Online Supplementary Material 1: Detailed Information on Data and Methods

Data

This study involves data collected from two cohorts of undergraduates and their research advisors who participated in the SOEST Scholars Program in 2016–2017 and 2017–2018. A total of 36 undergraduates participated in the SOEST Scholars program over these two years, and 31 completed student surveys (response rate of 86%). Of the 31 students who completed surveys, all but one of their advisors completed the advisor survey (response rate of 97%). Thus, our data set comprises paired survey responses from 30 of 36 student-advisor pairs (response rate of 83%).

Of these 30 students, 16 (53%) identified as women; the remaining 14 (47%) identified as men. No students reported other genders. They represent various ethnicities, including 14 Native Hawaiian, 1 Pacific Islander, 2 African-American, 5 Asian, 3 Filipino, and 5 Caucasian. Thus, 15 (50%) of the students are Native Hawaiians and/or Pacific Islander (NHPI), with the other 15 representing a range of non-indigenous identities.

Study Design

Two sets of survey items (which we term "Absolute" and "Growth") are identical on student and advisor surveys (see Online Supplementary Material 2). In the Absolute set, students and advisors evaluate the students' skills and performances in 10 areas (e.g., amount of work accomplished, quality of work performed) at the end of the research experience along a five-point Likert scale, ranging from Unsatisfactory to Excellent. In the Growth set, students and advisors evaluate the extent to which the students changed or grew over the course of the undergraduate research experience in nine areas (e.g., works more independently, takes more initiative to problem-solve) along a five-point Likert scale, ranging from Strongly Disagree to Strongly Agree.

The student and advisor surveys were approved as exempt in 2017 by the UH Institutional Review Board (Protocol # 2017-00612). Prior to 2017, survey items were pre-tested (informally validated) with previous years' cohorts of students and advisors and revised accordingly.

All surveys were administered online using Survey Monkey software. Each student was emailed the uniform resource locator (URL) of the student survey along with a unique student code. Similarly, each advisor was emailed the URL for the advisor survey and a unique advisor code, as well as their student's code. The unique student codes were used to pair the student and advisor surveys. No identifying information was collected, and students and advisors were asked to refrain from including names or other identifying information in their open-ended comments. Two reminders were sent to non-respondents, spaced about a month apart.

Our null hypothesis is that there is no statistically significant difference between the student vs. advisor assessments of students' skills and performance, as measured by Absolute and Growth survey items. We test this hypothesis in two ways: (1) comparing the student vs. advisor responses to each individual survey item, and (2) comparing the student vs. advisor responses to each dataset (Absolute and Growth) as a whole. For the former analysis, we perform a paired, two-tailed t-test. For the latter, we apply a non-parametric permutation test.

T-test on Individual Survey Items

As t-tests can only be performed on quantitative data, our first step was to convert the qualitative Likert responses to integers. This conversion is shown in Table S1. Limitations of this approach are explored in the Discussion section below.

Table S1. Quantification of Likert responses to Absolute and Growth survey items on a scale of 1 to 5.

	Likert Scale Responses				
Absolute Survey Items	Unsatisfactory	Fair	Satisfactory	Very Good	Excellent
Growth Survey Items	Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree
Quantified as:	1	2	3	4	5

Following quantification of Likert scale responses, for each survey item, we computed the student (S) and advisor (A) mean, the standard error of these means, and the mean difference, D=A-S. As all items were stated positively (none had reverse wording), a positive D indicates that the advisors, on average, assessed the students more highly than the students assessed themselves. To test the statistical significance of any non-zero D, we applied a paired, two-tailed t-test to each survey item. Ideally, for more robust results, we would apply a t-test to the entire Absolute and Growth data sets of student vs. advisor responses, rather than to each individual survey item. However, such an analysis is precluded by the t-test assumption that each sample be independent.

Permutation Test Analysis on Combined Survey Items

Unlike t-tests that calculate a theoretical probability value based on an assumed distribution, permutation tests (also called exact or randomization tests) calculate p-values empirically by resampling the data numerous times. A comparison between the actual survey responses vs.

randomized data iterations reveals the probability of any observed student-advisor differences being significant (that is, non-random).

Permutation tests make no assumptions regarding probability distributions, so they can be applied to any data set. Thus, we invoked a permutation test to compare the mean advisor vs. student response to the 10 Absolute survey items combined (300 student-advisor pairs), as well as to the nine Growth survey items combined (269 student-advisor pairs, due to one Growth item not being answered by one respondent). The coding, which was done in Matlab, is described below for the Absolute data set. The coding for the Growth data set is identical except for the number of pairs (269 vs. 300)

First, we calculated the mean advisor-student difference for each pair of Absolute responses using the quantification shown in Table S1. For example, if a student rated herself or himself on a given survey item as Satisfactory (3) while the advisor rated that student as Excellent (5) on that same survey item, then the difference would be recorded as +2. If the student's and advisor's ratings were reversed, the difference would be –2. This was done for each of the 300 student-advisor pairs of responses to the 10 Absolute survey items, and the results were averaged to get a mean difference, D.

Then, we performed a permutation test. Instead of doing a normal randomization (which would have randomized all data), we preserved the pairings of the student-advisor data. We randomly swapped the student vs. advisor responses within each pair, calculated the mean difference and iterated 100,000 times. This yielded a normal frequency distribution of mean differences. The fraction of results that were tailward of the observed mean difference yielded an empirical p-value. For example, if the observed mean difference is 0.10, then p is calculated as the number of mean differences less than -0.10 or greater than 0.10, divided by 100,000.

Demographic Analyses

We then performed the permutation test again, on both the Absolute and Growth data sets through a demographic lens. We examined differences in student-advisor ratings by gender, ethnicity, and the intersectionality of these identities. This was motivated by previous studies that found women and certain minority groups—and particularly students at the intersection of those identities—often report lower self-efficacy.

For gender, we compared men and women, as none of the students reported a non-binary gender. For ethnicity, we compared Native Hawaiians and Pacific Islanders (NHPI) to non-indigenous students (non-NHPI); this choice was determined by the data set rather than a priori, as 50% of our students were NHPI. No other ethnic or racial category had more than five students (17%), precluding meaningful analysis. Limitations of this approach are explored in the Discussion section in the main article. For the intersectionality analysis, we compared four categories: NHPI women, NHPI men, non-NHPI women, and non-NHPI men.