Confirmatory factor analysis of two self-efficacy scales for astronomy understanding and robotic telescope use

R. Freed,[1] D. H. McKinnon,[2] M. T. Fitzgerald,[3] and S. Salimpour[4]

[1]*University of North Carolina, Chapel Hill, North Carolina, USA*
[2]*School of Education, Edith Cowan University, Joondalup, Western Australia, Australia*
[3]*Las Cumbres Observatory, Goleta, California, USA and Deakin University, Burwood, Victoria, Australia*
[4]*International Astronomical Union, Office of Astronomy for Education, Heidelberg, Germany and Max Planck Institute for Astronomy, Heidelberg, Germany*

(Received 26 July 2022; accepted 9 November 2023; published 5 December 2023; corrected 9 January 2024)

This paper presents the results of a confirmatory factor analysis on two self-efficacy scales designed to probe the self-efficacy of college-level introductory astronomy (Astro-101) students ($n = 1381$) from 22 institutions across the United States of America and Canada. The students undertook a course based on similar curriculum materials, which involved students using robotic telescopes to support their learning of astronomical concepts covered in the "traditional" Astro-101 courses. Previous research by the authors using these self-efficacy scales within a pre-/post-test approach showed both high reliabilities and very high construct validities. However, the scale purporting to measure students' self-efficacy in relation to their use of the astronomical instrumentation associated with online robotic telescopes was particularly skewed and required further investigation. This current study builds on the previous work and shows how a slight adjustment of the survey items presents an improved and robust scale for measuring self-efficacy.

DOI: 10.1103/PhysRevPhysEducRes.19.020164

I. INTRODUCTION

Self-efficacy is defined as a person's belief that they can succeed in a particular task or activity [1–3]. Self-efficacy is domain-specific or even task-specific [4,5] and is thought to be influenced by four primary sources of information: personal performance accomplishments or mastery experiences, verbal or social persuasion, vicarious learning, and physiological and affective states and reactions [1,6]. Particularly in the domains of science, technology, engineering, and math (STEM), self-efficacy plays an important role in determining students' participation and persistence in these fields [7–11]. Research shows that students with higher self-efficacy are more likely to take courses or follow a STEM career pathway [12–13]. Additionally, higher self-efficacy leads to enhanced science identity which further increases participation and persistence in STEM [14].

With the recent proliferation of robotic telescopes in the context of astronomy education [15] and their potential to engage students in authentic scientific practices, there is a growing need to empirically verify the extent to which engagement with robotic telescopes affects student self-efficacy. Although there is an increasing number of self-efficacy instruments [16–19], measuring self-efficacy is domain-specific [1,4]. Therefore, off-the-shelf-broadly applicable STEM-wide self-efficacy instruments are hard to find and do not really address the construct in the domain in which we were interested: a reconceptualization of the introductory astronomy curriculum typically offered in the United States of America and Canada to make robotic telescope use an integral part of the coursework. This new curriculum with robotic telescope use is entitled Our Place in Space! (hereafter OPIS!). Based on the above, we designed an instrument specific to the domain of robotic telescope usage and astronomy courses.

A. Previous research

Our previous pilot research and exploratory factor analysis (EFA) (EFA pilot henceforth), showed two latent variables were being measured: astronomy personal self-efficacy (APSE) and instrumental self-efficacy (ISE), relating to the use of telescopes and image analysis [20]. In order to robustly assess the adequacy of the hypothesized model for these two latent variables, this current study uses confirmatory factor analysis (CFA) to understand the relationship between the two self-efficacy factors that emerged in the EFA pilot. The EFA pilot study detailed the EFA computed on 27 items [20] we had written in an attempt to measure the task-specific domain of using robotic telescopes. Students used robotic telescopes to capture data, which they examined in an

educational setting to uncover the concepts typically taught in courses commonly described as Astro-101, ASTRO-101, or Astronomy-101 in colleges and universities in the United States of America and Canada. These are typically very broad introductory university courses covering the Solar System, stars, and galaxies with some covering the expansion of the universe.

The analysis in the EFA pilot 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 of an online robotic telescope. We named this scale "Instrumental Self-Efficacy (ISE)." Despite the high reliability, we noted two problems with the ISE scale: (i) it was highly skewed for the students who had used robotic telescopes in their lab work; (ii) it appeared to suffer from a ceiling effect given that there were only five items in the scale and that once the student had learned how to use the technology (mastery), that was it, they knew.

It became clear on the postoccasion of data collection after the students had completed the course (hereafter referred to as "postoccasion") that the skewed nature of the original five-item ISE scale was a problem for our eventual aim of investigating causal path models that can help explain relationships among the constructs of self-efficacy, attitudes toward science, science identity, science performance, and career intentions in the STEM domain. A major requirement of structural equation modeling (SEM) is that statistical distributions of the scales have to be multivarietally normal. The skewed ISE scale clearly could not meet this condition.

Mindful of these caveats, in addition to the original items from the EFA paper, we included a further 11 modified statements modeled on the original items all of which we distributed in the Fall 2021 iteration of the survey. These additional statements probed the students' *confidence* in dealing with the task-specific aspects of the instrumentation associated with robotic telescopes manipulated through the Internet.

We recognize that self-confidence and self-efficacy are not quite the same construct but are related in a task-specific environment. A student's academic self-confidence refers to a student's self-perception of his or her academic abilities [21–22], while his or her self-efficacy refers to an individual's expectations that he or she can succeed in the completion of a specific academic task [23]. The literature states that self-confidence is the more stable construct compared with self-efficacy, which is "context specific" [1,24].

Despite this difference, our reasoning was that the earlier "efficacy" items were phrased in an almost dichotomous manner that attracted responses that were at the extremes of the Likert scale. In contrast, the use of the word "confident" in a context-specific way in conjunction with the same items could be used to evoke a response ranging from "very low" to "very high." That is to say, an individual's level of confidence is not a dichotomous condition. It is more likely to occur on a continuum from "none" to "completely." For example, an earlier item for instrumental self-efficacy asked students to respond to the statement "I am able to request telescope images through a web-based portal" while the new item is "I feel confident that I could show someone how to request an image from a remote telescope using an online portal." The former item infers a "Yes or No" response while the latter requires some reflection. Here, the use of the words "I feel confident that …" is not used to infer "self-confidence." Rather, it is asking the respondent to react to how confident they feel about doing the action of what follows in the item. In this way, we hoped to make their reaction to what follows much more context specific in the same way as self-efficacy items. We distributed the new questionnaire at the beginning of the fall semester of 2021 via Qualtrics to the new cohort of students.

II. METHOD

We distributed the extended questionnaire online to a projected enrollment of almost 1600 students in 22 universities and colleges in the United States of America and Canada at the beginning of the fall semester of 2021 and again at the end. The respondents were students enrolled in an Astro-101 course at their respective institutions. The instructors at these institutions are also participants in the project being conducted by the University of North Carolina-Chapel Hill (UNC-CH) who were to use SKYNET [25] to access a large number of robotic telescopes around the world. All participating students in these classes experienced the use of robotic telescopes, which replaced the "normal" observing experiences of looking through a telescope during an observation session organized by the instructors and tutors of their courses.

Students indicated their agreement to participate in responding to the questionnaire by reading the opening page and clicking the button to continue. Participants were guaranteed complete confidentiality of their identities and responses in the Statement of Ethical Clearance granted by UNC-CH. We collected the students' identification number at their institution so that we could match the pre- to the postoccasion data supplied. Once matched, we deleted the student ID from the dataset.

On the pre-course-occasion of testing (hereafter referred to as preoccasion), we received a total of 1264 responses from the 22 universities and colleges, and 801 on the postoccasion of testing. We undertook extensive data cleaning involving both visual and automatic searches using the Statistical Package for the Social Science v28 (SPSS). We looked for such things as duplicate entries, incomplete responses, and various forms of pattern marking for which we had written extensive syntax using SPSS. We also checked all of the variables simultaneously in the questionnaire for anomalous responses using the automated

SPSS routine included in the package. Those cases found by the software were inspected visually before either accepting or deleting them.

After this checking, the preoccasion data comprised 1117 cases, and the postdata comprised 705 cases where the responses appeared to have been generated in a conscientious manner by the participants. When we cross matched the preoccasion responses to the postoccasion, we found that 521 students had completed both the pre- and the postquestionnaires. This number represents an approximate completion rate of 40% of the total who supplied data on at least one occasion ($N = 1381$). Some, for example, only supplied data on the postoccasion but had not completed the preoccasion questionnaire, while others had provided data on the preoccasion but not on the postoccasion.

A. Confirmatory factor analysis and criteria for goodness of fit

This study used a CFA approach, which is theory driven [26]. In our case, the *theory* was the factor structure we had hypothesized from the EFA we had computed. In short, we employed CFA to test the ability of our previous factor model (the theory) to fit the *observed* data collected in our current study.

There is continuing debate over which measures are relevant to testing CFA models and the criteria against which any model is to be judged [27]. Normally, only four criteria are reported in testing CFA models: the value and probability of the total Chi-square statistic, the goodness of fit index (GFI), the adjusted goodness of fit index (AGFI), and the root mean square error of approximation (RMSEA). Jain and Chetty [28] published a larger list of criteria together with the values specified for the adequacy of fit of CFA models. They note that the total Chi-Square and its p value, which are normally included in earlier publications, are not included in their criteria given that the overall Chi-square is heavily influenced by the N of cases. That is to say, if N is large, then the Chi-square is also likely to be large and return a p value that is significant, which indicates that the discrepancy between the model and the data being analyzed is not a good fit. Instead, Jain and Chetty divide the total Chi-square by the *degrees of freedom* to obtain a value for CMIN/d.o.f. with the criterion that the result should be less than 5.0 to indicate a reasonable fit of the model to the data. They further suggest using the AGFI for which the criterion is greater than 0.9, the comparative fit index (CFI) again greater than 0.9 and the RMSEA with a value of less than 0.4. These four values are reported for the CFA models below.

III. RESULTS

We computed various CFAs using amos v27, first using the preoccasion data collected in the fall semester of 2021 from the 1117 respondents who had supplied good data. Subsequently, we checked the best factor structure using the postoccasion data supplied by 705 of the 801 respondents who had supplied good data. Subsequently, we report the reliability analyses followed by convergent and discriminant validity analyses of the potential scales.

The goodness of fit statistics for the *original* factor structure reported by the authors in their previous paper on this subject of self-efficacy: the astronomical personal self-efficacy (APSE) scale comprising eight items, and the instrument self-efficacy (ISE) scale comprising five items were found to be inadequate as indicated by Jain and Chetty [28] (AGFI = 0.905; CFI = 0.947; RMSEA = 0.083; CMIN/d.o.f. = 8.765). The value of the CMIN/d.o.f. does not meet the criterion of being <5 and the RMSEA is high while the two fit indices are acceptable. These statistics indicate that the original two-factor model comprising eight APSE and five ISE items is *not* a good fit.

It should also be noted that a number of the items had to be allowed to covary with each other in order to achieve these goodness-of-fit statistics. Indeed, this is not unexpected given that the item responses are purported to be influenced by the latent variables hypothesized in our original model and that they are either related to each other through the ISE factor or the APSE factor. That is to say, the items are likely to covary.

A. Exploration of the new items

In our final analysis, we found that 8 of the 11 new ISE items could be used to very good effect in the ISE scale. To be parsimonious, we also eliminated three of the weaker statements from the APSE factor leaving eight items. That is to say, a total of 16 items were identified in this final model with eight items for each of the two scales. This model yielded much superior goodness of fit statistics (AGFI = 0.978; CFI = 0.966; RMSEA = 0.041; CMIN/d.o.f. = 2.904). We also noted that all of the absolute and incremental fit measures were better for the new model of eight confidence-related items and the eight APSE items, while only being slightly worse for parsimonious fit measures.

We further tested the validity of the two factors by computing a CFA using the postoccasion data from the 705 respondents who had supplied good data. The goodness-of-fit statistics for this postoccasion analysis were acceptably good (AGFI = 0.933, CFI = 0.982; RMSEA = 0.052; CMIN/d.o.f. = 2.917) meeting the criteria for adequacy set by Jain and Chetty [28]. The goodness-of-fit statistics suggest that the two-factor model for efficacy is an adequate fit to the data collected on both the pre- and postoccasions while the validity of the model is indicated by the satisfactory criteria on both occasions of testing.

B. Reliability of the two scales

We next proceeded to test the reliability of the two constructs using SPSS v28 on both the pre- and postoccasions. The results of the statistical output for each of the two potential scales on both occasions are presented in Table I.

These results show that three of the four potential scales each comprising eight items can form scales by simply

TABLE I. Reliability statistics for the individual and combined cohorts.

	New scales with eight items in each	Cronbach's alpha	F test	p value	Tukey index
Preoccasion $N = 1117$	Astronomy personal self-efficacy	0.895	$F(1, 242) = 3.834$	$p = 0.050$	1.128
	Instrumental self-efficacy	0.920	$F(1, 242) = 1.254$	$p = 0.263$	0.964
Postoccasion $N = 705$	Astronomy personal self-efficacy	0.917	$F(1, 704) = 4.414$	$p = 0.036$	1.233
	Instrumental self-efficacy	0.929	$F(1, 704) = 2.668$	$p = 0.102$	1.167

adding together the raw individual-item scores. The postoccasion *astronomy personal self-efficacy* is the exception. The null hypothesis that the raw item scores can be added together to form a scale has to be rejected [$F(1, 704) = 4.414$, $p = 0.036$]. While this statistic reveals that the items are close to being additive, we investigated the properties of the transformed scale created using the item scores raised by the Tukey index (1.233). Our analysis indicated that the transformed item scores could be added together to form a scale (Cronbach's alpha $= 0.918$, $F(1, 704) = 0.095$, $p = 0.779$, Tukey index $= 0.973$). While the goal of such a procedure is to produce scale scores that are normally distributed, our analysis revealed that there was an insufficient difference to merit this transformation and that the raw item scores could be added together to produce a reasonably normally distributed scale score for the APSE factor on the postoccasion of testing. This makes using the APSE scale much easier to use by simply adding the raw-item scores together. We will check this situation as more data become available. The distributions of the scale scores on the pre- and postoccasions for both the personal efficacy and instrumental scales are illustrated in Fig. 1 below.

The skewness and kurtosis statistics for these two scales were found to be adequately "normal" for both the pre- and postoccasion data and could be used in inferential multivariate statistical analysis procedures such as multivariate analysis of variance (MANOVA). Visual inspection of both distributions demonstrates that these scales are amenable to detecting changes from the pre- to the postoccasion.

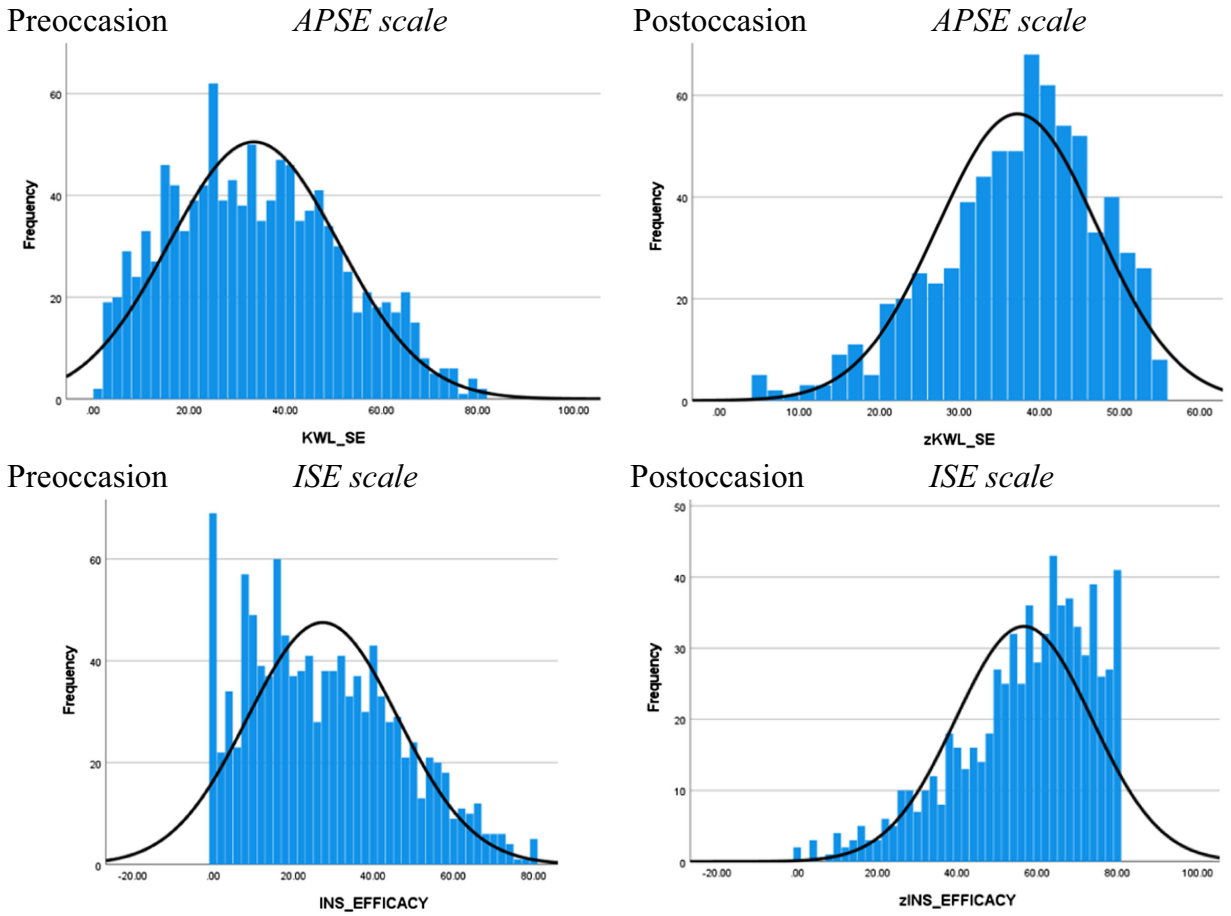

FIG. 1. Frequency distributions of the pre- and postoccasion APSE and ISE scales.

TABLE II. Summary of convergent and discriminant validities of APSE and ISE scales.

Scale	Convergent validity (Pearson ρ, p value)	Discriminant validity (Pearson ρ, p value)
APSE	$\rho = 0.695, p < 0.0001$	$\rho = 0.015, p = 0.741$
ISE	$\rho = 0.710, p < 0.0001$	$\rho = 0.049, p = 0.085$

C. Convergent and discriminant construct validity

Construct validity involves testing to see whether a scale purporting to measure a particular construct is indeed measuring that construct. We have already shown above in the CFA using data drawn from the pre- and postoccasion datasets that the two factors probably possess a reasonable degree of validity. That is to say, the same items load on the same factors in a consistent way on both the pre- and postoccasions of data collection. Convergent- and discriminant-validity analyses are two aspects of demonstrating that a scale possesses construct validity. Moreover, we consulted with two astronomers and four members of the research team to answer the question "Do these items reflect the construct (APSE or ISE) that we are attempting to measure?" The answer to this question was "Yes." We did not consult with respondents to the questionnaire.

Demonstrating convergent validity involves taking two measures that are supposed to be measuring aspects of the same construct and showing that they are related. Conversely, discriminant validity involves taking two measures that are not supposed to be related and demonstrating that they are, in fact, unrelated. Both types of validity are a requirement for indicating that the scale possesses a degree of construct validity. In our wider NSF study, we collected data on a number of associated constructs such as science identity and attitudes toward STEM, as well as constructs that had little or nothing to do with the efficacy scales or STEM fields, for example, career intentions other than in the STEM domain. This allowed us to explore the convergent and discriminant validities of the two scales.

We first investigate the convergent validity of the APSE scale and the discriminant validity of the ISE scale simultaneously using a correlational analysis with responses to the item "I can use the orbital period and semimajor axis of planets to work out the mass of the central body." We hypothesized that there should be a significant correlation for the APSE scale and close to a zero correlation for the ISE scale given that the former relates to astronomical knowledge and the latter to using the instrumentation. The criterion for convergent validity is to have a correlation coefficient (Pearson) greater than 0.5 and preferably greater than 0.7. Table II summarizes the outcomes of the correlational analyses. Table II shows that the Pearson correlation coefficient for the APSE scale is highly significant at $\rho = 0.695 (p < 0.0001)$. The Pearson correlation coefficient for the ISE scale is close to zero ($\rho = 0.049$) and not significant ($p = 0.085$).

To test the convergent validity of the ISE scale, we used the variable "I am confident that I can determine whether a certain object is currently in the sky at night." In this case, the Pearson correlation coefficient for the ISE scale is highly significant ($\rho = 0.710, p < 0.000$) demonstrating a high degree of convergent validity. To test the discriminant validity of the APSE scale, we used the variable of career intention to become a lawyer. This item was part of a set of career intention questions collected at the same time as the efficacy data. Here, we would hypothesize that there should be no correlation between the APSE scale and this career intention. Consistent with expectation, the Pearson correlation coefficient is not significant ($\rho = 0.015$, $p = 0.741$). Table III presents a summary of the convergent and discriminant correlation statistics for the two scales.

These demonstrations of construct validity through analyses of the convergent and discriminant validity

TABLE III. Correlation analysis of APSE and ISE for both occasions of testing.

		Pre-APSE	Pre-ISE	Post-APSE	Post-ISE
Pre-APSE	Pearson correlation	…			
	N	1117			
Pre-ISE	Pearson correlation	0.599	…		
	Significance (one-tailed)	<0.0001			
	N	1117	1117		
Post-APSE	Pearson correlation	0.431	0.256	…	
	Significance (one-tailed)	<0.0001	<0.0001		
	N	521	521	705	
Post-ISE	Pearson correlation	0.276	.279	0.690	…
	Significance (one-tailed)	<0.0001	<0.0001	<0.0001	
	N	521	521	705	705

TABLE IV. Means and standard deviations for the efficacy scales on the pre- and postoccasions of testing.

		Pre-APSE	Post-APSE	Pre-ISE	Post-ISE
N	Valid	1117	705	1117	705
	Missing	184	596	184	596
Mean		33.466	37.224	27.369	56.621
Standard deviation		17.641	9.989	18.749	17.016

components suggest to us that the two scales validly measure what we hypothesized these constructs to be: astronomical personal self-efficacy and instrumental self-efficacy. Moreover, the scales do not appear to suffer greatly from a ceiling effect. The reliability analyses demonstrated that the scales formed by adding both the individual item scores possess a high internal consistency of response with Cronbach's alphas close to 0.9, or greater, depending on the occasion of testing. Finally, the distributions showed that both scales were approximately normal on both the preoccasion and postoccasion of testing. The skewness and kurtosis statistics for both scales on both occasions of testing are within acceptable bounds of normality (i.e., ± 2 and ± 7, respectively) for use in exploring structural equation models. Indeed, the skewness and kurtosis for both scales were less than one.

We computed a correlation analysis of the two scales for both occasions of testing. All correlations shown in Table III are highly significant. The different *N*s in the cells reflect the number of responses on each occasion of testing (1117 and 705) while the $N = 521$ reflects the number of respondents who supplied data on both occasions.

One might infer that the significant correlation between the preoccasion APSE and ISE scales ($\rho = 0.599$) reflects the fact that respondents knew little and hence that their self-efficacy in both APSE and ISE was low. The fact that the correlation rises to 0.690 on the postoccasion of testing may indicate that the respondents' self-efficacy improves in both domains. Indeed, Table IV shows that this is the case where the means and standard deviations of the two self-efficacy scales on both occasions of testing are presented. The ISE scale improves dramatically (29.3 units) while the APSE scale improves by a much lesser amount (3.8 units). It would appear that the respondents are much more confident about their skill level in manipulating the robotic telescopes than they are about their astronomical knowledge.

The use of these efficacy scales to probe covarying changes with the introduction of the new curriculum and its associated labs involving the use of robotic telescopes has the potential to probe other aspects of Astro-101 course design such as course implementation integrity, laboratory interventions, and the depth of treatment instructors decide to employ in their institutions. Another interesting variable will be instructor expertise as indicated by the number of times an instructor has implemented the OPIS! curriculum as well as the extent to which robotic telescopes have been used in the laboratory sessions. The outcomes of such analyses will be reported as the studies are undertaken.

D. The items

Table V shows the final 16 items for the two scales investigated in this paper. These appear to be much more robust than the ones we proposed in our original paper. On three of the four occasions, the raw item scores on the 0–10 Likert scale could simply be added to form a scale score. On the postoccasion for the APSE, we used minor data transformations to render the scale additive and found

TABLE V. Final survey items.

Astronomy personal self-efficacy	
1	I can do astronomy
2	I can explain how the length of the day changes with latitude
3	I can explain how eclipses occur
4	I can explain why stars are different colors and brightnesses
5	I have a good grasp of what objects exist within and around our galaxy
6	The current scientific model of the origin and evolution of the universe is clear to me
7	I can explain why planets move faster and slower in their orbits
8	Most astronomy concepts are easy to learn
Instrumental self-efficacy	
9	I feel confident that I could show someone how to request an image from a remote telescope using an online portal
10	I feel confident that I could learn how to use a remote telescope
11	I feel confident that I could, with relative accuracy, visualize the universe at all different scales
12	I feel confident that I would be able to use parallax measurements of objects within our solar system to measure the astronomical unit
13	I feel confident that I could explain how some variable stars change brightness over time
14	I feel confident that I could distinguish between a globular cluster and galaxy in a telescope
15	I am confident that I can measure the angular diameter of an object using astronomical image processing software
16	I am confident that I can compare my image to a reference image to look for changes

that this made little difference to the scale when compared with one created by simply adding the raw Likert-scale scores together. We caution any potential users to check the reliability of the additivity and Tukey index of the eight APSE items to ensure that the resulting scale score can be obtained by adding the raw Likert-scale item scores together.

IV. DISCUSSION

This paper builds on initial work to establish efficacy scales that deal with astronomy concepts typically encountered in Astro-101 courses in the United States of America and Canada in the context of using robotic telescopes. We were originally aware of the skewed nature of one of these scales and carried out a study to investigate ways to improve it.

In this study, we have established that the two new scales of astronomy personal self-efficacy and instrumental self-efficacy are reliable and with the CFA demonstrating that they are valid constructs as shown by the CFA computations for both the pre- and postoccasion data producing similar outcomes. In particular, the skewness present in the original instrumental self-efficacy scale comprising five items and reported in our original paper [20] has been markedly improved. We have also demonstrated the high construct validity of the scales where convergent and discriminant analyses are computed using other items that are not part of either scale.

The importance of having such reliable instrumentation to probe the effectiveness of undergraduate courses in bringing about changes to students' self-efficacy, and Astro-101 in particular, cannot be understated. Anecdotally and hypothetically, it has long been considered that self-efficacy in such courses was a powerful contributing factor to the success, or otherwise, of a given course. However, previously this has not been probed to any significant extent. With this new robust instrumentation, Astro-101 course instructors can probe the effectiveness of their courses in raising the self-efficacy of their students which, it can be reasonably argued, is a more important factor than improving simple content knowledge at this stage in an undergraduate students' career.

Beyond probing individual course effectiveness, these tools can contribute to larger-scale effectiveness studies. This can go some way to provide an evidential basis for the effectiveness of Astro-101 (and beyond) courses in addressing scientific literacy, scientific identity, and plugging the "leaky pipeline" [29,30] in STEM as well as investigating the nature of the path models one of which is presented in Fig. 1 in our original paper [20].

With the recent proliferation of programs having students access remote and robotic telescopes, along with the growth and expansion of course-based undergraduate research experiences (CURES), research experiences for undergraduates (REUs), and research experiences for teachers (RETs), it is important to evaluate the effectiveness of such programs and the impacts of telescope use. Over the past three years, there has been a large increase in the number of community colleges and universities implementing the Our Place in Space! (OPIS!) curriculum [31] which integrates the use of Skynet Robotic Telescopes in Astro-101 lab courses. The use of the self-efficacy instrument proposed here will help to inform the refinement of the curriculum and supporting materials as well as the different pedagogical approaches employed during the implementation of the OPIS! curriculum. In addition, the Global Sky Partners [32] program has a large and diverse range of programs around the world with learners of all ages using the Las Cumbres Observatory robotic telescopes. There is thus a need for a large-scale understanding of the impact on students using such technology across nations and educational designs. This underscores the importance of having well-validated instruments to probe educational change in the move to improve science literacy and increase the STEM workforce in the United States of America and abroad. Furthermore, a well-validated quantitative instrument can lend support to the qualitative research on the impacts of astronomy research experiences on students and teachers [33–35].

A. Limitations

While the sample size in this study was significant, the fact that not all students in each class completed the survey could lead to bias in the results. A further limitation is that the APSE scale score on the postoccasion needed to be transformed slightly to form a scale although the statistics for additivity were very close to acceptable. This can make the scale a little less straightforward to apply than is desirable. However, we suggest that researchers check the reliabilities of the pre- and postscale items as it might not be necessary to carry out this transformation. The APSE portion of the survey instrument is tailored to Astro-101 courses in the United States. It would be useful to have a similar instrument that was less course specific, and which could be used in a wider range of settings. For example, a survey constructed based on the International Astronomical Union's Astronomy Literacy recommendations [36] could be globally applicable.

V. CONCLUSION

In this paper, the confirmatory factor analysis of the results of two self-efficacy scales was presented. Both possess high validity and reliability suitable for broader analysis. With the recent and continued growth of astronomy programs employing the use of both remote and robotic telescopes, there is increasing opportunity to assess the impacts of such use on students at varying levels in their educational pathways. This arena of student access to scientific tools may help promote increased

science literacy and may also play a role in keeping students in the STEM pipeline, therefore mitigating the problem of the lack of STEM workers in the United States and beyond.

ACKNOWLEDGMENTS

This research was conducted with a National Science Foundation Grant No. 2013300.

[1] A. Bandura, *Self-Efficacy: The Exercise of Control* (Macmillan, London, 1997).
[2] A. Bandura, Self-efficacy: Toward a unifying theory of behavioral change, Psychol. Rev. **84**, 191 (1977).
[3] S. L. Britner and F. Pajares, Sources of science self-efficacy beliefs of middle school students, J. Res. Sci. Teach. **43**, 485 (2006).
[4] R. W. Lent, S. D. Brown, and P. A. Gore Jr., Discriminant and predictive validity of academic self-concept, academic self-efficacy, and mathematics-specific self-efficacy, J. Counsel. Psychol. **44**, 307 (1997).
[5] F. Pajares, Self-efficacy beliefs in academic settings, Rev. Educ. Res. **66**, 543 (1996).
[6] A. Bandura, *Social Foundations of Thought and Action: A Social Cognitive Theory* (Prentice-Hall, Englewood Cliffs, NJ 1986).
[7] S. T. Baier, B. S. Markman, and F. M. Pernice-Duca, Intent to persist in college freshmen: The role of self-efficacy and mentorship, J. Coll. Student Dev. **57**, 614 (2016).
[8] M. Ghee, M. Keels, D. Collins, C. Neal-Spence, and E. Baker, Fine-tuning summer research programs to promote underrepresented students' persistence in the STEM pathway, CBE Life Sci. Educ. **15**, ar28 (2016).
[9] M. J. Graham, J. Frederick, A. Byars-Winston, A.-B. Hunter, and J. Handelsman, Increasing persistence of college students in STEM, Science **341**, 1455 (2013).
[10] V. Sawtelle, E. Brewe, and L. H. Kramer, Exploring the relationship between self-efficacy and retention in introductory physics, J. Res. Sci. Teach. **49**, 1096 (2012).
[11] A. Sithole, E. T. Chiyaka, P. McCarthy, D. M. Mupinga, B. K. Bucklein, and J. Kibirige, Student attraction, persistence and retention in STEM programs: Successes and continuing challenges, Higher Educ. Stud. **7**, 46 (2017), https://eric.ed.gov/?id=EJ1126801.
[12] S. L. Britner, Motivation in high school science students: A comparison of gender differences in life, physical, and earth science classes, J. Res. Sci. Teach. **45**, 955 (2008).
[13] M.-T. Wang and J. Degol, Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields, Dev. Rev. **33**, 304 (2013).
[14] J. E. Stets, P. S. Brenner, P. J. Burke, and R. T. Serpe, The science identity and entering a science occupation, Soc. Sci. Res. **64**, 1 (2017).
[15] E. Gomez and M. Fitzgerald, Robotic telescopes in education. ECU Publications Post 2013. (2017), 10.1080/21672857.2017.1303264.
[16] J. A. Baldwin, D. Ebert-May, and D. J. Burns, The development of a college biology self-efficacy instrument for nonmajors, Sci. Educ. **83**, 397 (1999).
[17] 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).
[18] H. S. Fencl and K. R. Scheel, Pedagogical approaches, contextual variables, and the development of student self-efficacy in undergraduate physics courses, AIP Conf. Proc. **720**, 173 (2004).
[19] K. L. Wester, L. Gonzalez, L. D. Borders, and T. Ackerman, Initial development of the faculty research self-efficacy scale (FaRSES): Evidence of reliability and validity, J. Professoriate **10**, 78 (2019), https://web.s.ebscohost.com/abstract?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=15567680&AN=141332787&h=%2bcjxPVMwvgRSIgnyWyfQD7nWeYH0ZW%2foHda6q1UAopT2Hc9jX7JdLqIryn4R53e8a0LbdzoRWr09XSmkALouLw%3d%3d&crl=c&resultNs=AdminWebAuth&resultLocal=ErrCrlNotAuth&crlhahurl=login.aspx%3fdirect%3dtrue%26profile%3dehost%26scope%3dsite%26authtype%3dcrawler%26jrnl%3d15567680%26AN%3d141332787.
[20] 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).
[21] J. Eccles, Who am I and what am I going to do with my life? Personal and collective identities as motivators of action, Educ. Physcol. **44**, 78 (2009).
[22] R. Sheldrake, Confidence as motivational expressions of interest, utility, and other influences: Exploring under-confidence and over-confidence in science students at secondary school, Int. J. Educ. Res. **76**, 50 (2016).
[23] R. W. Lent, S. D. Brown, and K. C. Larkin, Self-efficacy in the prediction of academic performance and perceived career options, J. Counsel. Psychol. **33**, 265 (1986).
[24] P. R. Pintrich and T. Garcia, Self-regulated learning in college students: Knowledge, strategies, and motivation, in *Student Motivation, Cognition, and Learning* (Routledge, London, 2012), pp. 129–150.
[25] D. Reichart *et al.*, PROMPT: panchromatic robotic optical monitoring and polarimetry telescopes, arXiv:astro-ph/0502429.
[26] R. O. Mueller and R. O. Mueller, Confirmatory factor analysis, in *Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS* (Springer, New York, 1996), pp. 62–128.

[27] D. Shi and A. Maydeu-Olivares, The effect of estimation methods on SEM fit indices, Educ. Psychol. Meas. **80**, 421 (2020).

[28] R. Jain and P. Chetty, Confirmatory factor analysis (CFA) in SEM using SPSS Amos. Knowledge tank; Project Guru (2022), https://www.projectguru.in/confirmatory-factor-analysis-cfa-in-sem-using-spss-amos/.

[29] A. M. Atkin, R. Green, and L. McLaughlin, Patching the leaky pipeline, J. Coll. Sci. Teach. **32**, 102 (2002), https://my.nsta.org/resource/8218/patching-the-leaky-pipeline.

[30] C. Bennett, Beyond the leaky pipeline: Consolidating understanding and incorporating new research: About women's science careers in the UK, Brussels Econ. Rev. **54**, 149 (2011), https://econpapers.repec.org/article/bxrbxrceb/2013_2f108939.htm.

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

[32] Global Sky Partners. (n.d.). Retrieved July 23, 2022, from https://lco.global/education/partners/.

[33] A. Beltzer-Sweeney and S. White, The Double STARS Research Seminar: An analysis of its effects and methodologies, RTSRE Proc. **2**, 1 (2019), https://rtsre.org/index.php/rtsre/article/view/62.

[34] S. R. Buxner, Exploring how research experiences for teachers changes their understandings of the nature of science and scientific inquiry, J. Astron. Earth Sci. Educ. **1**, 53 (2014).

[35] R. Freed, Evaluation of the Astronomy Research Seminar, RTSRE Proc. **2**, 1 (2019), https://www.rtsre.org/index.php/rtsre/article/view/61.

[36] J. Retrê, P. Russo, H. Lee, E. Penteado, S. Salimpour, M. Fitzgerald, J. Ramchandani, M. Pössel, C. Scorza, L. Christensen, E. Arends, S. Pompea, and W. Schrier, Big Ideas in Astronomy Home. (n.d.). (2019) Retrieved July 11, 2022, from https://astro4edu.org/bigideas/.

Correction: A typographical error in the first sentence of the abstract has been fixed.