Statistically Adrift: Why A Central Conclusion in *Academically Adrift* is Faulty

David M. Lane and Fred L. Oswald
Rice University

One of the most cited findings reported in the book *Academically Adrift: Limited Learning on College Campuses* by Arum and Roska (2011a) reflects the main theme on the cover of the book, namely that: "...we observe no statistically significant gains in critical thinking, complex reasoning, and writing skills for at least 45 percent of the [college] students in our study" (p. 36). Similarly, Arum and Roska (2011b) state "Forty-five percent of students did not demonstrate any significant improvement in learning, as measured by CLA performance, during their first 2 years of college." (p. 204).

This gain refers to change in a measure called the CLA, which the authors administered to college students both at the beginning of their freshman year and again at end of their sophomore year, with the difference between CLA scores across years as the indicator of change, where a positive and significant change was determined to be an indicator of learning. Astin (2011) predicted that "this 45-percent conclusion is well on its way to becoming part of the folklore about American higher education," and the year since his publication has borne this out. A Google search for the key words:

"Academically Adrift" Arum "45 percent of the students"

conducted on August 6th, 2012, revealed 13,100 results. As a basis for comparison, a search using the key words:

"Academically Adrift" Arum "47 percent of the students"

revealed no results.

Astin (2011) provided an astute critique of Arum and Roska's use of significance testing to measure improvement in student learning. His two major arguments were (1) learning should be measured in terms of practically significant gains; a learning standard based on statistical significance is vaguely defined at best, and (2) the reliability of measured student gains was low, making it very difficult to find significant differences, even if important gains in learning actually existed. Initially, we felt that Astin dealt with these issues sufficiently, and that no further comments were necessary.

Upon closer examination, however, we discovered that even the specific nature of the significance tests by Arum and Roska is problematic and merits a comment in order to correct the scientific record. These authors – and subsequently the media – were shocked at their finding that 45% of students failed to show statistically significant gains, because this was interpreted to reflect poorly on college teaching. This 45% finding is, indeed, shocking – but for a completely different reason. Considering that each significance test was based on a sample size of 1 (i.e., each student's change in

the CLA measure)[1], it is hard to imagine that as many as 55% of students would show statistically significant gains. Indeed, one would expect to find an order of magnitude fewer significant improvements, based on the mean difference between the pre- and post-tests the authors reported in their study.

The reason Arum and Roska found that so many (not so few) students improved significantly is that they computed the wrong significance test. These authors computed a margin of error appropriate to test to the mean difference *across* students but then they misapplied that margin of error in testing the improvement of *individual* students. More specifically, their key error was to include the sample size in calculating the margin of error, leading to a margin of error that was much lower (i.e., making the sample estimate appear more accurate) than the margin of error appropriate for each individual student. Clearly, the number of students in the study should not affect the significance of tests for the improvement scores of individual students, but with Arum and Roska's analysis it does. For example, if 100,000 students had been tested, virtually every gain (or loss) would have been significant using Arum and Roska's approach, and the authors would have concluded that a much higher percentage of students showed learning gains – even though nothing changed except the number of students that were tested. Using the appropriate method, an increase in sample size would not change the significance test for each student's change in his/her CLA score across two time points.

Thus, we agree with Astin that the proportion of students improving significantly is hardly one of the better ways to measure the size of an effect. However, if this measure is to be used, it should at least be calculated correctly, and we wanted to explain the correct statistical test.

Our analysis leads us to question the generality of the concept of critical thinking. Despite the controversy surrounding their work, Arum and Roska's (2011) book shows clearly that these authors are highly capable of critical thinking in many contexts. However, the context of significance testing may not be one of them. In short, critical thinking may be context dependent.

References

Arum, R., & Roska, J. (2011a). *Academically adrift: Limited learning on college campuses.* Chicago: University of Chicago Press.

Arum, R., & Roska, J. (2011b). Limited learning on college campuses, *Society*, *48*, 203-207.

Astin, A. W. (2011, February). "In 'Academically Adrift,' data don't back up sweeping claim." *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Academically-Adrift-a/126371/.

[1] The fact that the "error term" was based on the whole sample does not materially affect this argument.