DEVELOPMENT OF A TOOL TO ASSESS STUDENTS' CONCEPTUAL UNDERSTANDING IN INTRODUCTORY STATISTICS

Nathan Tintle[1] and Jill VanderStoep[2]
[1]Dordt College, Sioux Center, Iowa, USA
[2]Hope College, Holland, Michigan, USA
nathan.tintle@dordt.edu

Few tools exist to assess students' conceptual understanding in post-secondary, introductory statistics courses. The CAOS test is widely considered to be the gold standard, but was first published in 2007 and does not necessarily reflect some of the changes in student learning at the secondary level. Furthermore, it may not be sensitive enough to measure student conceptual understanding in modern post-secondary statistics courses (e.g., simulation-based inference). In this paper we will describe the process of developing a new instrument which uses some CAOS items, as well as additional new items to improve validity and reliability. We will share the validity and reliability results across n=3,833 students at 49 institutions, as well as information about external factors associated with student performance (e.g., test setting, question order).

INTRODUCTION

The algebra-based introductory statistics course continues to grow in popularity and remains a highly enrolled class around the world. For over a decade the Comprehensive Assessment of Outcomes in a first Statistics course (CAOS; delMas et al. 2007) has been the gold-standard valid and reliable instrument for evaluating student conceptual knowledge in the course (delMas et al. 2014). Recently, however, based on student performance data cited in a number of papers (e.g., Tintle et al. 2011), individuals have recognized an opportunity to modify the CAOS to reflect a number of observed trends in student performance and hopefully provide a more discriminating instrument to evaluate student performance. This has resulted in a handful of efforts to revise the CAOS test (e.g., GOALS (Sabbag and Zieffler 2015); Beckman et al. 2017, among others). In an effort to provide a more independent instrument adaptation and to best meet the timing and needs of a large-scale assessment project we initiated in 2012, we, too, initiated efforts to revise the CAOS instrument. Since 2012, we have been using this revised instrument in a large scale assessment project involving thousands of introductory statistics students in the United States. This manuscript presents validity and reliability results from this effort.

METHODS

Development of the instrument

The tool considered here is a modified version of the CAOS test (delMas et al. 2007), we note that similar modifications were also made resulting in the GOALS instrument (Sabbag and Zieffler 2015). A more detailed overview of the changes made with additional rationale and question verbiage can be found in Chance et al. (2016), but we briefly summarize these changes here. First, we eliminated a handful of questions we determined to be lacking discriminatory power based on either (a) excellent student pre-test performance (e.g., the students learned the concepts in high school) or (b) consistently poor performance both pre-course and post-course (potentially reflecting a concept that was poorly assessed by the question being asked). Second we added or modified the language for a handful of additional questions covering topics including sample size impacts and generalizability, and we included a few additional questions on p-value interpretation/inference. Items and field testing results on over 500 students were shared with an advisory board of statistics educators in 2012 before being implemented on a large scale beginning in 2013. Minor modifications in question wording were made in 2014 and 2015 to enhance readability. Results from the modified, 32-question, multiple choice concept inventory are presented in Chance et al. (2016), other papers being presented as part of ICOTS 10 (Chance et al. 2018, Roy et al. 2018, VanderStoep et al. 2018, among others), with additional papers anticipated in the future based on data collection from 2014–present.

Description of the sample

The primary sample that we will reference here was gathered during the 2016-2017 academic year from institutions primarily located in the United States. The dataset consists of n = 3,833 students, across 49 institutions including community colleges, comprehensive universities, four-year colleges, two year colleges and universities, and 190 separate instructor-sections. A variety of straightforward data cleaning rules were applied to the data to yield a clean and complete dataset of n = 3,833 students including removing students with large amounts of missing data, students who did not give permission to the use the data for research purposes, students who failed to complete either the pre-test or post-test, etc.

Administration of the instrument

The conceptual inventory was administered in conjunction with the Survey of Attitudes Towards Statistics (SATS-36; Schau 2003) and a brief demographic questionnaire. To test for potential order effects with the SATS, students were randomly assigned to receive either the SATS or the conceptual inventory first. All students in the sample took the instrument at the start of the semester (within the first week), and again at the end of the semester (during the last week of the course or during finals week). The vast majority of test administrations were unsupervised, outside of class using an internet accessible survey with a link specific to the instructor that was provided to the students. Most instructors provided a small grade incentive (e.g., homework points) for completion of the survey, but not for performance on the survey; instead students were typically encouraged by their instructors to 'try their best.'

RESULTS

Scale reliability

We started by reducing the 32-multiple choice questions to 24 unique questions, by collapsing highly-related questions resulting from the same question stem. The average score on the 24 pre-test items across all n = 3,833 students was 46.4% (SD = 11.3%) correct (Min = 10.4%, Q_1 = 38.5%, Median = 45.8%, Q_3 = 53.8%, Max = 88.5%), while the average score on the post-test was 54.1% (SD = 13.5%) correct (Min = 14.6%, Q_1 = 44.4%, Median = 53.5%, Q_3 = 63.2%, Max = 97.2%). The overall reliability of the scale (Cronbach's alpha = 0.65 at post-course) was reasonable for a scale comprised of dichotomous items and, notably, was higher at post-course than pre-course. We sequentially dropped each of the 24 items in the scale and recomputed alpha. In only two cases did the alpha increase after dropping the item with the alpha increasing to 0.648 and 0.652, respectively, from 0.647, for a single question in the significance scale and a single question on the data collection scale. Item-total correlations ranged from 0.09 to 0.40 pre-course and 0.15 to 0.47 post-course, with 19 of the 24 items' correlations strengthening from pre- to post-course.

Table 1. Summary statistics overall and by scale[1]

	Number of items	**Pre-course Mean (SD; α)**	**Post-course Mean (SD; α)**
Overall	24	46.4% (11.3%; 0.47)	54.1% (13.5%; 0.64)
Data Collection and Scope	4	52.0% (20.2%; -0.03)	58.4% (20.6%; 0.15)
Descriptive Statistics	5	49.5% (21.7%; 0.36)	54.9% (21.2%; 0.41)
Confidence Intervals	5	34.4% (19.5%; 0.20)	46.1% (22.4%; 0.33)
Significance	7	57.4% (19.5%; 0.34)	62.8% (21.0%; 0.42)
Simulation	3	30.6% (24.0%; 0.16)	34.7% (24.3%; 0.19)

[1]Pre- to post-course comparisons using paired t-tests were statistically significant overall and for each scale ($p<0.001$).

Subscale reliability

In line with prior work with analyses of this scale and the CAOS scale, we grouped items into five subscales of related items (see Table 1). In general, the alphas within each subscale were lower than overall, as would be expected due to the lower number of items contributing to each subscale. Notably, however, the alphas increased from the pre-course administration of the scale to the post-course administration of the scale for each of the five subscales.

Construct and predictive validity

Whereas the construct validity of the CAOS test on which this inventory is based has been demonstrated elsewhere, there are a few important observations worth making here. First, the reliability of the scale is stronger overall and within subscales at the end of the course compared to the beginning of the course. Second, stronger item-total correlations were also observed at the end of the course for the vast majority of items. Finally, student performance overall and on each subscale was better at the end of the course. These pieces of evidence point to some amount of learning occurring in the course as measured by the instrument.

Scores on the post-test showed good predictive validity by being moderately correlated with standardized math quantitative scores (ACT/SAT; $r = 0.39$, $p < 0.001$), college GPA before the course started ($r = 0.30$, $p < 0.001$) and most measured attitudes towards statistics at the end of the course by the Survey of Attitudes Towards Statistics (Value: $r = 0.27$ ($p < 0.001$); Difficulty: $r = 0.25$ ($p < 0.001$); Cognitive Competence: $r = 0.33$ ($p < 0.001$); Affect: $r = 0.28$ ($p < 0.001$); Interest: $r = 0.24$ ($p < 0.001$)). With only one of the six SATS subscales not showing association with post-test scores (Effort: $r = 0.002$ ($p = 0.51$)).

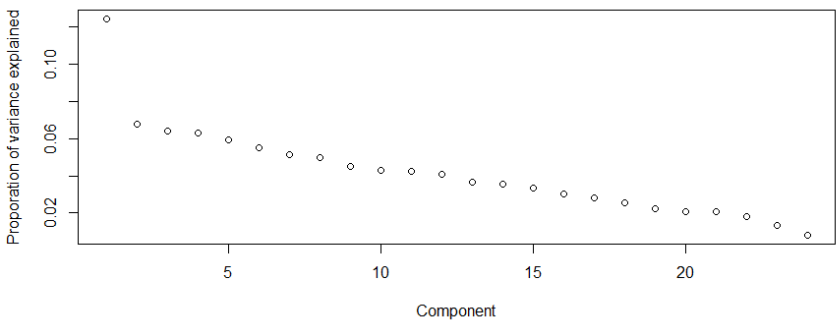
Order effects

The concept scale was part of a longer pre-course and post-course survey which included both demographic questions and the Survey of Attitudes Towards Statistics. In order to control for potential order effects of the conceptual inventory and the SATS, students were randomly assigned to either take the SATS first or the conceptual inventory first. On the pre-test there was no statistically significant impact of taking the concepts inventory before or after the SATS (0.004 estimated effect on overall score of taking concepts first; $p = 0.28$). There also was no evidence of an effect on the overall post-test score (-0.0001 estimated effect on overall score if take concepts inventory first; $p = 0.98$). No significant differences were observed on subscales on either the pre-course or post-course administration of the test (details not shown).

Administration environment

The majority of the students who took the exam did so out of class (82.4% pre-test; 74.8% post-test). As expected location of administration was associated with student performance, with scores on the pre-test and post-test both approximately 2.1 percentage points higher when taking the test in-class vs. an out-of-class administration ($p < 0.001$ for both pre- and post-course). Impact on pre-test subscale scores was similar with four of the five scales showing better scores (1.9 to 4.7 percentage points higher; $p < 0.05$ in all four cases), with the lone exception the confidence intervals scale (0.008 improvement on pre-test for in-class administration; $p = 0.48$). A similar pattern was observed on the post-test with improvement ranging from 2.2 to 4.1 percentage points for four of the five subscales, ($p < 0.05$), with the lone exception being the descriptive statistics subscale (0.009 improvement; $p = 0.41$).

Figure 1. Scree plot after principal components analysis on the twenty-four item scale:



Factor structure

Principal components analysis (PCA) was run on the post-test data, suggesting a single component model was the best explanation for the observed data (see scree plot of post-data in

Figure 1). The first component explained 12.4% of the total variation, with the next component explaining only 6.8% of total variation. After varimax rotation, 18 of the 24 items loaded positively on the first component. When PCA was run on the pre-test data a single model still appeared best, though the proportion of variation explained by the first component was less (10.0%) and only 14 of the items loaded positively on the first component.

CONCLUSIONS

The modified concept inventory presented here shows strong reliability and validity, and should be considered ready for widespread use in assessing student performance in introductory statistics courses. The single factor structure suggested by the modified concept inventory suggests that the instrument is likely evaluating overall student abilities in an introductory statistics course. Subscales show lower overall reliability, but may be useful in focusing in on particular areas of student improvement when evaluating curricular, pedagogical, instructor, or student effects. It was promising to note that order used when jointly administering the conceptual inventory with the Survey of Attitudes Towards Statistics showed little impact on performance on the conceptual inventory. Expectedly, the environment in which the inventory was administered did have modest impacts on student performance and should be accounted for when interpreting results. Future work is needed to further evaluate the validity and reliability of the instrument in more diverse samples (e.g., beyond just the USA). Consideration of whether a short-version of the scale (we have begun pilot-testing) would work as effectively as the 32-question inventory used here, and the impact of incentives for student participation/performance on the inventory are both needed areas of future research.

ACKNOWLEDGEMENTS

We acknowledge funding support from grant NSF-DUE-1323210 and the tireless efforts of Beth Chance in designing, refining, and interpreting the instrument, Cindy Nederhoff in coordinating assessment at dozens of institutions, and numerous undergraduate assistants in assembling the dataset, data cleaning, preparing instructor reports, and initial data analyses.

REFERENCES

Beckman, M., delMas, B., & Garfield, J. (2017). Cognitive transfer outcomes for a simulation-based introductory statistics curriculum. *Statistics Educ Research Journal*. 16(2): 419-440.

Chance, B., Wong, J., & Tintle, N.L. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Educ.* 24(3): 114-126.

Chance, B., Mendoza, S., & Tintle, N.L. (2018). Student gains in conceptual understanding in introductory statistics with and without a curriculum focused on simulation-based inference. *Proceedings of the International Conference on Teaching Statistics*, 10. Kyoto, Japan, 2018.

delMas, R. (2014). Trends in students' conceptual understanding of statistics. Proceedings of the Ninth International Conference on Teaching Statistics. http://iase-web.org/icots/9/proceedings

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

Roy, S. & Mcdonnel, T. (2018). Assessing simulation-based inference in secondary schools, *Proceedings of the International Conference on Teaching Statistics*, 10. Kyoto, Japan, 2018.

Sabbag, A. G., & Zieffler, A. (2015). Assessing learning outcomes: An analysis of the GOALS-2 instrument. *Statistics Educ Research Journal*, 14(2): 93–116.

Schau, C. (2003). *Survey of Attitudes Toward Statistics* (http://evaluationandstatistics.com/)

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).

VanderStoep, J.L., Couch, O., Lenderink, C. (2018). Assessing the association between quantitative maturity and student performance in an introductory statistics class: simulation-based vs. non simulation based. *Proceedings of the International Conference on Teaching Statistics*, 10. Kyoto, Japan, 2018.