noticeable reduction in both the RMSEA and SRMR measures. The CFI and TLI values for the one-factor testlet model were noticeably higher (both greater than 0.95), indicating a good fit between the model and the covariances. None of the correlation residuals had absolute values greater than 0.10.

Table 4. Factor loadings for one-factor testlet CFA model of the CAOS test

Items	Loading	SE	P(>|Z|)	Items	Loading	SE	P(>|Z|)
1, 2	0.150	0.007	0.000	21	0.163	0.007	0.000
3, 4, 5	0.547	0.005	0.000	22	0.289	0.007	0.000
6	0.429	0.007	0.000	23, 24	0.247	0.007	0.000
7	0.285	0.009	0.000	25, 26, 27	0.439	0.007	0.000
8, 9, 10	0.495	0.007	0.000	28, 29, 30, 31	0.381	0.007	0.000
11, 12, 13	0.391	0.006	0.000	33	0.234	0.007	0.000
14, 15	0.465	0.006	0.000	34, 35	0.297	0.007	0.000
16	0.517	0.006	0.000	36	0.364	0.006	0.000
17	0.357	0.007	0.000	37	0.302	0.008	0.000
18	0.221	0.007	0.000	38	0.344	0.007	0.000
19	0.346	0.006	0.000	39	0.225	0.008	0.000
20	0.178	0.007	0.000	40	0.163	0.006	0.000

Inspection of the squared loadings indicated that variables with factor loadings below 0.3 (i.e., squared correlations < 0.09) are associated with topics that may not be covered in all introductory statistics courses (e.g., item 2: ability to recognize two different graphical representations of the same data; item 18: recognizing a context where differences in variability is important) or items with low (item 20: matching a scatterplot with a verbal description of the relationship) or high (item 7: understanding the purpose of random assignment) difficulty. A two-factor model was fit where variables with loadings above 0.3 were fit to the first factor and all other variables were fit to the second factor. The fit statistics for the two factor model were similar to those for the one-factor model. In addition, the correlation between the two factors was very high (r = 0.975) and a statistical test did not indicate that the two-factor model fit the data better than the one-factor model ($\chi^2(1) = 0.095$, p = 0.758). The factor rho coefficient (Raykov, 2004), a measure of construct measurement reliability, was estimated at 0.75, indicating acceptable reliability for research purposes (Pedhazur & Schmelkin, 1991). Therefore, the one-factor model based on the 24 observed variables was retained.

CAOS Test and Sub-topic Scores

Figure 2 displays mean scores on the CAOS test and the six sub-topic areas over a nine academic year period. In general, scores have been stable over the nine-year period for the overall CAOS factor score and each of the sub-topic areas, with a possible increasing trend in mean scores for items assessing students' understanding of tests of significance. Students tended to score around 50% correct on the CAOS test across all years. Performance was higher on items assessing understanding of bivariate data and variability, with items assessing understanding of sampling variability and data collection showing the lowest mean performance.

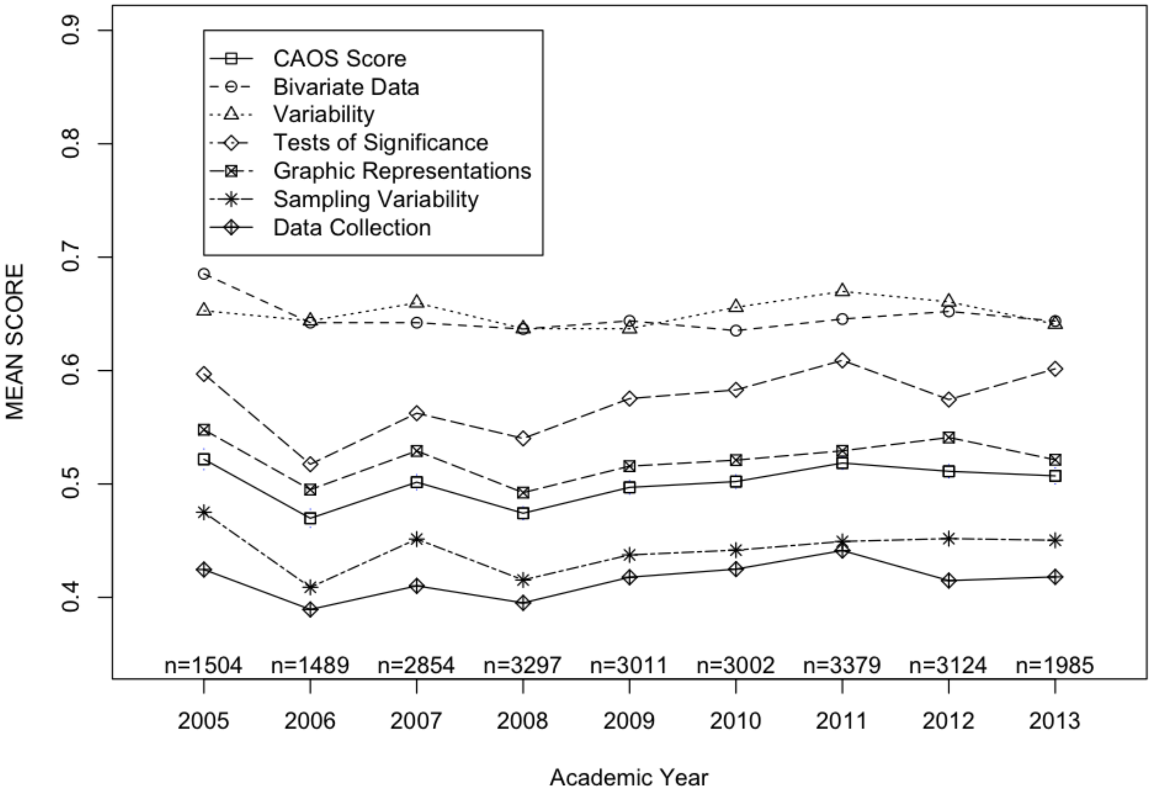

Figure 2. Mean scores on the CAOS test and sub-topics by academic year

DISCUSSION

Results from confirmatory factor analyses provides evidence that, after removing a single item (item 32), the CAOS test measures a single construct of statistical understanding of concepts covered in introductory statistics courses with sufficient internal measurement reliability for research purposes. Trends in mean scores on the CAOS test and six sub-topic areas assessed by the CAOS test show very stable mean test scores over a nine-year period. While the results do not indicate that students' statistical understanding has not increased over the study period, the sample was not collected specifically to assess the effects of the statistics reform movement. Therefore, there is still a need for studies that purposefully sample courses that differ in the degree to which they are reform based to see if performance is related to type of course. See Fry (2014), presented at this conference, for an example of a survey instrument that is being developed by researchers at the University of Minnesota to assess the degree of statistical reform in a course. The instrument, called the Statistics Teaching Inventory (STI), could be used in conjunction with assessments such as the CAOS test to study the effect of recommendations made by the statistics reform movement.

REFERENCES

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

Cobb, G. (1992). Teaching statistics. In *Heeding the Call for Change: Suggestions for Currricular Action, MAA Notes, Vol. 22*, 3-33.

Cobb, G. (1993). Reconsidering statistics education: A National Science Foundation conference. *Journal of Statistics Education 1*(1). http://www.amstat.org/publications/jse/v1n1/cobb.html

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28-58. http://iase-web.org/documents/SERJ /SERJ6(2)_delMas.pdf

Fry, E. B. (2014). *Introductory statistics instructors' practices and beliefs regarding technology and pedagogy*. Paper presented at the 9th International Conference on Teaching Statistics (ICOTS-9), Flagstaff, Arizona, USA.

Hogg, R. (1992). Report of workshop on statistics education. In *Heeding the Call for Change: Suggestions for Curricular Action, MAA Notes, Vol. 22*, 34-43.

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

Kline, R. B. (2011). *Principles and practice of structural equation modeling (3rd Edition).* New York: The Guilford Press.

Pedhazur, E. J., and Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy, 35*, 299-331.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36.

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal, 11*(1), 21-40.
http://iase-web.org/documents/SERJ/SERJ11(1)_Tintle.pdf

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*(3), 197-219.