Development and Validation of a Research-based Assessment: Reasoning about *P*-values and Statistical Significance

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Sharon Jacqueline Lane-Getaz

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Department of Educational Psychology: Quantitative Methods in Education
Statistics Education

Advisers
Joan B. Garfield, Ph.D.
Robert C. delMas, Ph.D.

June 2007

ABSTRACT

This study developed the Reasoning about *P*-values and Statistical Significance (RPASS) scale and provided content- and some construct-related validity evidence. The RPASS scale was designed to assess conceptual understanding and misunderstanding of *P*-values and statistical significance and to facilitate future research about the effects of instructional approaches on this understanding.

During Phase I, a test blueprint was developed based on difficulties identified in the literature. RPASS-1 was piloted across four courses at the University of Minnesota, assessing five correct conceptions and 12 misconceptions ($N = 333$). In Phase II, learning goals were added to the blueprint from the ongoing literature review. Incorporating modifications from the blueprint, the pilot, and suggestions from five statistics education advisors produced RPASS-2.

During Phase III, RPASS-2 was administered. Feedback from two field tests and 13 student interviews ($n = 61$) produced a 25-item RPASS-3A. Ten experts from four colleges and universities rated RPASS-3A content and made modification suggestions. After individual meetings to review an interim RPASS-3B, all ten experts *agreed* or *strongly agreed* that the two subscales (correct conceptions and misconceptions) assessed the stated learning objectives or misconceptions. Deleting one redundant item produced RPASS-4.

In Phase IV, RPASS-4 was administered to students across five introductory courses at California Polytechnic State University, assessing 13 correct conceptions and 14 misconceptions ($N = 224$). On average, respondents answered 16 items correctly.

Results showed a higher mean proportion of correct responses for correct conception items versus misconception items. Statistical literacy items were the least difficult, and statistical thinking items were the most difficult.

RPASS-4 total score reliability was low (Cronbach's coefficient $\alpha = .42$, $N = 224$). Convergent and discriminant measurements were gathered in two courses to provide some evidence of construct-related validity ($n = 56$). Correcting validity coefficients for attenuation, RPASS-4 correlated moderately with the convergent and weakly with the discriminant measure.

In Phase V, a subsequent item analysis identified a 15-item subset of RPASS-4 items (designated RPASS-5) with estimated internal consistency reliability of $\alpha = .66$. RPASS-5 retained weak evidence of construct validity as obtained for RPASS-4. Inferences about respondents' understandings and misunderstandings were drawn from these 15 items.