Self-efficacy changes and gender effects on self-efficacy in a large-scale robotic telescope focused curriculum

Rachel Freed,[1] David McKinnon,[6,*] Saeed Salimpour,[2,3] Michael Fitzgerald,[4] Dan Reichart,[5] and Christina Norris[1]

[1]*Edith Cowan University, Joondalup, Australia*
[2]*International Astronomical Union, Office of Astronomy for Education, Heidelberg, Germany*
[3]*Max Planck Institute for Astronomy, Heidelberg, Germany*
[4]*Las Cumbres Observatory, Goleta, California, USA*
[5]*Univeristy of North Carolina, Chapel Hill, North Carolina, USA*
[6]*Charles Sturt University, Bathurst, NSW, Australia*

(Received 19 December 2023; accepted 18 March 2024; published 8 May 2024)

In this paper, we present the results of an investigation into the effects of engaging with robotic telescopes during an Astronomy 101 (Astro101) course in the United States and Canada on the self-efficacy of students. Using an astronomy self-efficacy survey that measures both astronomy personal self-efficacy and instrumental self-efficacy, the authors probed their covariance with the respondents' experience of an Astro101 course that uses robotic telescopes to collect astronomical data. Strong effects on both self-efficacy scales were seen over the period of a semester utilizing a scalable educational design using robotic telescopes. After participation in the course, the results show that the gender gap in self-efficacy between self-identified men and women is largely reduced to statistically insignificant differences compared to the initial large significant difference.

DOI: 10.1103/PhysRevPhysEducRes.20.010137

I. INTRODUCTION

A. Astro101 and robotic telescopes

First-year nonscience-major astronomy courses at undergraduate institutions in the United States, herein referred to as "Astro101," have a comparatively long history as a focus for broad scientific literacy in the populace [1–3]. Estimates in the past have shown that in any given year, roughly a quarter of a million students take such courses as part of the general requirements of their degree [4,5]. These courses are available at local community colleges all the way up to ivy league R1 research institutions. The majority of students taking these courses are not majors in any of the science, technology, engineering, and mathematics (STEM) fields and are spread across the humanities, business, law, medicine, and various other fields of study [6]. Not only are these subjects seen as providing an arena for broadening scientific literacy but also potentially influencing students' making career decisions more closely aligned to the STEM fields, hence influencing the "leaky pipeline" [7,8] issue in a positive manner.

Similar to Astro101, a *course-focused* change agent in educational programs, robotic telescopes are an *instrument-focused* change agent for scientific literacy and career clarification. Robotic telescopes have a long history of being extolled as a potential game-changer for science education [9–11], providing direct access, via the Internet, for students to collect their own novel data affording them a much more powerful sense of ownership over their own learning. However, robotic telescopes are "just" an instrument. Placing a robotic telescope in a field and simply waiting for students or instructors to use it will not change science education. Bain and Weston [12] articulated the idea that there is an implicit "belief that teachers and students with access to and mastery of technology would transform education." Bain and Weston [12] see this as a failure of the education system and not of technology. They argue that education research needs to inform the manifestation of technology in education settings, rather than education settings adopting every new technology. In the context of schools, there is often a quick rush toward certain technologies without the pedagogical considerations or evidence that they have a positive impact on student outcomes. One example of this is the various manifestations of 1:1 laptop initiative in schools, which has had its fair share of criticism and support [13]. The key here is that any new technology or rather any new pedagogical intervention requires the appropriate support for teachers to allow them to use it in their classroom. Therefore, robotic telescopes have to be embedded appropriately in a well

*Present address: Charles Sturt University Panorama Avenue, Bathurst 2795, Australia.

thought-through educational design and setting, supported by robust pedagogical theory and research in order to have any chance of achieving such a goal [14].

Within the context of instrument-focused Astro101 courses, the University of North Carolina (UNC) at Chapel Hill has been developing a unique astronomy curriculum—"Our Place In Space!" (OPIS), primarily for undergraduate students, for the past 13 years. The goal of this curriculum is to significantly boost STEM enrolments on a national scale as well as boost students' technical and research skills. This curriculum leverages "Skynet"—a global network of about two dozen professional-grade, robotic telescopes that we have deployed across four continents and five countries. The provision of OPIS! to more than two dozen institutions nationally has allowed more than 3500 students to have authentic astronomy experiences in courses that have access to these telescopes. The broad reach of this program allows for a deep investigation into the development of self-efficacy and change in motivation to persist with science due to using astronomy instrumentation in an authentic way. This then provides the basis for the development of survey instruments in which we situate this paper.

OPIS! [15] is a sequence of eight laboratory activities (labs) in which students use the same research instrumentation as professionals to collect their own data. They then use this self-collected data (astronomical images and spectra) to reproduce some of the greatest astronomical discoveries of the past 400 years, such as measuring the orbits of the Galilean moons around Jupiter and the use of Cepheid variable stars to measure distances to objects within the Milky Way galaxy. In addition, they gain technical and research skills at the same time. Although students are not carrying out cutting-edge research, they are using cutting-edge research instrumentation. In addition, they are collecting and analyzing their own data and working collaboratively with peers. Consequently, there is great overlap with the course-based undergraduate research experience (CURE) pathway model [16], where these labs including observational experiences are specifically designed to pair with standard introductory astronomy curricula. Thereby the design of the courses allows for the facilitation of widespread adoption.

OPIS! is built around the cosmic distance ladder, the method by which astronomers successively measure the distance to more and more distant objects in the universe using the previous step as the calibration for the next step and which serves as an organizing principle in most introductory astronomy courses or sequences, and as such, it reinforces students' classroom experiences. The goal of OPIS! is to move beyond laboratory experiences in which students learn how to use a telescope for its own sake, instead of using these instruments to enhance the learning of science concepts. This is in line with other research on the implementation of new technologies such as expressed by Saubern *et al.* [14].

B. Self-efficacy, motivation, and the pathway to change

Influencing the development of scientific literacy and student career decisions is not a simple one-step process. As hypothesized by Wooten *et al.* [16], there are numerous "pathways" through which this can be developed. These include increased technical skills or increased content knowledge through increased self-efficacy or increased motivation to the endpoint of enhanced science identity and career clarification. For our purposes, we simplify this path model to that presented in Fig. 1.

It can readily be seen that self-efficacy is a key hypothesized gatekeeper toward the intended long-term outcomes of these courses alongside motivation [16,17]. Self-efficacy is influenced by four sources: mastery experiences [18,19], vicarious experiences [20], social persuasion from important or similar others [21,22], and from physiological and affective states [18]. Because self-efficacy is domain and even task specific [23], it needs to be studied within each science field. It has only recently been studied within the domain of undergraduate astronomy courses [24–26]. Bailey *et al.* [24] studied the interrelationship between self-efficacy, interest, and knowledge gains relating to star properties in Astro101 classes. For the courses in which there were self-efficacy increases, interviews with instructors indicated there was a higher level of scaffolding and more purposefully planned opportunities for students to experience mastery of course tasks. Hewitt *et al.* [25] found increases in research self-efficacy for students in a course-based undergraduate research experience in an online astronomy class. Additionally, in their latent profile analysis, Galano *et al.* [27], using a modified version of our self-efficacy instrument [28], found a significant positive relationship between attitudes toward astronomy and astronomy self-efficacy.

As a quantitative statistical tool to measure such changes in self-efficacy had not been developed, we created an instrument to do so with the initial exploratory factor analysis presented in an earlier published paper [28].

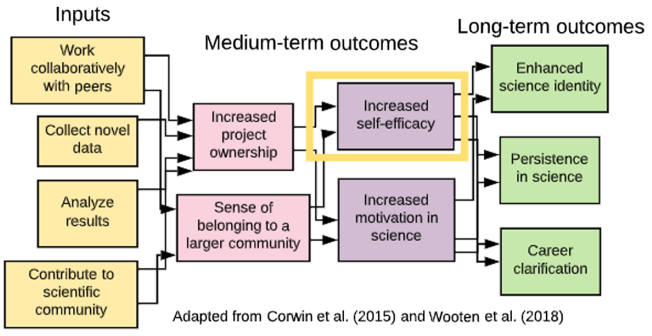

FIG. 1. Diagram of how research experience inputs influence medium- and long-term outcomes for students. Increased self-efficacy leads to enhanced science identity, persistence in science, and increased STEM career choices.

The development of this instrument led to further querying a hypothesized link between the variety of Astro101 inputs and short-term outcomes (e.g., work collaboratively or increased project ownership) and long-term outcomes (e.g., enhanced science identity and persistence in science). However, currently, there is no evidence of a robust way to measure this at scale. It is the development of this tool that we use in this paper to attempt a longitudinal investigation into the covarying effects on the self-efficacy of students through engaging with robotic telescopes in the same Astronomy 101 course adopted by tertiary institutions in the United States and Canada.

C. Gender differences in science self-efficacy

Research into gender differences in STEM participation goes as far back as the early 1970s [29], and to date, there is an extensive body of research that continues to explore various constructs of self-efficacy, ability, self-esteem, perceptions, and gender in the context of STEM from psychological, sociological, and educational perspectives [30–34]. Although studies such as the meta-analysis conducted by Huang [35] reported a difference in self-efficacy between women (art, language) and men (mathematics, computer, and social sciences), there is much more to this than a simple demarcation between disciplines.

There is a distinct lack of women in astronomy, both professionally [36,37] and in the large amateur astronomy community in the United States [38], although there seems to be a potential increase in the field of astronomy education research [39]. There are many research papers documenting that women have lower self-efficacy in STEM fields than men at all levels of their educational and career trajectories [40–44]. Women in science face not only a masculine-oriented environment [45,46] but also stereotype threat [47] and microaggressions [48]. A 2007 National Academies review of the literature found that women are lost to science and engineering careers at every educational transition [49]. More recently, White [50] has shown that there continue to be gender disparities in all aspects of physics education and careers with women earning fewer degrees, having less support in graduate school, and earning less money in their careers. Studies by Nissen [51] and Nissen and Shemwell [52] suggest that inequities in self-efficacy are a systemic feature of physics education and addressing these inequities at all levels is a critical component of educational design practices.

There have been various explanations provided to account for the gender gap in STEM [36,48,53–55]. There is an association between STEM attrition and declining self-efficacy in STEM early in college [56,57]. A meta-analysis covering four decades identified "five meta-narratives: individual background characteristics; structural barriers in K-12 education; psychological factors, values, and preferences; family influences and expectations; and perceptions of STEM fields" [58] (p. 137). The key here is that the gender gap is a complex interaction of various constructs, some of which can evolve with time. Therefore, linking gendered differences in self-efficacy to any discipline is perhaps not the solution, but rather a step toward exploring why this is the case.

II. CONTEXT

A. Context of the scale under study

Our previous paper detailed the exploratory factor analysis (EFA) computed on 27 items [28] intended to measure the level of students' self-efficacy in both astronomy and dealing with robotic telescope systems. We identified two scales of high reliability. The first, with a Cronbach's alpha of 0.93, measured students' sense of self-efficacy in relation to the current state of their astronomical knowledge. We named this scale *Astronomy Personal Self-Efficacy* (APSE). The second, with a Cronbach's alpha of 0.88, measured students' sense of efficacy in utilizing the associated hardware and software associated with online robotic telescopes. We named this scale *Instrumental Self-Efficacy* (ISE).

In the EFA paper, two problems with the ISE scale were noted: it was highly skewed for the students who had used robotic telescopes in their lab work and it appeared to suffer from a ceiling effect given that there were only five items in the scale that related to this construct with those students' prior experiences influencing their initial self-efficacy. Since the eventual aim of our project is to investigate causal path models that can help explain relationships among various constructs such as self-efficacy, attitudes toward science, science identity, science performance, and career intentions in the STEM domain, we had a problem, *viz.*, the scales have to be largely multivariately *normal*. The original ISE scale clearly did not meet these requirements.

In subsequent research, we explored how a number of items probing the level of students' *confidence* in dealing with aspects of robotic telescope operations and the images they produced could improve the construct of ISE. Using the outcomes of our previous EFA study, our subsequent approach involved a confirmatory factor analysis (CFA), before undertaking reliability analyses and construct validity analyses [59]. Such a CFA approach is driven by theory with collected data being evaluated in ways as to how well the model fits the data. We demonstrated that the two modified factors of APSE and ISE were very robust. The scales possessed high reliabilities (Cronbach's alphas of 0.895 and 0.917 for the APSE scale and 0.920 and 0.929 for the ISE) on the pre- and postoccasions, respectively. We also demonstrated that both scales possessed high construct validities on both occasions of testing.

B. Context of the curriculum

In this paper, we report the results of a repeated measures investigation into the changes in these two aspects of

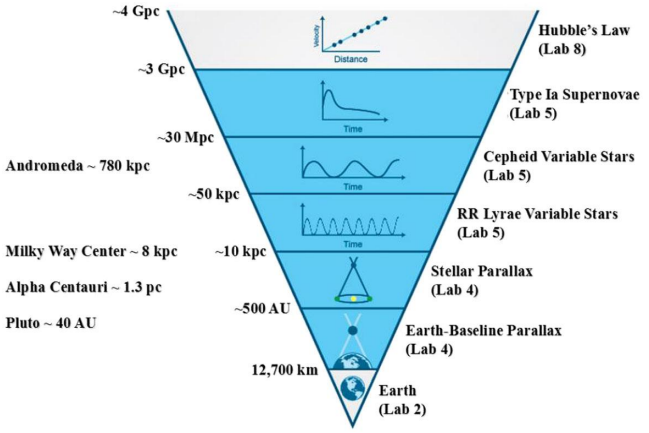

FIG. 2. OPIS! is a series of eight, interconnected labs that teach the evolution of our understanding of Earth's place in the universe, and the cosmic distance ladder. Lab 6 makes use of what was learned in lab 5 to teach the Great Debate. In lab 7, students use Skynet to collect 21-cm spectroscopy from Green Bank Observatory's 20 m-diameter radio telescope to measure the Galaxy's rotation curve and mass distribution [60].

self-efficacy of students who had been involved in using Skynet. Accompanying this system is a specialist curriculum (OPIS!) that makes extensive use of Skynet to generate astronomical data for students to use in their Astro101 course and also addresses the common topics typically found in such courses.

Within the OPIS! course, there is an introductory lab where students learn how to use (i) Skynet and (ii) the image-analysis application—Afterglow. Students (individually or collaboratively as groups or as a whole class) collect and make measurements of their images to distinguish between geocentric and heliocentric models. To do so, they use the phase and angular size of Venus (lab 3), measure the mass of a Jovian planet using the orbit of one of its moons using Newton's modification of Kepler's third law (lab 3), measure the distance to an asteroid using parallax measured simultaneously by Skynet telescopes in different hemispheres (lab 4) and measure the distance to a globular cluster using RR Lyrae stars as standard candles (lab 5) to address the Great Debate about the nature of "spiral nebulae." Further work is done with archival data that takes longer than a semester to collect (e.g., cepheid stars, type Ia supernovae, etc.). See Fig. 2 for more.

III. METHOD

Using the initial astronomy self-efficacy instrument that we developed [28], data were collected from students enrolled in Astro101 courses in the Fall of 2020, Spring 2021, and Summer 2021 across 22 schools, colleges, and universities in the United States of America and Canada. Consent was obtained from participants by clicking on the link after the plain language statement of purpose that took them to the questionnaire. The plain language statement informed them of the purpose of the research and that this had been cleared by the administering university under IRB Protocol 20-2062.

Through the analysis of the self-efficacy data, we found that the skewed nature of the ISE scale was going to be a problem in subsequent structural equation modeling analyses. Consequently, in the Fall semester of 2021, changes were made to the self-efficacy instrument by adding additional questions that probed students' confidence in using aspects of robotic telescopes. We tested these items using a confirmatory factor analysis approach. The successful outcomes of this investigation are published in Freed *et al.* (accepted). The two modified self-efficacy scales reported in that paper are used to report the results in this paper.

On the preoccasion of data collection in Fall 2021, a total of 1264 responses were received representing approximately 84% of the 1500 expected participants and 801 on the post-occasion. Extensive data cleaning was performed involving a variety of approaches. The first level of cleaning involves a custom python script developed by and based on the work of Salimpour *et al.* [61,62]. The automated python script removes extra columns generated by Qualtrics (e.g., external reference, distribution channel, user language, Q recaptcha score) from the raw data; recodes required written script to numerical codes; and, labels the variable columns accordingly. Having an automated cleaning script means that new data can easily and consistently produce a comma-separated values file for import into the Statistical Package for the Social Science v28 (spss) [63] where more detailed data cleaning and statistical analyses can be affected. This next step of data cleaning involved deleting incomplete responses followed by detection and deletion of any duplicate entries detected by spss where an individual with the same student ID from the same institution had attempted the questionnaire twice or more. We also undertook further data cleaning through visual scanning, automatic detection of anomalous responses, and detection of various forms of pattern marking for which we had written extensive spss syntax. We visually inspected those cases identified by spss and our syntax before either accepting or deleting them.

The net effect of this extensive data cleaning yielded 1117 cases (approximately 75% of the expected participants) on the preoccasion of data collection before students had begun to engage with the Astro101 course. The cleaning process was repeated again at the end of their course to produce the postoccasion dataset of 705 valid cases from the original 801 responses submitted. The amalgamated data-matching procedure employed by spss revealed a total of 1301 students who had supplied a response on at least one of the two occasions of data collection. Of these, 521 students completed both the pre- and the postquestionnaires.

Before proceeding any further, we investigated the data to see if there was any selection bias given that some institutions had offered course credit for completing the

questionnaire. Here, we compared the means and standard deviations for all those who had submitted a response on both occasions of testing with those who provided a response only on the preoccasion and those who had submitted a response only on the postoccasion.

We were also interested in institutional differences in any self-efficacy changes that may be apparent. This led to us eliminating those institutions where only small numbers of respondents had submitted both pre- and postdata. This led to the elimination of a further 170 cases from 16 institutions leaving a total of 326 matched cases from the remaining 5.

Of this set of 326 matched responses, participants identified their gender as either "prefer not to say" (4) or "other" (1). Given that we also intended to probe the effect of gender on any changes in self-efficacy, we removed these five students from our analyses because the N in each of these two groups does not meet the criterion that any group size should be large enough to provide a reasonable estimate of the mean and standard deviation [64]. Consequently, the removal process left 321 student responses for which we report the results below.

In our analyses, we employ a multivariate analysis of variance (MANOVA) with repeated measures on the occasion of testing for both the pre- and postoccasions of testing for the two self-efficacy dependent variables simultaneously and using institution and gender as the independent variables in attempts to search for changes in the students' reported sense of self-efficacy that may covary with the intervention of the OPIS! curriculum and the use of robotic telescopes. Where a *between-group* analysis of this nature is concerned, the statistic called Box's M is required to ensure that the distributions of variables in the cells of the MANOVA meet the requirements of statistical mathematics. Statistical mathematics relies on the distributions within each cell of the computation being normally distributed both within a cell and overall. That is to say, the distributions of the dependent variables (DVs) individually grouped by the independent variable(s) (IVs) and collectively must all be distributed approximately normally. In short, Box's M tests the null hypothesis that the observed "covariance matrices of the dependent variables are equal across groups" [63]. If the covariance matrices are not equal, then the significance of any main effects or interactions cannot be interpreted with any great degree of confidence. When Box's M is significant, mathematical transformations of the errant DVs should be calculated so that the covariance matrices are not significantly different from equality across groups. It is only then that the statistical output can be regarded with any confidence.

We employ a graphical approach to explore significant interactions in the output of the MANOVA. For example, as we illustrate below in the *between-groups analysis*, there are significant first-order interactions between the occasion of testing and institution and for the occasion of testing and gender. To investigate the interactions, we first plot the means of the four DVs for the different institutions on one graph and for women and men on both occasions of testing on a second graph. We then inspect the respective gradients of Astronomy Personal Self-Efficacy (APSE) and Instrumental Self-Efficacy (ISE). This approach allows the reader to assess our claims quickly and visually rather than poring over the tables of numbers that we also supply. In addition, we also compute *Cohen's d* effect sizes to explore the magnitude of any changes that have occurred and to triangulate our interpretation of the graphical output.

IV. RESULTS

Table I shows the means, standard deviations, and Ns of both self-efficacy variables for the two occasions of testing. In order to probe whether there were any *selection effects* or *coercion effects*, we introduced an independent variable to reflect the occasions for which we received a valid response (1 = preoccasion only, 2 = postoccasion only, 3 = both occasions).

We computed four analyses of variance using the responses supplied for each of the two self-efficacy variables. That is to say, the group of 596 who supplied preoccasion-only data is compared with the 521 who supplied data on both occasions as well as those 184 who responded only on the postoccasion compared with the 521 who supplied data on both occasions.

It should be noted that because four univariate statistical tests are being computed, we modified the p value below which significance is indicated rather than using the "normal" 95% level of confidence where an apparent claim that something is statistically significant is often made. The reason for the modification given the four univariate tests is because we would not have the 95% level of confidence in saying that something was significant when in fact the actual level of confidence is only approximately 81% given that four univariate tests are being computed (i.e., the actual confidence level is $(0.95)^4$). Thus, in order to maintain an overall 95% level of confidence, the p value for each test has to be lowered. Thus, we apply a Sidak's adjustment to the p value of 0.05 because the DVs are correlated with each other (mean Pearson's $\rho = 0.422$). This yields a value of $p < 0.023$ below which significance can be claimed.

The analysis revealed that there were no significant differences in the means of the APSE and ISE self-efficacy variables between the 596 respondents who supplied data only on the preoccasion and the 521 who supplied data on both occasions. However, on the postoccasion only, there was a significant difference in the APSE variable $[F(1, 704) = 9.892,\ p = 0.002]$ but not for the ISE variable $[F(1, 704) = 4.862,\ p = 0.028]$ for the 184 who supplied data on only the postoccasion and the 521 who supplied data on both occasions. Those who supplied post-only data had a mean score of 2.7 APSE points lower than those who supplied data on both occasions.

If there had been a *selection effect* for this group (the 184 respondents), then they may have been motivated to

TABLE I. Results of significance tests to evaluate selection effects in the pre- and postmeans on both self-efficacy instruments, APSE and ISE, using the entire study population (both matched and unmatched).

Response occasion		Pre-APSE	Post-APSE	Pre-ISE	Post-ISE
Preoccasion only	Mean	34.148		27.5369	
	Standard deviation	17.678		18.715	
	N	596		596	
Postoccasion only	Mean		35.246*		54.250
	Standard deviation		10.597		17.585
	N		184		184
Both occasions	Mean	32.687	37.923*	27.177	57.459
	Standard deviation	17.583	9.680	18.804	16.748
	N	521	521	521	521
Total	Mean	33.466	37.224	27.369	56.621
	Standard deviation	17.641	9.989	18.749	17.016
	N	1117	705	1117	705

*Indicates a significant difference with $p < 0.023$ (Sidak's adjustment).

respond to gain the extra credit offered by instructors while not realizing that supplying only one set of data made them ineligible to receive the five or ten points of credit on offer. There may be a small selection effect for the group who supplied data on both occasions and who wished to receive credit points for completing both questionnaires. This cannot be interpreted as *coercion* for at least three reasons: given that the research team was not involved in the collection of data; the fact that a large number of respondents (596) completed only the preoccasion questionnaire while a smaller number (184) completed only the postoccasion version; and no institution provided a complete set of data on both occasions of testing. Thus, we can reasonably conclude that there is little, if any, selection effect with respect to the changes in self-efficacy detected in the matched pre- and postdata. That is, the changes are likely to be real.

The three research questions that we were interested in probing for the matched dataset involved computing a MANOVA with repeated measures on the occasion of testing involving both the efficacy dependent variables (DVs) and gender of the respondents simultaneously. The first research question relates to the different ways that the OPIS! program and robotic telescope use could be implemented by the different institutions and the potential effects on self-efficacy. That is to say, "Are there any institutional differences that covary with the implementation of the OPIS! program?". The second question relates to the gender of the respondents: "Is there a difference in the way that women and men react to the OPIS! program?". The third research question relates to any potential effects of implementing the OPIS! program on the self-efficacy of respondents. That is, "Does engagement in the OPIS! program involving robotic telescopes in their Astro101 course *covary* with any changes in the self-efficacy of students who engage with it?" That is to say, is there a *Main Effect* due to the occasion of testing? These questions can be answered simultaneously using just one MANOVA. We do not mean to imply that any main effect is "caused" by the OPIS! program because we do not have a control group who did not use the OPIS! course or robotic telescopes.

In the analysis that follows, we have eliminated those institutions where the number of responses was too low to make any safe inferences about the findings and have matched the data supplied by the respondents from the pre- to the postoccasion in five institutions. We have also eliminated those individuals who do not identify their gender in a binary way (female or male) because the numbers in the other categories are too low. This yielded five institutions with large enough *N*s on the pre- and postoccasions of testing to produce an observed power >0.6 for main effects and interactions indicating that the likelihood of the actual differences indicated in the analysis is real. In particular, the observed power for all main effects was greater than 0.97, and significant first- and second-order interactions were greater than 0.65. We can thus be reasonably confident in the findings as they relate to this research.

Table II shows the means, standard deviations, and *N*s of the 321 students in the five institutions who identified their gender in a binary way and who supplied data on both the pre- and postoccasions of testing. Examination of Table II can lead the observer to note that the means of both scales for all of the institutions differ on both the preoccasion and the post-occasion. Moreover, the means increase from the pre- to postoccasion of testing with the ISE means increasing more than the APSE means. In addition, the mean scale scores are lower for women on the preoccasion of testing than for men but the gap closes quite markedly on the postoccasion of testing.

For this between-groups MANOVA with repeated measures on the occasion of testing using institution and gender

TABLE II. Matched pre- and postdescriptive statistics of the self-efficacy measures (APSE and ISE) by gender and institution.

		Pre-APSE			Post-APSE	
Institution	Gender	Mean	Standard deviation	*N*	Mean	Standard deviation
1	Female	31.024	15.946	42	37.403	11.608
	Male	38.568	18.435	37	36.930	10.323
	Total	*34.557*	*17.460*	*79*	*37.182*	*10.958*
16	Female	26.932	14.746	59	37.097	9.006
	Male	34.897	17.077	58	40.832	6.796
	Total	*30.880*	*16.371*	*117*	*38.948*	*8.171*
17	Female	30.630	15.252	27	36.014	10.115
	Male	43.292	17.033	24	41.082	8.496
	Total	*36.588*	*17.180*	*51*	*38.399*	*9.640*
20	Female	22.484	15.015	31	38.005	8.562
	Male	38.053	12.638	19	39.614	7.995
	Total	*28.400*	*15.968*	*50*	*38.617*	*8.306*
21	Female	23.786	15.473	14	29.316	12.380
	Male	25.500	10.395	10	30.045	10.470
	Total	*24.500*	*13.355*	*24*	*29.620*	*11.387*
Total	Female	27.451	15.390	173	36.535	10.212
	Male	36.946	16.907	148	39.012	8.857
	Total	*31.829*	*16.766*	*321*	*37.677*	*9.676*

		Pre-ISE			Post-ISE	
		Mean	Standard deviation	*N*	Mean	Standard deviation
1	Female	25.000	17.518	42	54.286	16.407
	Male	33.973	20.582	37	53.081	18.087
	Total	*29.203*	*19.420*	*79*	*53.722*	*17.113*
16	Female	17.525	15.133	59	63.441	11.856
	Male	30.431	16.781	58	63.155	12.084
	Total	*23.923*	*17.172*	*117*	*63.299*	*11.919*
17	Female	26.556	20.083	27	54.370	19.109
	Male	44.583	19.633	24	60.667	17.074
	Total	*35.039*	*21.671*	*51*	*57.333*	*18.277*
20	Female	24.097	19.393	31	62.935	12.091
	Male	27.737	13.678	19	61.526	10.532
	Total	*25.480*	*17.383*	*50*	*62.400*	*11.434*
21	Female	19.000	11.293	14	45.786	22.192
	Male	32.000	12.832	10	49.800	15.533
	Total	*24.417*	*13.393*	*24*	*47.458*	*19.413*
Total	Female	22.046	17.333	173	58.283	16.172
	Male	33.372	18.290	148	59.122	15.276
	Total	*27.268*	*18.631*	*321*	*58.670*	*15.746*

as the IVs, Box's $M = 117.409$, $p = 0.089$, which means that the equality of the covariance matrices can be considered equal thus allowing us to interpret the output with confidence. The null hypotheses in this analysis are as follows:

(i) There is no significant difference in the means of the self-efficacy DVs across the various institutions;

(ii) There is no significant difference in the means of the self-efficacy DVs of women and men; and,

(iii) There is no significant difference between the means of the self-efficacy DVs from the pre- to the postoccasion of testing for those who supplied data on both occasions.

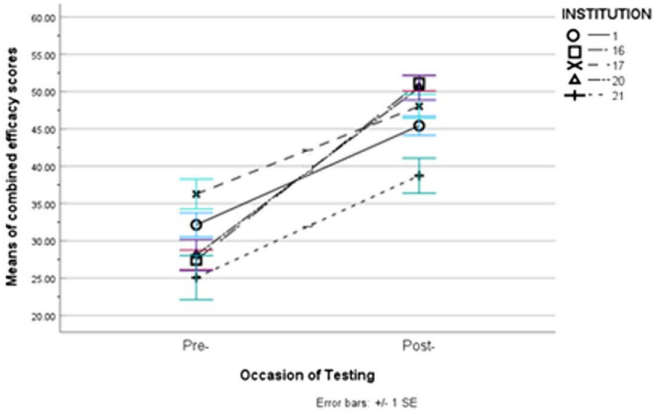

FIG. 3. First-order interaction of combined mean pre- and postscores of efficacy variables by institution.

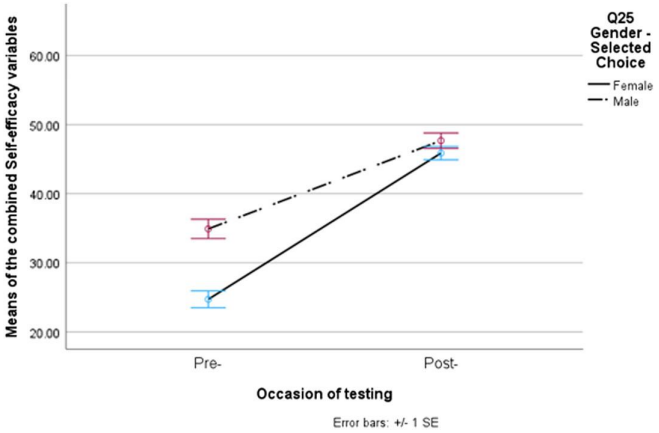

FIG. 4. First-order interaction of the pre- and post-self-efficacy scores and gender.

There are two *between-groups main effects*. The first is due to the institution to which respondents belong $[F(4, 311) = 3.763,\ p = 0.005]$. This result indicates the mean scores of the combined self-efficacy variables are significantly different across the institutions. This result leads us to reject the first null hypothesis. The second between-groups main effect is due to gender $[F(1, 311) = 19.057,\ p < 0.0001]$. This indicates that the mean scores of the combined self-efficacy variables for the women and the men are significantly different from each other. This result leads us to reject the second null hypothesis of no difference between the genders.

In this between-group analysis, there is no significant first-order interaction between institution and gender $[F(4, 311) = 0.855,\ p = 0.492]$. That is to say, the pattern of results for women and men is largely the same across all institutions. While at first glance, these results are a significant finding, in the MANOVA with repeated-measures domain, the main effects require further careful examination, more especially in the within-groups analysis where significant first and higher-order interactions are found. Indeed, the story is slightly more complicated than these between-groups main effects would indicate.

The analysis showed that there are two significant within-groups main effects. The first is due to the occasion of testing $[F(1, 311) = 320.761,\ p < 0.0001]$. This result leads us to reject the third null hypothesis of no difference between the pre- and post-test self-efficacy scores. This means that there is a significant change in the means of the combined self-efficacy variables from the pre- to the postoccasion for all respondents. The second *within-groups main effect* is due to the efficacy variables $[F(1, 311) = 143.142,\ p < 0.0001]$. That is to say, there is a significant difference in the means of the two combined self-efficacy variables over both occasions of testing across the five institutions.

As noted above, of greater interest to us are the within-group interactions. There are three first-order interactions that we explore in Figs. 3–5. There is one significant second-order interaction that we explore in Fig. 6. Please note that all error bars mean ± 1 standard error of the mean.

The first significant first-order interaction is between the occasion of testing and the institution $[F(4, 311) = 10.349,\ p < 0.0001]$. This indicates that respondents in the different institutions vary in their responses to the self-efficacy variables. In this component of the computation, MANOVA computes the mean score of each individual's self-efficacy score for both the pre- and postoccasion of testing before computing these single scores as an ANOVA by institution as the IV.

The significant first-order between-groups interaction $[F(4, 311) = 10.349,\ p < 0.0001]$, illustrated in Fig. 3, may be attributed to the differing gradients for the five institutions from the pre- to the postoccasion of testing. Two of the institutions (16 and 20) start near the bottom of the institution mean scores and end up at the top on the postoccasion. The other three institutions (1, 17, and 21) have shallower gradients and are largely parallel. Taken together with the between-subject main effect for institution, this illustrates the potentially differing covarying effects of the OPIS! program in the different institutions.

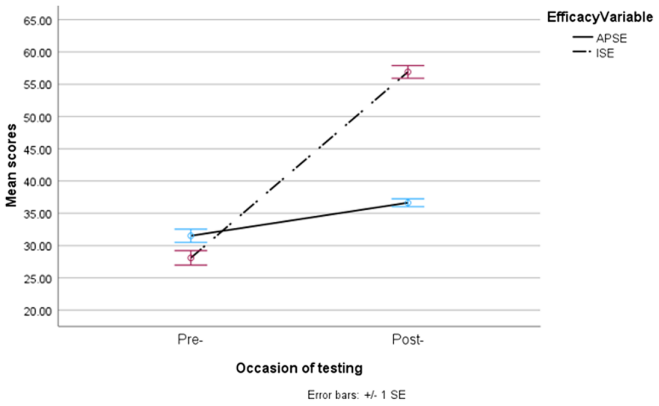

FIG. 5. Mean scores showing the first-order interaction of occasions and the two self-efficacy variables.

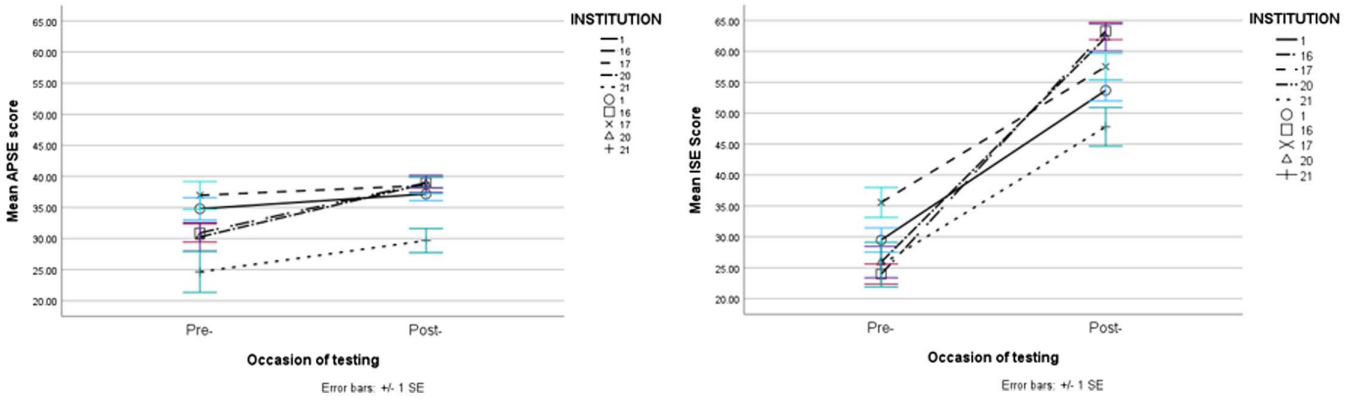

FIG. 6. Second-order interaction of occasion of testing, the self-efficacy scales and institution. Left: APSE; right: ISE.

Figure 4 above illustrates the probable reason for the second significant first-order interaction between the occasion of testing and gender $[F(1,311) = 19.620, p < 0.0001]$. Here the likely cause of the interaction is the fact that the mean of the combined self-efficacy scores for the women is much lower on the preoccasion of testing compared with the men while on the postoccasion, the difference is much less and the standard errors of the means overlap. Another way of expressing this significant interaction is to say that by the end of the semester, there is no significant difference in the combined mean self-efficacy scores of women and men.

Figure 5 helps illustrate the probable cause of the significant first-order interaction $[F(1,311) = 381.231, p < 0.0001]$ between the occasion of testing and the two self-efficacy variables. One can see that the gain in the ISE variable is much greater than the gain in the APSE variable. Thus, the slope for the ISE variable is much greater than the slope for the APSE variable. While this is an expected outcome given that the respondents had not used robotic telescopes before, the magnitude of the difference in ISE is large, while simultaneously their self-efficacy in relation to their astronomical knowledge has increased by a much smaller margin.

The significant second-order within-groups interaction for the occasion of testing, the two self-efficacy variables, and the institution to which respondents belong $[F(4,311) = 5.791,\ p < 0.001]$ are explored in Fig. 6. Here two graphs are employed side-by-side with one for each efficacy variable. The graph on the left shows the mean scores of the APSE scale for each institution on the two occasions of testing while the one on the right shows the mean scores for the ISE.

Observation of Fig. 6 shows that the probable cause of the significant second-order interaction is likely due to the major difference in gradients for the APSE and ISE scales, and the fact that two of the five institutions (16 and 20) have greater gains in both the self-efficacy scales compared with the other three. This probably indicates that the respondents in the different institutions reacted to the two efficacy scales in different ways.

Of particular note in this analysis is the lack of a significant fourth-order interaction of occasion of testing, efficacy variables, institution, and gender $[F(4,311) = 2.025, p = 0.091)$. This indicates the men and women in all of the institutions reacted to the two efficacy scales on the two occasions of testing in similar ways. That is to say, while there were institutional differences as indicated by the significant between-groups main effect as well as gender differences in the two self-efficacy scales, using the analyses of the significant first- and second-order interactions, we are able to conclude that the introduction of the OPIS! program covaried with major changes in self-efficacy and with greater changes in the self-efficacy of women in all of the institutions compared with their male peers.

A. Effect sizes

In this section, we explore the *effect sizes* for the difference in means of both the institutions and the women's and men's scores using the two scales for self-efficacy. "Effect size" is a measure of how large the difference is in the means of two scores. We use Cohen's d because it expresses the difference in means in terms of a standard deviation (σ or sigma) carefully chosen as the divisor of the difference in the two means on the *Occasions of Testing*. We have chosen to use the "pooled standard deviation" because the standard deviations are different on the two occasions of testing [65].

Cohen defines a value for the *Effect Size* as being "*very small or inconsequential*" if *Cohen's* $d < 0.2$, *small* if it is in the range of 0.2 to 0.4, *moderate* if d is between 0.4 and 0.7, and *large* if d lies between 0.7 and 1. The effect size can be described as *very large* if Cohen's d is greater than 1.0 [65,66]. To obtain a Cohen's d of greater than "2" in a group learning situation is highly unusual.

Bloom [67] refers to very large effect sizes greater than 2 as the "2-sigma problem" where the outcomes of the learning are mostly achieved by one-to-one tutoring while

TABLE III. Cohen's *d* effect sizes for the self-efficacy variables, by gender and occasion of testing.

Efficacy scale	Gender-occasion	Mean	σ	N	Pooled σ	Cohen's d
APSE	Female-pre	27.451	15.390	173	13.060	0.696
	Female-post	36.535	10.212	173		
	Male-pre	36.946	16.907	148	13.496	0.153
	Male-post	39.012	8.857	148		
	All-pre	31.829	16.766	321	13.688	0.427
	All -post	37.677	9.676	321		
ISE	Female-pre	22.046	17.333	173	16.763	2.162
	Female-post	58.283	16.172	173		
	Male-pre	33.372	18.290	148	16.851	1.528
	Male-post	59.122	15.276	148		
	All-pre	27.268	18.631	321	17.249	1.821
	All-post	58.670	15.746	321		

sigmas greater than "1" are achieved mostly through "mastery learning" [67]. The challenge for educators is to design learning environments and programs that can produce the learning gains corresponding to one-to-one tutoring but within an educational framework that is economically justified and sustainable.

Table III shows the magnitude of the effect sizes for each of the two self-efficacy variables: APSE and ISE for women and men separately in the five institutions and of note is the 2-sigma effect size for women in the Instrumental Self-Efficacy scale. Also of interest is the effect size for men in the same ISE variable (*Cohen's* $d = 1.528$) which is also "very large" and symptomatic of approaches that involve mastery learning. Also of interest are the much lower effect sizes for the APSE scale for both women and men. Nonetheless, the effect size for women can be described as "moderate" while for men it is "very small" (i.e., <0.2). We will address these issues in the discussion.

Table IV shows the descriptive statistics for each of the five institutions and the 321 cases that we used to compute the MANOVA with repeated measures on the occasion of testing for the APSE and ISE variables. The table shows varying sizes of gains. Some of the institutions appear to have made larger gains in the ISE scale, such as the ones coded as "16" (Gain = 40.1) and "20" (Gain = 36.9), while others such as "1" and "17" made more modest gains (24.3 and 22.3, respectively)., In their APSE scores, all institutions made much more modest gains with the institutions labeled as "16" and "20" making the largest (8.1 and 10.2, respectively).

We can also probe the graphical evaluations of the first-order interactions presented in the above figures using the effect sizes achieved by each institution. Table IV shows

TABLE IV. Descriptive statistics of APSE and ISE by institution with Cohen's *d* effect size.

	APSE pre- and post-				ISE pre- and post-		
Institution-occasion	N	Mean	Standard deviation	Cohen's d	Mean	Standard deviation	Cohen's d
1-pre	79	34.557	17.460	0.180	29.203	19.420	1.340
1-post	79	37.182	10.958		53.722	17.113	
16-pre	*117*	*30.880*	*16.371*	*0.624*	*23.923*	*17.172*	*2.664*
16-post	*117*	*38.948*	*8.171*		*63.299*	*11.919*	
17-pre	51	36.588	17.180	0.130	35.039	21.671	1.112
17-post	51	38.399	9.640		57.333	18.277	
20-pre	*50*	*28.400*	*15.968*	*0.803*	*25.480*	*17.383*	*2.509*
20-post	*50*	*38.617*	*8.306*		*62.400*	*11.434*	
21-pre	24	24.500	13.355	0.413	24.417	13.393	1.382
21-post	24	29.620	11.387		47.458	19.413	
Total-pre	321	31.829	16.766	0.427	27.268	18.631	1.821
Total-post	321	37.677	9.676		58.670	15.746	

Cohen's *d* for each institution for both efficacy scales over the two occasions of testing. Two of the five institutions have effect sizes greater than 2 sigma on the ISE scale. Indeed, Table IV shows that the institution identified as "16" for the ISE variable shows the largest effect size of 2.664 and a moderate effect size of 0.624 for the APSE variable, while the one identified as "20" shows an effect size of 2.509 for changes in the ISE scale and a large effect size of 0.803 in the APSE scale. The examination of the effect sizes presents a consistent picture of our graphical interpretations of the first- and second-order interactions computed in the MANOVA for the APSE and ISE variables by institution and gender.

V. DISCUSSION

We have used the two self-efficacy scales of APSE and ISE, developed by the authors, to probe their covariance with the respondents' experience of an Astro101 course that employs robotic telescopes for students to collect astronomical data. Students then use these data in their laboratory sessions to cover the normal content of such a course. We cannot say that the robotic telescope experiences have caused these changes in the self-efficacy scales since we did not have a control group over the extent to which robotic telescopes were used or the extent to which the OPIS! program was implemented by instructors. As we progress this research with future cohorts of students using a multiple-baseline, multiple-probe research design, we will be able to attribute a degree of causality to any changes that consistently appear from semester to semester. It will also allow for further interrogations into the delivery of the course across multiple institutions and determine how the depth of delivery impacts the self-efficacy of the students.

We have demonstrated the usefulness of the two self-efficacy scales, shown to be reliable and valid in an earlier paper [28], in probing changes that occur over the course of a semester of study with women and men as well as with different institutions. Future research planned by the research team will allow us to probe such things as the effectiveness of the approach; the effect of instructor experience through repetition of delivery; the emergence of science identity and testing of the CUREs pathway of Wooten *et al.* [16].

In the context of this current study, the decrease in the self-efficacy gap between women and men on the post-occasion provides further evidence that authentic experiences with regard to STEM may support the positive changes observed here in self-efficacy although we do not mean to imply causality. This echoes the work of Bandura [18,68], with regard to mastery experiences, and the work by Lent *et al.* [21], who showed such experiences as being the strongest contributor to positive changes in mathematics self-efficacy. However, the work of Zeldin and Pajares [69] showed that "vicarious experiences and verbal persuasions were instrumental sources for the development and maintenance of self-efficacy beliefs for women in mathematics-related careers." (p. 227). In the context of our study, perhaps there is an interaction between the mastery required to use robotic telescopes and the other vicarious experiences that contribute to the positive changes in self-efficacy. This may be due to the nature of the OPIS! curriculum, the use of research-grade instruments, and the fact that the students come from a variety of non-STEM majors. There is an extensive body of research spanning decades that highlights the fact that students' experiences and perceptions of science are not so positive during school [70–75].

Nonetheless, the positive change in self-efficacy evidenced by the decrease in the gender self-efficacy gap is an important finding. This is because irrespective of the factors that negatively impact self-efficacy, and that self-efficacy is shaped early [76], it would appear that there are ways to change female students' perceptions through such authentic experiences more than males. Therefore, authentic telescope use within this collaborative Astro101 setting appears potentially to help redress gender biases in physics education.

VI. LIMITATIONS

Gender self-efficacy is a complex construct encompassing a variety of factors, which may impact it. While this was not a direct focus of this study, it would lend itself to future investigations in our research context. Furthermore, while the delivery method within the different institutions was not a focus of this paper; it may impact the generalizability of the effect sizes we found. Implementation integrity is thus important, which lends itself to significant further studies. Another constraint on the generalizability of findings in this research relates to the fact that the statistical analysis was constrained to those institutions that provided a large enough N of matched responses on both occasions of testing. The situation may become clearer and more generalizable as we amalgamate further data over subsequent semesters from the smaller institutions to achieve large enough Ns as we progress the research using the same efficacy scales.

VII. CONCLUSION

Astronomy 101 courses with remote access to telescopes may play a significant role in increasing self-efficacy in this STEM domain with a special emphasis on closing the gender gap in STEM self-efficacy. This study covered a broad and diverse range of institutions, including community colleges and four-year research-focused universities, and found self-efficacy increases in all cases. While female respondents started out with lower self-efficacy in both the APSE and ISE constructs, the gender gap was closed, with no significant difference in either measure at the end of a semester with the implementation of robotic

telescope-based labs. Importantly, this research provides an encouraging insight into ensuring that engaging young women in authentic science experiences, while increasing their self-efficacy, may work toward developing a greater flow through the STEM pipeline.

ACKNOWLEDGMENTS

This work was supported by NSF Award No. 2013295, Project Title: Collaborative Research: Exploring the Impact of Robotic Telescope-Based Observing Experiences on Students' Learning and Engagement in STEM.

[1] S. R. Buxner, C. D. Impey, J. Romine, and M. Nieberding, Linking introductory astronomy students' basic science knowledge, beliefs, attitudes, sources of information, and information literacy, Phys. Rev. Phys. Educ. Res. **14**, 010142 (2018).

[2] R. A. Duschl, H. A. Schweingruber, and A. W. Shouse, *Taking Science to School: Learning and Teaching Science in Grades K-8* (National Academies Press, Washington, DC, 2007).

[3] B. Partridge and G. Greenstein, Goals for "Astro 101:" Report on workshops for department leaders, Astron. Educ. Rev. **2**, 46 (2003).

[4] A. Fraknoi, Insights from a survey of astronomy instructors in community and other teaching-oriented colleges in the United States, Astron. Educ. Rev. **3**, 7 (2004).

[5] C. Impey, Science literacy of undergraduates in the United States, Organ. People Strategies Astron. **2**, 353 (2013), https://www.researchgate.net/profile/Chris-Impey/publication/258843477_Science_Literacy_of_Undergraduates_in_the_United_States/links/5512e9980cf270fd7e33e3c6/Science-Literacy-of-Undergraduates-in-the-United-States.pdf.

[6] A. L. Rudolph, E. E. Prather, G. Brissenden, D. Consiglio, and V. Gonzaga, A national study assessing the teaching and learning of introductory astronomy part II: The connection between student demographics and learning, Astron. Educ. Rev. **9**, 5 (2010).

[7] M. A. Cannady, E. Greenwald, and K. N. Harris, Problematizing the STEM pipeline metaphor: Is the STEM pipeline metaphor serving our students and the STEM workforce?, Sci. Educ. **98**, 443 (2014).

[8] L. Linnenbrink-Garcia, T. Perez, M. M. Barger, S. V. Wormington, E. Godin, K. E. Snyder, K. Robinson, A. Sarkar, L. S. Richman, and R. Schwartz-Bloom, Repairing the leaky pipeline: A motivationally supportive intervention to enhance persistence in undergraduate science pathways, Contemp. Educ. Psychol. **53**, 181 (2018).

[9] E. Gomez and M. Fitzgerald, Robotic telescopes in education, Astron. Rev. **13**, 28 (2017).

[10] M. T. Fitzgerald, R. Hollow, L. M. Rebull, L. Danaia, and D. H. McKinnon, A review of high school level astronomy student research projects over the last two decades, Pub. Astron. Soc. Aust. **31**, e037 (2014).

[11] P. M. Sadler, R. R. Gould, P. S. Leiker, P. R. A. Antonucci, R. Kimberk, F. S. Deutsch, B. Hoffman, M. Dussault, A. Contos, K. Brecher, and L. French, MicroObservatory Net: A network of automated remote telescopes dedicated to educational use, J. Sci. Educ. Technol. **10**, 39 (2001).

[12] A. Bain and M. E. Weston, *The Learning Edge: What Technology Can Do to Educate All Children* (Teachers College Press, Columbia University, New York, 2011).

[13] M. E. Weston and A. Bain, The end of techno-critique: The naked truth about 1:1 laptop initiatives and educational change, J. Technol. Learn. Assess. **9** (2010), https://ejournals.bc.edu/index.php/jtla/article/view/1611.

[14] R. Saubern, M. Henderson, E. Heinrich, and P. Redmond, TPACK—time to reboot?, Australas. J. Educ. Tech. **36**, 1 (2020).

[15] D. E. Reichart, Robotic telescope labs for survey-level undergraduates, Phys. Teach. **59**, 728 (2021).

[16] M. M. Wooten, K. Coble, A. W. Puckett, and T. Rector, Investigating introductory astronomy students' perceived impacts from participation in course-based undergraduate research experiences. Phys. Rev. Phys. Educ. Res. **14**,010151 (2018).

[17] S. Bartlett, M. T. Fitzgerald, D. H. McKinnon, L. Danaia, and J. Lazendic-Galloway, Astronomy and science student attitudes (ASSA): A short review and validation of a new instrument, J. Astron. Earth Sci. Educ. **5**, 1 (2018).

[18] A. Bandura, *Self-Efficacy: The Exercise of Control* (W. H. Freeman and Company, New York, 1997).

[19] E. L. Usher and F. Pajares, Sources of self-efficacy in mathematics: A validation study, Contemp. Educ. Psychol. **34**, 89 (2009).

[20] C. B. Hodges and P. F. Murphy, Sources of self-efficacy beliefs of students in a technology-intensive asynchronous college algebra course, Internet Higher Educ. **12**, 93 (2009).

[21] R. W. Lent, F. G. Lopez, and K. J. Bieschke, Mathematics self-efficacy: Sources and relation to science-based career choice—PsycNET, J. Counsel. Psychol. **38**, 424 (1991).

[22] H. P. Phan, Relations between informational sources, self-efficacy and academic achievement: A developmental approach, Educ. Psychol. **32**, 81 (2012).

[23] F. Pajares, Self-efficacy beliefs in academic settings, Rev. Educ. Res. **66**, 543 (1996).

[24] J. M. Bailey, D. Lombardi, J. R. Cordova, and G. M. Sinatra, Meeting students halfway: Increasing self-efficacy and promoting knowledge change in astronomy, Phys. Rev. Phys. Educ. Res. **13**, 020140 (2017).

[25] H. B. Hewitt, M. N. Simon, C. Mead, S. Grayson, G. L. Beall, R. T. Zellem, K. Tock, and K. A. Pearson, Development and assessment of a course-based undergraduate research experience for online astronomy majors, Phys. Rev. Phys. Educ. Res. **19**, 020156 (2023).

[26] M. Simon, E. Prather, I. Rosenthal, M. Cassidy, J. Hammerman, and L. Trouille, A new curriculum

development model for improving undergraduate students' data literacy and self-efficacy in online astronomy classrooms, Astron. Educ. J. **2**, 1 (2022).

[27] S. Galano, L. Palazzo, and I. Testa, A latent profile analysis of students' attitudes towards astronomy across grades 9–13, Int. J. Sci. Educ. **45**, 1 (2023).

[28] R. Freed, D. McKinnon, M. Fitzgerald, and C. M. Norris, Development and validation of an astronomy self-efficacy instrument for understanding and doing, Phys. Rev. Phys. Educ. Res. **18**, 010117 (2022).

[29] J. R. Cole and S. Cole, *Social Stratification in Science* (University of Chicago Press, Chicago, 1973).

[30] J. R. Cordova, G. M. Sinatra, S. H. Jones, G. Taasoobshirazi, and D. Lombardi, Confidence in prior knowledge, self-efficacy, interest and prior knowledge: Influences on conceptual change. Contemp. Educ. Psychol. **39**, 164 (2014).

[31] L. A. Corwin, M. J. Graham, and E. L. Dolan, Modeling course-based undergraduate research experiences: An agenda for future research and evaluation, CBE—Life Sci. Educ. **14**, es1 (2015).

[32] T. Honicke and J. Broadbent, The influence of academic self-efficacy on academic performance: A systematic review, Educ. Res. Rev. **17**, 63 (2016).

[33] A. D. Stajkovic, A. Bandura, E. A. Locke, D. Lee, and K. Sergent, Test of three conceptual models of influence of the big five personality traits and self-efficacy on academic performance: A meta-analytic path-analysis, Pers. Individ. Diff. **120**, 238 (2018).

[34] G. Trujillo and K. D. Tanner, Considering the role of affect in learning: Monitoring students' self-efficacy, sense of belonging, and science identity, CBE Life Sci. Educ. **13**, 6 (2014).

[35] C. Huang, Gender differences in academic self-efficacy: A meta-analysis, Eur. J. Psychol. Educ. **28**, 1 (2013).

[36] R. S. Barthelemy, M. McCormick, and C. Henderson, Gender discrimination in physics and astronomy: Graduate student experiences of sexism and gender microaggressions, Phys. Rev. Phys. Educ. Res. **12**, 020119 (2016).

[37] L. J. Kewley, Closing the gender gap in the Australian astronomy workforce, Nat. Astron. **5**, 615 (2021).

[38] C. Antolini, O. Katz, and H. Usher, Astronomy for all-all for astronomy? A pilot study of amateur astronomy community attitudes and experiences, Europlanet Science Congress 2020, online, EPSC2020-1084 (2020), 10.5194/epsc2020-1084.

[39] S. Salimpour and M. T. Fitzgerald, A glass ceiling in AER?: A preliminary glimpse at the distribution of authors by gender in the iSTAR (istardb.org) database, *RTSRE Proc.* 2 (2019).

[40] S. Cwik and C. Singh, Damage caused by societal stereotypes: Women have lower physics self-efficacy controlling for grade even in courses in which they outnumber men, Phys. Rev. Phys. Educ. Res. **17**, 020138 (2021).

[41] S. Hand, L. Rice, and E. Greenlee, Exploring teachers' and students' gender role bias and students' confidence in STEM fields. Soc. Psychol. Educ. **20**, 929 (2017).

[42] E. M. Marshman, Z. Y. Kalender, T. Nokes-Malach, C. Schunn, and C. Singh, Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm?, Phys. Rev. Phys. Educ. Res. **14**, 020123 (2018).

[43] J. M. Nissen, Gender differences in self-efficacy states in high school physics, Phys. Rev. Phys. Educ. Res. **15**, 013102 (2019).

[44] J. V. Patterson and A. T. Johnson, High school girls' negotiation of perceived self-efficacy and science course trajectories, J. Res. Educ. **27**, 79 (2017), https://eric.ed.gov/?id=EJ1142363.

[45] M. M. Williams and C. George-Jackson, Using and doing science: Gender, self-efficacy, and science identity of undergraduate students in STEM, J. Women Minorities Sci. Eng. **20**, 99 (2014).

[46] S. V. Rosser, Breaking into the lab: Engineering progress for women in science and technology, Int. J. Gender Sci. Technol. **10**, 213 (2018), https://genderandset.open.ac.uk/index.php/genderandset/article/view/490.

[47] A. Ottemo, A. J. Gonsalves, and A. T. Danielsson, (Dis) embodied masculinity and the meaning of (non)style in physics and computer engineering education, Gender Educ. **33**, 1017 (2021).

[48] R. Ivie and C. Tesfaye, Women in physics: A tale of limits, Phys. Today **65**, No. 2, 47 (2012).

[49] A. Agogino, Beyond bias and barriers: Fulfilling the potential of women in academic science and engineering, in *Proceedings of the APS April Meeting Abstracts* (2007), pp. K6-001.

[50] Women in STEM, edited by S. C. White, Phys. Teach. **57**, 235 (2019).

[51] J. M. Nissen, Are inequities in self-efficacy a systemic feature of physics education? arXiv:1612.09188.

[52] J. M. Nissen and J. T. Shemwell, Gender, experience, and self-efficacy in introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 020105 (2016).

[53] J. C. Blickenstaff, Women and science careers: Leaky pipeline or gender filter?, Gender Educ. **17**, 369 (2005).

[54] E. Seymour and N. M. Hewitt, *Talking about Leaving: Why Undergraduates Leave the Sciences* (Westview Press, Boulder, CO, 1997).

[55] C. Tobias, The gender gap on the Federal Bench, Hofstra Law Rev. **19**, 5 (1990), https://heinonline.org/HOL/LandingPage?handle=hein.journals/hoflr19÷=13&id=&page=.

[56] L. M. Larson, K. M. Pesch, S. Surapaneni, V. S. Bonitz, T.-F. Wu, and J. D. Werbel, Predicting graduation: The role of mathematics/science self-efficacy, J. Career Assess. **23**, 399 (2015).

[57] J. A. Raelin, M. B. Bailey, J. Hamann, L. K. Pendleton, R. Reisberg, and D. L. Whitman, The gendered effect of cooperative education, contextual support, and self-efficacy on undergraduate retention, J. Eng. Educ. **103**, 599 (2014).

[58] M. A. Kanny, L. J. Sax, and T. A. Riggers-Piehl, Investigating forty years of stem research: How explanations for the gender gap have evolved over time, J. Women Minorities Sci. Eng. **20**, 127 (2014).

[59] R. Freed, D. H. McKinnon, M. T. Fitzgerald, and S. Salimpour, Confirmatory factor analysis of two self-efficacy scales for astronomy understanding and robotic telescope use, Phys. Rev. Phys. Educ. Res. **19**, 020164 (2023).

[60] K. Williamson, D. Reichart, C. Wallace, E. E. Prather, and S. Hornstein, Mapping the Milky Way: A radio astronomy-directed investigation for lecture-based Astro 101 courses, RTSRE Proc. **1**, 282 (2018), https://rtsre.org/index.php/rtsre/article/view/18.

[61] S. Salimpour, Visualising the Cosmos: Teaching cosmology in high school in the era of big data, doctoral thesis, Deakin University, 2021.

[62] S. Salimpour, R. Tytler, B. Doig, M. T. Fitzgerald, and U. Eriksson, Conceptualising the Cosmos: Development and validation of the Cosmology Concept Inventory for High School, Int. J. Sci. Math. Educ. **21**, 251 (2023).

[63] SPSS MANOVA, Multivariate analysis of variance (MANOVA), https://www.ibm.com/docs/sl/spss-statistics/beta?topic=statistics-multivariate-analysis-variance-manova (2021).

[64] N. K. Dhand and M. S. Khatkar, Statulator: An online statistical calculator, Sample size calculator for estimating a single mean, Available on August 9, 2022, at http://statulator.com/SampleSize/ss1M.html(2014).

[65] J. C. Goulet-Pelletier and D. Cousineau, A review of effect sizes and their confidence intervals, Part I: The Cohen's *d* family, Quant. Methods Psychol. **14**, 242 (2018).

[66] J. Cohen, *Statistical Power Analysis for the Behavioural Sciences* (Academic Press, New York, 1969).

[67] B. S. Bloom, The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring, Educ. Res. **13**, 4 (1984).

[68] A. Bandura, Self-efficacy mechanism in human agency, Am. Psychol. **37**, 122 (1982).

[69] A. L. Zeldin and F. Pajares, Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers, Am. Educ. Res. J. **37**, 215 (2000).

[70] M. Braund and M. Driver, Pupils' perceptions of practical science in primary and secondary school: Implications for improving progression and continuity of learning, Educ. Res. **47**, 77 (2005).

[71] L. Danaia, M. Fitzgerald, and D. McKinnon, Students' perceptions of high school science: What has changed over the last decade?, Res. Sci. Educ. **43**, 1501 (2013).

[72] J. Osborne, S. Simon, and S. Collins, Attitudes towards science: A review of the literature and its implications, Int. J. Sci. Educ. **25**, 1049 (2003).

[73] P. Potvin and A. Hasni, Analysis of the decline in interest towards school science and technology from grades 5 through 11, J. Sci. Educ. Technol. **23**, 784 (2014).

[74] R. Sheldrake, T. Mujtaba, and M. J. Reiss, Students' changing attitudes and aspirations towards physics during secondary school, Res. Sci. Educ. **49**, 1809 (2019).

[75] S. Tröbst, T. Kleickmann, K. Lange-Schubert, A. Rothkopf, and K. Möller, Instruction and students' declining interest in science: An analysis of German fourth-and sixth-grade classrooms, Am. Educ. Res. J. **53**, 162 (2016).

[76] S. L. Britner and F. Pajares, Sources of science self-efficacy beliefs of middle school students, J. Res. Sci. Teach. **43**, 485 (2006).