Journal of Latinos and Education

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hjle20

Reliability Analysis of Psychological Measures Related to STEM Persistence in Undergraduate Students at a Hispanic Serving Institution

Rena Kirkland, Aaron Montoya, Brenna Oakey & Marlene Garcia Araiza

To cite this article: Rena Kirkland, Aaron Montoya, Brenna Oakey & Marlene Garcia Araiza (09 Jul 2024): Reliability Analysis of Psychological Measures Related to STEM Persistence in Undergraduate Students at a Hispanic Serving Institution, Journal of Latinos and Education, DOI: 10.1080/15348431.2024.2367626

To link to this article: https://doi.org/10.1080/15348431.2024.2367626

Published online: 09 Jul 2024.

Submit your article to this journal

Article views: 25

View related articles

View Crossmark data

JOURNAL OF LATINOS AND EDUCATION
https://doi.org/10.1080/15348431.2024.2367626

Reliability Analysis of Psychological Measures Related to STEM Persistence in Undergraduate Students at a Hispanic Serving Institution

Rena Kirkland, Aaron Montoya, Brenna Oakey, and Marlene Garcia Araiza

Psychology, Adams State University

ABSTRACT
Research suggests that psychological factors are related to persistence in science, especially for underrepresented students; however, most psychological instruments have been validated through studies conducted at predominately White institutions. In the current study, we report reliability estimates for measures of science identity, science motivation, and science self-efficacy with a sample of undergraduate college students from a Hispanic Serving Institution (*N* = 309). Internal consistency and test-retest reliability were estimated with Cronbach's alpha and intra-class correlation coefficients, respectively. We report Cronbach's alpha values separately for male (*N* = 152), female (*N* = 152), Hispanic (*N* = 111), and White (*N* = 115) students. We also examined whether there were statistically significant differences in the Cronbach's alpha values between these groups. The results demonstrated good to excellent reliability estimates for internal consistency (α ranged from .89 to .96) and test-retest reliability (ICC ranged from .76 to .80) for all groups. There were no significant differences in Cronbach's alpha values between students identifying as male versus female or between Hispanic and White identifying students. We conclude by urging science education researchers to examine, report, and interpret reliability estimates for their measures for each dataset.

KEYWORDS
Hispanic serving institution; internal consistency; reliability; STEM education; test-retest reliability

Postsecondary institutions have been increasingly attentive to increase underrepresented students in Science, Technology, Engineering, and Mathematics (STEM). One approach to improving racial minority, low-income, and first-generation students' persistence in STEM has been to explore strategies to enhance psychological factors related to student retention in STEM (e.g., Estrada et al., 2016; Jackson et al., 2016; Jordt et al., 2017). Despite the rising interest in measuring psychological factors related to STEM persistence, some fundamental questions remain regarding the psychometric structure of commonly used instruments. Since there are efforts to increase underrepresented students' success in STEM, it is critical that researchers use measures that have been validated with similar demographic characteristics.

The most common approach to measuring psychological variables is the use of multi-item measurement scales, which involves participants responding to several items that are intended to measure an unobservable construct (i.e., latent trait; Hayes & Coutts, 2020). The response choices typically follow a Likert (strongly agree to strongly disagree) or Likert-like format (e.g., never to always), and responses are averaged (or summed) to provide a numerical score for each participant. When developing measurement scales, best practice requires that researchers conduct a psychometric examination of the instrument to provide evidence that the numerical score represents a real individual difference in the underlying construct (Boateng et al., 2018). This process requires examining

CONTACT Rena Kirkland ✉ rkirkland@adams.edu Psychology, Adams State University, 208 Edgemont Blvd., Alamosa, CO 81101, USA

reliability and validity of the measurement scale; however, reliability and validity are not properties of a measurement scale that holds across populations (i.e., psychometric properties are sample dependent). Therefore, it is erroneous for researchers to claim that a scale is reliable and valid. Instead, examining the psychometric properties of an instrument should be an ongoing process (Lindell & Ding, 2013). Reliability estimates should be examined and reported for each dataset since low reliability may contribute to misleading results (i.e., increases risk of false positive and false negative findings or underestimates the true effect). Further, it is important to examine reliability estimates for specific samples since an instrument may demonstrate adequate reliability for some groups but not others. Since most instruments are developed with student samples that are 60% to 80% White or Caucasian (e.g., Glynn et al., 2009), the current study is motivated by a need to examine reliability estimates of the psychological measures that are predictive of success in science in a Hispanic sample of undergraduate students. We examined reliability estimates for three psychological measures by gender (male and female) and for Hispanic and White students separately. Before describing our approach to examining reliability, we provide a brief review of the literature examining the impact of psychological factors on students' success in STEM.

Science identity

Researchers have found that when students identify as scientists, they are more likely to choose optional science experiences in middle and high school (Vincent-Ruz & Schunn, 2018), persist in STEM fields in college (Estrada et al., 2018), and enter a science occupation (Stets et al., 2017). In a large study of 1420 minority STEM students at the undergraduate and graduate levels, science identity predicted science persistence for up to 4 years post-graduation (Estrada et al., 2018). In a more recent study, Chen et al. (2021) found that science identity predicted sense of belonging and grades in a sample of diverse college students. Further, results indicated that grade differences between students with low compared to high science identity were larger for racial-minority students compared to nonminority students. The authors concluded that having strong science identity is particularly important for racial-minority students because it helps them feel a sense of belonging in science classes. It should be noted, however, that although Chen et al. (2021) found that science identity predicted higher grades and sense of belonging for minority students, only a small percent of the samples were Hispanic students; the first study consisted of 66.85% White and 3.58% Hispanic or Latino students, and the second study consisted of 75.52% White and 2.62% Hispanic or Latino students. Underrepresented minority students may be at larger risk of not identifying as scientists; therefore, interventions have aimed at increasing underrepresented students’ science identity to strengthen their STEM commitment (e.g., Chemers et al., 2011).

Despite the recommendations to support science identity in underrepresented minority students, measures of science identity were often developed with primarily White students. For example, the Persistence in the Sciences (PITS) survey includes a subscale that measures science identity (Hanauer et al., 2016). The PITS survey was developed with a sample of 323 undergraduate students of which only 1% identified as Hispanic or Latino. McDonald et al. (2019) developed a 1-item science identity measure using a student sample, 52% of which were from underrepresented groups; however, 48% of the sample were African American, and no information was provided regarding Hispanic students. Pugh et al. (2009) developed a measure of science identity with a sample that was 80% Caucasian and none of the sample was reported to be Hispanic (the rest of the sample was 7% African American and 13% Asian, Pacific Islander, mixed, or chose not to report).

Science motivation

Motivation is a drive to initiate and persist in behavior and has been examined extensively in educational contexts (Howard et al., 2021). While there are dozens of motivation theories in education (see Turabik & Baskan, 2015; Urhahne & Wijnia, 2023 for reviews), one of the most ubiquitous distinctions is between intrinsic and extrinsic motivation. Intrinsic motivation is defined by doing something because it is inherently interesting or satisfying (without regard to external rewards), and research shows it is related to high levels of persistence, positive self-perceptions, and greater engagement (Ryan & Deci, 2000). Extrinsic motivation is demonstrated when individuals engage in an activity to pursue an external reward or outcome (Ryan & Deci, 2000). A meta-analysis examining motivation in educational contexts found a positive relationship between intrinsic motivation and educational outcomes (Howard et al., 2020).

Intrinsic motivation may also have positive correlates in science education. Students who engaged in more intrinsically rewarding science class activities reported higher enjoyment in their educational experience (Druger, 2006). A 2011 study, however, reported a negative relationship between intrinsic motivation and science achievement (Painter, 2011), which opposes prior research. Therefore, more research is needed to examine how motivation affects achievement in STEM education.

Regarding student groups for which measures of motivation have been developed, the Science Motivation Questionnaire II (Glynn et al., 2009) was developed with a sample of undergraduate students, comprised of 82.7% White and 2% Hispanic or Latino students. A follow-up study examining the validity of the measure with science and non-science majors used a sample that included 89.1% White and 3.1% Hispanic or Latino students (Glynn et al., 2011). Since the measure was validated with a sample of primarily White students, examining the reliability of the measure with a sample of Hispanic students would be useful for HSI institutions interested in measuring this variable.

Science-efficacy

Self-efficacy is an individual's belief in their ability to succeed in a specific area (Bandura, 1997) and has been found to be predictive of motivation, behavior, and achievement across many contexts (Schunk & DiBenedetto, 2021). A substantial amount of research shows that self-efficacy is related to academic achievement (e.g., Multon et al., 1991). Bandura and Locke's (2003) meta-analysis showed that students with high self-efficacy persist longer and put forth more effort in their studies. Another meta-analysis examined correlates of college students' academic success including demographic variables, cognitive ability, and psychosocial factors (Richardson et al., 2012). A moderately positive correlation was found between academic self-efficacy and GPA, and a strong positive correlation was found between performance self-efficacy and grades (from 50 correlation coefficients).

Since self-efficacy is domain specific, dozens of efficacy measures have been developed (e.g., sport-efficacy, academic-efficacy, and coaching-efficacy). Based on the predictive value of self-efficacy theory, science efficacy has been a topic of interest in science education studies (e.g., Ackert et al., 2021), and research shows that science self-efficacy is related to achievement outcomes. For instance, in a diverse sample of undergraduate and graduate students, researchers found that science self-efficacy was a predictor of science career commitment (Chemers et al., 2011). In a different study with a sample of primarily Hispanic undergraduate students, self-efficacy predicted success in a physics class (Sawtelle et al., 2012). Chemers et al. (2011) and Sawtelle et al. (2012) samples included 40% and 49% Hispanic or Latino students, respectively, however, it is unclear what the ethnic composition of the students was in the studies that developed the science self-efficacy measures. Stets et al. (2017) developed and validated a 14-item measure of science self-efficacy, but no ethnic information was provided about the sample.

To summarize, strong evidence suggests that psychological processes including science identity, motivation, and self-efficacy are predictive of STEM persistence, especially for underrepresented students; however, more research is needed to understand how to best support students through scalable interventions. While educators further develop interventions that enhance psychological processes in STEM, it is critical that researchers use psychometrically sound instruments. Of note, an instrument that shows strong evidence of reliability with one group of students may not generalize to a different group of students.

Reliability within classical test theory

There are two measurement models that underlie psychometrics: classical test theory and item response theory. Most researchers report reliability coefficients that are based on classical test theory (CTT; Doval et al., 2023). The basic assumption of CTT is that the variance in any observed score (X) is based on the true differences of the trait being measured (i.e., true score, T) and measurement error (E). Thus, an observed score is represented by the following model, $X = T + E$ (Raykov & Marcoulides, 2016). As can be discerned by the theoretical formula, to assess the precision of observed scores, it is essential to examine the amount of error present in a measurement.

Based on CTT, reliability is the inverse of the proportion of measurement error and is related to the proportion of true score. Therefore, reliability provides an estimate of the proportion of total variance that is due to true variance versus measurement error variance. It is important for researchers to report reliability estimates each time they use a measurement scale because reliability quantifies the amount of measurement error for a specific sample (Streiner, 2003), and the lower the reliability, the more error which attenuates effect sizes (Fan, 2003; Matheson, 2019). That is, low reliability underestimates true relationships between variables. Cole and Preacher (2014) described serious problems to path analysis due to low reliability and uncorrected measurement error. In some cases, especially for small sample sizes, measurement error can lead to greater variation in estimated effect sizes, and thus, low reliability can overestimate a true effect by chance (Loken & Gelman, 2017). In any case, when authors do not report reliability, it is impossible to estimate how much measurement error is present, and any further analysis may be biased. Further, adequate reliability is necessary, although not sufficient, to provide evidence of validity. Since some STEM education researchers do not report the reliability of their data (e.g., Aagaard & Hauer, 2003; Bogner, 2023; Ma & Xiao, 2021; McCartney et al., 2022; Salinitri, 2005), it is impossible to discern whether their results are valid.

There are many different reliability estimates, each of which examines a different aspect of measurement error (Cook & Beckman, 2006). The most reported for a single administration of a measurement scale is Cronbach's alpha (Cronbach, 1951), which is a measure of internal consistency and measures the relationship between multiple items on a measure at one measurement time (see Taber, 2018 for extensive review of alpha in science education). Cronbach's alpha is an appropriate estimate of internal consistency when there is evidence of unidimensionality (Doval et al., 2023; Raykov & Marcoulides, 2019). Based on CTT, unidimensionality indicates that all items in a measure are indicative of the same underlying variable and can be examined through factor analysis procedures (DeVellis, 2006).

In contrast to Cronbach's alpha, which measures internal consistency, test-retest reliability examines the consistency of responses overtime (Polit, 2014). That is, test-retest reliability examines the stability of participant responses across two administrations of the same measure. Test-retest reliability is reported much less frequently than Cronbach's alpha due to the challenge of administering the measure two times with the same participants. Retest reliability was not reported in the articles that described the development of the measures, and our literature review did not reveal any authors reporting retest reliability on the science identity or science self-efficacy scale. Wardhany et al. (2018) reported retest reliability for an Indonesian version of the Science Motivation Questionnaire-II (with intraclass correlation coefficients of .82 and .88 for intrinsic and career motivation respectively), and Dong et al. (2020) reported retest reliability for a Chinese version of the questionnaire (with intraclass correlation coefficients of .54 and .52 for intrinsic and career motivation respectively). We did not find any reports of retest reliability with the English version of the scale. Retest reliability is critical to examine when researchers are interested in testing interventions because if a measure has low retest reliability, then the true effects of the interventions are masked (Aldridge et al., 2017). Further, retest reliability is critical to examine for replication research (Leppink & Pérez-Fuster, 2017).

Most researchers use retest intervals of two to four weeks; however, Watson (2004) emphasized that the retest interval should be theoretically meaningful. When examining semester-long interventions for college students, the meaningful retest interval is about 16 weeks. Due to the practical challenges of setting up the same testing protocol when examining retest reliability over long periods of time, few researchers report test-retest coefficients that correspond to meaningful time intervals. Besides being the first study to report retest reliability for measures of science identity, intrinsic and career motivation for science, and science self-efficacy, one of the contributions of our study is that we examined retest reliability over a 16-week interval, which aligns with the time interval for semester-long interventions.

The purpose of our study was to examine the internal consistency and test-retest reliability of three psychological measures related to STEM retention and success. We expected that the measures would show adequate internal consistency as measured by Cronbach's alpha for the full sample as well as for males, females, Hispanics, and White subsamples. Further, we expected that the measures would show adequate stability over time as measured by the intraclass correlation coefficient (ICC). If reliability is not adequate, then researchers might need to develop and test measures specifically for the demographics of the students.

Method

Participants

The participants were college students from a small university in the Southwest region of the United States taking STEM courses and part of a larger study examining the efficacy of Course-based Undergraduate Research Experiences (CUREs). The university has on average 242 STEM students per year, of which 44% are female, 56% male, 47% Pell eligible, 30% low-income, 33% first-generation, 48% White, and 38% Hispanic. Participants completed the survey twice, once at the beginning and once at the end of the semester. We collected data across four years in several STEM classes, and since many students take the survey more than once, we excluded duplicate responses in all analyses. After removing duplicate responses, the baseline sample included 309 participants (mean age = 19.87, *SD* = 3.6). At follow up, the sample included 247 participants (mean age = 20.15, *SD* = 3.4). To compare reliability estimates for Hispanic and White participants separately, we removed participants who selected multiple racial identities. See Table 1 for demographic data including gender and Hispanic identity, year in college, and STEM versus non-STEM majors in the sample.

Table 1. Sociodemographic characteristics of participants.

	Baseline		Follow-up	
Characteristic	*N*	%	*N*	%
Total	309	-	247	-
Gender				
Male	152	52.1	107	43.3
Female	155	47.4	138	55.9
Other	2	00.4	2	00.8
Hispanic Identity[a]				
Hispanic	107	42.0	92	50.0
White	148	58.0	92	50.0
Year in college[b]				
First year	118	52.0	81	51.3
Sophomore	44	19.4	25	15.8
Junior	35	15.4	28	17.8
Senior	13	5.7	13	8.2
Other	17	7.5	11	7.0
Major				
STEM major	206	66.7	177	71.7
NonSTEM	103	33.3	70	28.3

Note. [a]We removed participants who selected more than one racial identity, which reduced the separate sample sizes of Hispanic and White students.
[b]Some of the surveys did not ask participants to provide year in college, so there are some missing values in these data points.

Table 2. Description of instrument development of the measures.

Name of Scale Subscales used (if any)	Authors and Year	Description
Science Identity	Hanauer et al. (2016)	● 5 items ● Validated with 323 undergraduate biology students in 9 different biology classes from a mid-sized university in western Pennsylvania ● 52% White; 1% Hispanic or Latino ● α = .87 ● No test-retest reliability reported
Science Motivation Questionnaire 2 subscales: intrinsic motivation and career motivation	Glynn et al. (2011)	● 10 items (5 items for each subscale) ● Likert scale 1–5 never – always ● Validated with 367 undergraduate science majors and 313 non-science majors from large-sized university in southern United States ● 89.1% White; 3.1% Hispanic or Latino ● Intrinsic motivation α = .89 ● Career motivation α = .92 ● No test-retest reliability reported
Science Self-Efficacy	Stets et al. (2017)	● 14 items ● Likert scale 1–5 not at all confident–very confident ● Validated with 1429 undergraduate students from 25 different institutions (including private and public; small, medium, and large institutions) ● Breakdown of racial identities of sample was not reported ● ω = .97 ● No test-retest reliability reported

α = Cronbach's alpha (Cronbach, 1951); ω = Omega (Hayes & Coutts, 2020).

Materials

For an overview of the instruments and reliability estimates reported in the original published articles see Table 2.

Science identity

We used a five-item scale to measure the degree students identify as scientists (Hanauer et al., 2016). Students responded to items on a 5-point Likert scale (strongly disagree, disagree, neither agree nor disagree, agree, strongly agree).

Science motivation questionnaire

The full measure includes five subscales including intrinsic motivation, self-efficacy and assessment anxiety, self-determination, career motivation, and grade motivation (Glynn et al., 2011). We used the intrinsic and career motivation subscales, which include five items for each subscale with five response choices (never; rarely; sometimes; often; always).

Science self-efficacy

To measure self-efficacy in the science domain, we used Stets et al. (2017) Science Self-Efficacy Scale. Students respond to 14 items on a 5-point Likert-like scale (not at all confident, somewhat confident, confident, mostly confident, absolutely confident).

Procedures

After obtaining Institution Review Board approval, the surveys were administered during classes through Qualtrics (an online survey collection tool). The consent process included a verbal description of the purpose, length, and type of questions included in the survey. Participants also read a consent form and had to select "agree" to continue with the survey. For students who were not in class, the second author emailed the students and encouraged them to complete the survey. The response rate averaged 80% over the four years of data collection.

Statistical analysis

All analyses were conducted with SPSS version 27. As described above, participants completed the surveys at the beginning and end of the semester. To examine whether each instrument was unidimensional, we ran exploratory factor analysis (EFA) on the baseline data. We first examined Kaiser–Meyer–Olkin (KMO) and Bartlett's Test of Sphericity to check that factor analysis was appropriate for the data. KMO values over .8 indicate the data are adequate for factor analysis (Guttman, 1954). Bartlett's test of Sphericity produces a Chi-square where significant Chi-square values suggest sampling is adequate (Tabachnick & Fidell, 2001). We used principal axis factoring for the extraction method. To determine the number of factors, we examined the scree plot and factor matrix (Costello & Osborne, 2005).

For the reliability analyses, we used the baseline data to examine internal consistency. We ran Cronbach's alpha (Cronbach, 1951) to estimate internal consistency for the full sample and by gender (males vs females) and Hispanic identity (Hispanic vs White). To test whether the alpha values were significantly different by groups we used cocron, which is a platform-independent R package (Diedenhofen & Musch, 2016). We applied Bonferroni correction to adjust for running eight comparisons; thus, the critical p-value was .00625 (.05/8 = .00625, Bland & Altman, 1995). To estimate test-retest reliability, we used participants' baseline and follow-up assessments, and we examined ICC using a two-way mixed model (participant effects are random and measure effects are fixed), and a consistency type where the between-measure variance is excluded from the denominator variance (Koo & Li, 2016).

Results

Preliminary analysis

For the EFA, KMO statistics ranged from .85 to .96, and Bartlett's Test of Sphericity were all significant suggesting that factor analysis was appropriate for the data. The scree plots showed evidence of 1 factor for each scale or subscale (science identity, intrinsic motivation, career motivation, and science self-efficacy). For the science identity measure, the factor loadings ranged from .78 to .86 providing evidence of unidimensionality. For intrinsic and career motivation, the factor loadings ranged from .77 to .91, and .82 to .94, respectively. For science self-efficacy, the factor loadings ranged from .67 to .87. Scree plots showed clear evidence of one factor for each of the four measures.

Main analysis

For descriptive statistics see Table 3. Reliability coefficients, including test-retest reliability as measured by ICC and internal consistency as measured by Cronbach's alpha, can be found in Table 4. Results examining whether Cronbach's alpha was significantly different between males and females and between Hispanic and White participants are in Table 5.

Discussion

Recent attention to inequities in STEM has generated innovative strategies to increase the success of students underrepresented in STEM fields. One approach has been to enhance student experiences in STEM by addressing psychological factors such as science identity, motivation, and self-efficacy, which have been shown to be related to persistence in STEM fields. More work has to be done to identify and implement interventions that enhance psychological factors for specific populations. This

Table 3. Descriptive statistics for full sample and by gender and Hispanic identity.

	Baseline			Follow-up		
Name of Scale/Subscale	*N*	*M*	*SD*	*N*	*M*	*SD*
Science Identity[a]	**266**	**3.31**	**.96**	**158**	**3.43**	**1.02**
Males	132	3.29	.99	68	3.38	.99
Females	131	3.33	.93	87	3.44	1.05
Hispanic	92	3.36	.90	60	3.31	1.04
White	126	3.22	1.00	75	3.47	.96
Intrinsic Motivation	**303**	**3.98**	**.86**	**246**	**4.02**	**.83**
Males	149	3.89	.90	107	4.04	.78
Females	151	4.07	.81	136	4.06	.87
Hispanic	105	4.02	.84	92	4.04	.82
White	148	3.87	.91	92	4.00	.89
Career Motivation	**304**	**4.29**	**.86**	**243**	**4.24**	**.83**
Males	149	4.14	.94	107	4.17	.74
Females	152	4.43	.76	134	4.27	.90
Hispanic	107	4.31	.88	89	4.23	.82
White	147	4.20	.92	91	4.19	.89
Science Self Efficacy	**284**	**2.96**	**.91**	**163**	**3.35**	**.93**
Males	137	3.09	.87	71	3.47	.90
Females	144	2.85	.92	90	3.25	.95
Hispanic	100	2.80	.82	59	3.31	.89
White	134	3.02	.98	75	3.41	.89

[a]Science identity has a smaller sample size since data was not collected one semester for this variable.

Table 4. Reliability estimates for full sample and by gender and Hispanic identity.

Name of Scale Subscales (if used)	Number of items	Test-retest (*N*)	Internal consistency (*N*)	95% CI LL	 UL
Science Identity	5	**.76 (128)**	**.91 (283)**	**.89**	**.93**
Males			.92 (132)	.89	.94
Females			.91 (131)	.88	.93
Hispanic			.89 (92)	.85	.92
White			.91 (98)	.88	.94
Science Motivation					
Intrinsic Motivation	5	**.77 (143)**	**.93 (303)**	**.91**	**.94**
Males			.93 (149)	.91	.95
Females			.92 (151)	.89	.94
Hispanic			.92 (105)	.89	.94
White			.93 (115)	.91	.95
Career Motivation	5	**.76 (140)**	**.94 (304)**	**.93**	**.95**
Males			.95 (149)	.93	.96
Females			.93 (152)	.91	.95
Hispanic			.94 (107)	.92	.96
White			.94 (114)	.92	.96
Science Self Efficacy	14	**.80 (127)**	**.96 (284)**	**.95**	**.97**
Males			.95 (137)	.95	.96
Females			.96 (144)	.95	.97
Hispanic			.94 (100)	.93	.96
White			.96 (104)	.95	.97

CI = confidence interval; LL = lower limit; UL = upper limit. *Test-retest examined with intraclass correlation coefficient; internal consistency examined with Cronbach's alpha. Internal consistency and corresponding CI values are reported for baseline assessment.

requires that researchers select measurement scales that are appropriate for the populations served. It is critical to examine the psychometric properties of instruments for different groups since reliability and validity is not invariant across samples (Lindell & Ding, 2013). Thus, the purpose of the current study was to examine the reliability of three psychological measures related to success in STEM in Hispanic and White college students separately. Specifically, we examined the internal consistency (Cronbach's alpha) and test-retest (ICC) of the Science Identity Scale (Hanauer et al., 2016), the

Table 5. Results of cocron analyses examining cronbach alpha group differences.

Name of Scale Subscales (if used)	Number of items	*N/N*	X^2(df)	*p*-value
Science Identity	5			
Males vs females		132/131	.34(1)	.563
Hispanic vs White		92/98	.97(1)	.323
Science Motivation				
Intrinsic Motivation	5			
Males vs females		149/151	.72(1)	.397
Hispanic vs White		105/115	.72(1)	.397
Career Motivation	5			
Males vs females		149/152	1.54(1)	.215
Hispanic vs White		107/114	.02(1)	.883
Science Self Efficacy	14			
Males vs females		137/144	1.81(1)	.178
Hispanic vs White		100/104	4.79(1)	.028*

N/N= sample size for each group, respectively. X^2 = Chi-square results of cocron analyses examining Cronbach alpha group differences; cocron is an independent R-package program (Diedenhofen & Musch, 2016).

intrinsic and career motivation subscales from the Science Motivation Questionnaire (Glynn et al., 2009, 2011), and the Science Self-Efficacy Scale (Stets et al., 2017). We also examined whether there were significant differences in alpha coefficients between males and females and between Hispanic and White participants.

Cronbach's alpha values are typically interpreted as follows: values between .6 and .69 are interpreted as questionable; values between .7 and .79 are interpreted as adequate; values between .8 and .89 are considered good; and values higher than .9 are excellent (Cronbach, 1951; Nunnally, 1978). For the full sample, alpha values ranged from .91 to .96 indicating excellent reliability. We also examined whether reliability estimates differed based on gender and Hispanic identities. Results demonstrated reliability estimates ranging from good to excellent for males, females, Hispanic and White subsamples. There were no significant differences between the alpha values for male versus female or for Hispanic versus White students.

ICC values are interpreted more leniently compared to Cronbach's alpha since changes from pre- to post-test could be caused by true changes in the underlying construct (Cicchetti, 1994). In other words, changes from pre- to post-test may not only be due to measurement error but may be caused by actual changes in individual differences in the construct being measured. Participant scores may also change due to differences in the testing environment as it is difficult to set up identical testing procedures from the first to second assessment, and the longer the interval, the more difficult it is to exactly replicate procedures. Further, participant scores may change due to state differences (e.g., mood changes, fatigue) between pre- and post-test (Watson, 2004). Cicchetti (1994) suggested that ICC values from .4 to .59 are fair, values from .60 to .74 are good, and values above .75 are excellent. For our sample, ICC estimates ranged from .76 to .80 indicating that the measures demonstrated excellent test-retest reliability.

Recommendations and implications

Though it is important for science education researchers to scrutinize the quality of instrument development, it is erroneous to infer that a measure is reliable based solely on previously reported reliability estimates. Instead, researchers must examine, report, and interpret reliability for each dataset. This is essential because low reliability attenuates true effects; thus, researchers should examine reliability before proceeding with analyses.

Researchers should not mistake high reliability as an indication of unidimensionality. Rather, unidimensionality is a necessary assumption for Cronbach's alpha, and therefore should be examined through factor analysis methods prior to running alpha (Doval et al., 2023; Raykov & Marcoulides, 2019). Researchers should clearly explain which reliability estimate they examined and provide guidelines for interpreting values. In an extensive review of reliability estimates in four leading science education journals published across a year, Taber (2018) reported that Cronbach's alpha values were reported most frequently; however, some authors reported a reliability coefficient without explaining which estimate they examined, and some authors did not describe how to interpret the values (or interpreted the values inappropriately). Science education researchers should clearly report which reliability estimate they examined, report the precise values for their data, and provide interpretations of the values in relation to the purpose of the measures.

It is important to note that the guidelines to interpret reliability estimates are rules of thumb. Researchers should consider the context and use of the measures when determining whether their data shows adequate reliability. For instance, when important decisions are being made, such as in educational and clinical contexts, it has been recommended that Cronbach's alpha values should be no lower than .9 (Matheson, 2019). Test-retest values are typically interpreted more leniently; however, guidelines vary greatly depending on the context. For instance, for retest reliability of test scores in educational contexts values > .8 are considered necessary (Norcini, 1999). In clinical contexts, authors describe that values between .5 and .74 are poor-to-moderate, values between .75 and .9 are good, and values > .95 are excellent (Portney & Watkins, 2015). Finally, the interval between testing is also critical to consider when interpreting retest reliability. The longer time-interval, the more likely the true score and other situational factors will change, which will lead to smaller coefficients (Duff, 2012). Thus, the time-interval should be considered when interpreting the magnitude of test-retest reliability.

Although it is recommended to use theoretically meaningful test-retest intervals (Watson, 2004), few researchers examine retest reliability with an interval the length of a semester. We recommend that more research be conducted examining retest reliability of psychological measures used in science education, and that researchers use time intervals that align with the timeframe that is typical between pre- and post-assessment. In the current study, we report retest reliability estimates for the measures across a 16-week interval, which corresponds to the length of semester-long interventions. The ICC coefficients ranged from .76 to .80; therefore, we found evidence that the measures show stability across the timeframe of a semester-long class.

In sum, it is paramount for researchers to examine the reliability of their data prior to running statistical analysis. In CTT, upon which the coefficient alpha rests, when reliability is low at a minimum the estimated effect is less precise, and low reliability increases the risk of making type I and type II errors (Matheson, 2019). Even when statistically significant results are found, high measurement error will lead to an underestimate of the effect size. It is important to note that, in addition to increased chance of false positive findings, low reliability also leads to more variable measurement, and thus can sometimes overestimate effects based on chance. Therefore, it is critical that researchers report reliability estimates of their data and provide readers with an explanation of how to interpret the estimated coefficient(s). Though reliability is sample specific and should be examined and reported with every new sample, our findings provide evidence that the Science Identity Scale (Hanauer et al., 2016), the intrinsic and career motivation subscales from the Science Motivation Questionnaire (Glynn et al., 2009, 2011), and Science Self-Efficacy (Stets et al., 2017) demonstrated good to excellent internal consistency and test-retest reliability for a sample of Hispanic and White undergraduate students from a HSI institution.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This project was supported by a National Science Foundation (NSF), Improvement of Undergraduate STEM Education (IUSE), Hispanic Serving Institution (HSI) grant [number 1928405] aimed at increasing underrepresented students' success in STEM. We gratefully acknowledge Dr Steven Pulos for his comments and review of the paper.

References

Aagaard, E. M., & Hauer, K. E. (2003). A cross-sectional descriptive study of mentoring relationships formed by medical students. *Journal of General Internal Medicine*, *18*(4), 298–3022. https://doi.org/10.1046/j.1525-1497.2003.20334.x

Ackert, E., Snidal, M., & Crosnoe, R. (2021). The development of science, technology, engineering, and mathematics (STEM) efficacy and identity among Mexican-origin youth across Latino/a destinations. *Developmental Psychology*, *57*(11), 1910–1925. https://doi.org/10.1037/dev0001251

Aldridge, V. K., Dovey, T. M., & Wade, A. (2017). Assessing test-retest reliability of psychological measures: Persistent methodological problems. *European Psychologist*, *22*(4), 207–218. https://doi.org/10.1027/1016-9040/a000298

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W H Freeman. https://psycnet.apa.org/record/1997-08589-000

Bandura, A., & Locke, E. A. (2003). Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology*, *88*(1), 87–99. https://doi.org/10.1037/0021-9010.88.1.87

Bland, J. M., & Altman, D. G. (1995). Statistics notes: Multiple significance tests: The Bonferroni method. *British Medical Journal*, *310*(6973), 170. https://doi.org/10.1136/bmj.310.6973.170

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 1–18. https://doi.org/10.3389/fpubh.2018.00149

Bogner, F. X. (2023). Open schooling matters: Student effects in science motivation, intrinsic motivation and state emotions. *Journal of Higher Education Theory & Practice*, *23*(2). https://articlegateway.com/index.php/JHETP/article/view/5813

Chemers, M. M., Zurbriggen, E. L., Syed, M., Goza, B. K., & Bearman, S. (2011). The role of efficacy and identity in science career commitment among underrepresented minority students. *Journal of Social Issues*, *67*(3), 469–491. https://doi.org/10.1111/j.1540-4560.2011.01710.x

Chen, S., Binning, K. R., Manke, K. J., Brady, S. T., McGreevy, E. M., Betancur, L., Limeri, L. B., & Kaufmann, N. (2021). Am I a science person? A strong science identity bolsters minority students' sense of belonging and performance in college. *Personality and Social Psychology Bulletin*, *47*(4), 593–606. https://doi.org/10.1177/0146167220936480

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. https://doi.org/10.1037/a0033805

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, *119*(2), 166–e7. https://doi.org/10.1016/j.amjmed.2005.10.036

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(10), Article 7. https://doi.org/10.7275/jyj1-4868

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*(11), S50–S59. http://www.jstor.org/stable/41219505

Diedenhofen, B., & Musch, J. (2016). Cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science*, *11*(1), 51–60. https://rb.gy/0fy9qg

Dong, Z., Li, M., Minstrell, J., & Cui, Y. (2020). Psychometric properties of science motivation questionnaire II-Chinese version in two waves of longitudinal data. *Psychology in the Schools*, *57*(8), 1240–1256. https://doi.org/10.1002/pits.22370

Doval, E., Viladrich, C., & Angulo-Brunet, A. (2023). Coefficient alpha: The resistance of a classic. *Psicothema*, *35*(1), 5–20. https://doi.org/10.7334/psicothema2022.321

Druger, M. (2006). Experiential learning in a large introductory biology course. In J. J. Mintzes & W. H. Leonard (Eds.), *Handbook of college science teaching* (pp. 37–43). National Science Teachers Association Press.

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*(3), 248–261. https://doi.org/10.1093/arclin/acr120

Estrada, M., Burnett, M., Campbell, A. G., Campbell, P. B., Denetclaw, W. F., Gutiérrez, C. G., Hurtado, S., John, G. H., Matsui, J., McGee, R., Okpodu, C. M., Robinson, T. J., Summers, M. F., Werner-Washburne, M., Zavala, M. E., & Marsteller, P. (2016). Improving underrepresented minority student persistence in STEM. *CBE - Life Sciences Education*, *15*(3), es5. https://doi.org/10.1187/cbe.16-01-0038

Estrada, M., Hernandez, P. R., Schultz, P. W., & Herrera, J. (2018). A longitudinal study of how quality mentorship and research experience integrate underrepresented minorities into STEM careers. *CBE - Life Sciences Education*, *17*(1), Article 9. ar9. https://doi.org/10.1187/cbe.17-04-0066

Fan, X. (2003). Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational and Psychological Measurement*, *63*(6), 915–930. https://doi.org/10.1177/0013164403251319

Glynn, S. M., Brickman, P., Armstrong, N., & Taasoobshirazi, G. (2011). Science motivation questionnaire II: Validation with science majors and nonscience majors. *Journal of Research in Science Teaching*, *48*(10), 1159–1176. https://doi.org/10.1002/tea.20442

Glynn, S. M., Brickman, P., & Taasoobshirazi, G. (2009). Science motivation questionnaire: Construct validation with nonscience majors. *Journal of Research in Science Teaching*, *46*(2), 127–146. https://doi.org/10.1002/tea.20267

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149–161. https://doi.org/10.1007/BF02289162

Hanauer, D., Graham, M. J., & Hatful, G. F. (2016). A measure of college student persistence in the sciences (PITS). *CBE - Life Sciences Education*, *15*(4), Article 54. ar54. https://doi.org/10.1187/cbe.15-09-0185

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. *But . . . Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Howard, J. L., Bureau, J. S., Guay, F., Chong, J. X. Y., & Ryan, R. M. (2021). Student motivation and associated outcomes: A meta-analysis from self-determination theory. *Perspectives on Psychological Science*, *16*(6), 1–36. https://doi.org/10.1177/1745691620966789

Howard, J. L., Chong, J. X. Y., & Bureau, J. S. (2020). The tripartite model of intrinsic motivation in education: A 30-year retrospective and meta-analysis. *Journal of Personality*, *88*(6), 1268–1285. https://doi.org/10.1111/jopy.12570

Jackson, M. C., Galvez, G., Landa, I., Buonara, P., Thoman, D. B., & Gibbs, K. (2016). Science that matters: The importance of a cultural connection in underrepresented students' science pursuit. *CBE - Life Sciences Education*, *15*(3), Article 42. https://doi.org/10.1187/cbe.16-01-0067

Jordt, H., Eddy, S. L., Brazil, R., Lau, I., Mann, C., Brownell, S. E., King, K., Freeman, S., & Schinske, J. (2017). Values affirmation intervention reduces achievement gap between underrepresented minority and white students in introductory biology classes. *CBE - Life Sciences Education*, *16*(3), Article 41. https://doi.org/10.1187/cbe.16-12-0351

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Leppink, J., & Pérez-Fuster, P. (2017). We need more replication research – a case for test-retest reliability. *Perspectives of Medical Education*, *6*(3), 158–164. https://doi.org/10.1007/s40037-017-0347-z

Lindell, R., & Ding, L. (2013). Establishing reliability and validity: An ongoing process. *American Institute of Physics Conference Proceedings*, *1513*(1), 27–29. https://doi.org/10.1063/1.4789643

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ*. Article e6918. https://doi.org/10.7717/peerj.6918

Ma, Y., & Xiao, S. (2021). Math and science identity change and paths into and out of STEM: Gender and racial disparities. *Socius: Sociological Research for a Dynamic World*, *7*, 7. https://doi.org/10.1177/23780231211001978

McCartney, M., Roddy, A. B., Geiger, J., Piland, N. C., Ribeiro, M. M., & Lainoff, A. (2022). Seeing yourself as a scientist: Increasing science identity using professional development modules designed for undergraduate students. *Journal of Microbiology & Biology Education*, *23*(1), Article e00346–21. https://doi.org/10.1128/jmbe.00346-21

McDonald, M. M., Zeigler-Hill, V., Vrabel, J. K., & Escobar, M. (2019). A single-item measure for assessing STEM identity. *Frontiers in Education*, *4*, 78. https://doi.org/10.3389/feduc.2019.00078

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology*, *38*(1), 30–38. https://doi.org/10.1037/0022-0167.38.1.30

Norcini, J. J. (1999). Standards and reliability in evaluation: When rules of thumb don't apply. *Academic Medicine*, *74*(10), 1088–1090. https://doi.org/10.1097/00001888-199910000-00010

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Painter, J. (2011). *Autonomy, competence, and intrinsic motivation in science education: A self-determination theory perspective* [Doctoral dissertation]. University of North Carolina Chapel Hill. Carolina Digital Repository. https://doi.org/10.17615/f3yw-8912

Polit, D. F. (2014). Getting serious about test-retest reliability: A critique of retest research and some recommendations. *Quality of Life Research*, *23*(6), 1713–1720. https://doi.org/10.1007/s11136-014-0632-9

Portney, L. G., & Watkins, M. P. (2015). *Foundations of clinical research: Applications to practice* (3rd ed.). F.A. Davis Company.

Pugh, K. J., Linnenbrink-Garcia, L., Koskey, K. L. K., Stewart, V. C., & Manzey, C. (2009). Motivation, learning, and transformative experience: A study of deep engagement in science. *Science Education*, *94*(1), 1–28. https://doi.org/10.1002/sce.20344

Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement*, *76*(2), 325–338. https://doi.org/10.1177/0013164415576958

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement*, *79*(1), 200–210. https://doi.org/10.1177/0013164417725127

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353–387. https://doi.org/10.1037/a0026838

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Salinitri, G. (2005). The effects of formal mentoring on the retention rates of first-year, low achieving students. *Canadian Journal of Education*, *28*(4), 853–873. https://www.jstor.org/stable/4126458

Sawtelle, V., Brewe, E., & Kramer, L. H. (2012). Exploring the relationship between self-efficacy and retention in introductory physics. *Journal of Research in Science Teaching*, *49*(9), 1096–1121. https://doi.org/10.1002/tea.21050

Schunk, D. H., & DiBenedetto, M. K. (2021). Self-efficacy and human motivation. *Advances in Motivation Science*, *8*, 153–179. https://doi.org/10.1016/bs.adms.2020.10.001

Stets, J. E., Brenner, P. S., Burke, P. J., & Serpe, R. T. (2017). The science identity and entering a science occupation. *Social Science Research*, *64*, 1–14. https://doi.org/10.1016/j.ssresearch.2016.10.016

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Allyn & Bacon.

Taber, K. S. (2018). The use of cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Turabik, T., & Baskan, G. A. (2015). The importance of motivation theories in terms of education systems. *Procedia - Social & Behavioral Sciences*, *186*, 1055–1063. https://doi.org/10.1016/j.sbspro.2015.04.006

Urhahne, D., & Wijnia, L. (2023). Theories of motivation in education: An integrative framework. *Educational Psychology Review*, *35*, Article 45. https://doi.org/10.1007/s10648-023-09767-9

Vincent-Ruz, P., & Schunn, C. D. (2018). The nature of science identity and its role as the driver of student choices. *International Journal of STEM Education*, *5*, Article 48. https://doi.org/10.1186/s40594-018-0140-5

Wardhany, I. I., Subita, G. P., & Maharani, D. A. (2018). Cross-cultural adaptation and psychometric properties of the science motivation questionnaire-II: Indonesian version. *Pesquisa Brasileira em Odontopediatria e Clinica Integrada*, *18*(1), Article e4294. https://doi.org/10.4034/PBOCI.2018.181.111

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*(4), 319–350. https://doi.org/10.1016/j.jrp.2004.03.001