ON THE UTILIZATION OF THE COMPREHENSIVE ASSESSMENT OF OUTCOMES IN STATISTICS (CAOS)

BACKGROUND

CAOS developmental goals

(1) An assessment with questions understandable by all students finishing any introductory statistics course, whose scores are reliable

(2) Identify areas where students do or do not improve in their understanding and statistical reasoning

Initial validity evidence

- Experts agree that CAOS is well suited for students completing the *consensus curriculum* (Cobb, 2007)
- Internal consistency = 0.82 (delMas et al., 2007)
- CFA confirms unidimensionality (delMas, 2014)

Research questions

(1) How has CAOS been used in teaching and research?

(2) To what extent have CAOS users established validity evidence for their specific interpretations and uses?

Use in teaching Fall '07 – Spring '18

38,519 students, 169 instructors, 116 institutions

- 45% of students took CAOS as extra credit
- 25% of students took CAOS as an exam
- 10% of students took CAOS as exam review

Use in research 6 most common uses:

- Base for developing ad hoc items
- Measure association with other constructs
- Evaluate unique student populations
- Compare different classroom formats
- Compare different statistics curricula
- Base for developing new assessments

Unforeseen uses

- Two studies calculated and analyzed CAOS subscores defined by topics (e.g., sampling variability). Standard 1.14 of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) states that subscores should be distinct and reliable.
- Nine studies used CAOS to compare unique student populations to US national baseline results. No explicit evaluation of differential item functioning (DIF) was conducted on initial tests.

"The job of validation is not to support an interpretation, but to find out what might be wrong with it."
(Cronbach, 1980, p. 103)

Follow-up Study

Research Question:

- What might be wrong with calculating subscores for CAOS by topic?
- What might be wrong with using CAOS to measure understanding in diverse student populations?

Methods:

- 6-topic bifactor model to assess distinctiveness
- Unidimensional testlet model for data collected between fall '17 and spring '18 to test for DIF

Researchers using CAOS should ensure their intended use is supported by a preponderance of validity evidence.

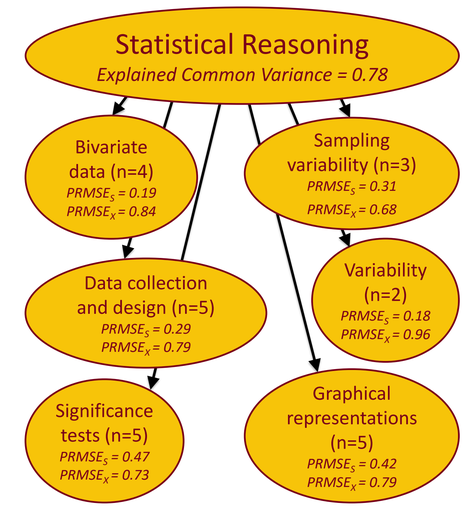
V.N. Vimal Rao and Carlos Chávez

COLLEGE OF EDUCATION
+ HUMAN DEVELOPMENT

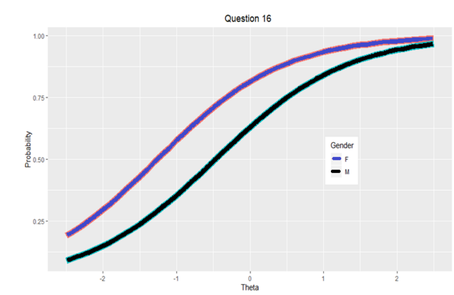
Department of Educational Psychology

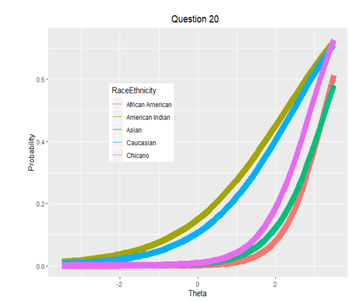
UNIVERSITY OF MINNESOTA
Driven to Discover®

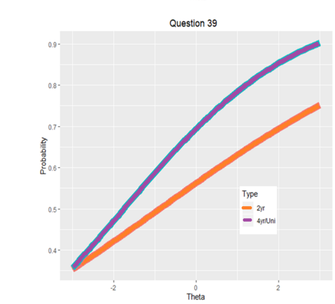
SUBSCORES



DIF







REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bargagliotti, A., Anderson, C., Casey, S., Everson, M., Franklin, C., Gould, R., ... & Watkins, A. (2014, July). Project-SET materials for the teaching and learning of sampling variability and regression. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics, Flagstaff, Arizona, USA*. Voorburg, The Netherlands: International Statistical Institute.

Cobb, G. W. (2007). The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1).

Cronbach, L. J. (1980). Validity on parole: How can we go straight. *New directions for testing and measurement, 5*(1), 99-108.

delMas, R. (2014). Trends in students' conceptual understanding of statistics. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics, Flagstaff, Arizona, USA*. Voorburg, The Netherlands: International Statistical Institute.

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.

Fabrizio, M., López, M. V., & Plencovich, M. C. (2011). Statistics in teacher training colleges in Buenos Aires, Argentina: Assessment and challenges. *Proceedings of the 56th Session of the International Statistics Institute. Lisbon: Portugal.*

Feinberg, R. A., & Jurich, D. P. (2017). Guidelines for interpreting and reporting subscores. *Educational Measurement: Issues and Practice*, *36*(1), 5-13.

Hannigan, A., Gill, O., & Leavy, A. M. (2013). An investigation of prospective secondary mathematics teachers' conceptual knowledge of and attitudes towards statistics. *Journal of Mathematics Teacher Education, 16*(6), 427-449.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate behavioral research, 47*(5), 667-696.

Tintle, N.L., Rogers, A., Chance, B., Cobb, G., Rossman, A. (2014). Quantitative evidence for the use of simulation and randomization in the introductory statistics course. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics, Flagstaff, Arizona, USA*. Voorburg, The Netherlands: International Statistical Institute.