Maric *et al.*
International Journal of STEM Education (2023) 10:39
https://doi.org/10.1186/s40594-023-00430-x

International Journal of STEM Education

REVIEW

Open Access

Measurement in STEM education research: a systematic literature review of trends in the psychometric evidence of scales

Danka Maric[1*], Grant A. Fore[1], Samuel Cornelius Nyarko[1] and Pratibha Varma-Nelson[1,2]

Abstract

Background The objective of this systematic review is to identify characteristics, trends, and gaps in measurement in Science, Technology, Engineering, and Mathematics (STEM) education research.

Methods We searched across several peer-reviewed sources, including a book, similar systematic reviews, conference proceedings, one online repository, and four databases that index the major STEM education research journals. We included empirical studies that reported on psychometric development of scales developed on college/university students for the context of post-secondary STEM education in the US. We excluded studies examining scales that ask about specific content knowledge and contain less than three items. Results were synthesized using descriptive statistics.

Results Our final sample included the total number of $N = 82$ scales across $N = 72$ studies. Participants in the sampled studies were majority female and White, most scales were developed in an unspecified STEM/science and engineering context, and the most frequently measured construct was attitudes. Internal structure validity emerged as the most prominent validity evidence, with exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) being the most common. Reliability evidence was dominated by internal consistency evidence in the form of Cronbach's alpha, with other forms being scarcely reported, if at all.

Discussion Limitations include only focusing on scales developed in the United States and in post-secondary contexts, limiting the scope of the systematic review. Our findings demonstrate that when developing scales for STEM education research, many types of psychometric properties, such as differential item functioning, test–retest reliability, and discriminant validity are scarcely reported. Furthermore, many scales only report internal structure validity (EFA and/or CFA) and Cronbach's alpha, which are not enough evidence alone. We encourage researchers to look towards the full spectrum of psychometric evidence both when choosing scales to use and when developing their own. While constructs such as attitudes and disciplines such as engineering were dominant in our sample, future work can fill in the gaps by developing scales for disciplines, such as geosciences, and examine constructs, such as engagement, self-efficacy, and perceived fit.

*Correspondence:
Danka Maric
dmaric@iu.edu
[1] STEM Education Innovation and Research Institute, Indiana University-Purdue University, 755 W. Michigan, UL1123, Indianapolis, IN 46202, USA
[2] Department of Chemistry and Chemical Biology, Indiana University-Purdue University, 402 N. Blackford Street, LD 326, Indianapolis, IN 46202, USA

Background

Measurement of students' experiences and instructor interventions continues to be an important aspect of Science, Technology, Engineering and Mathematics (STEM) education. Given the enormous educational and research efforts devoted to understanding students' development in STEM, advancing strategies that measure how well interventions and experiences work

is imperative. Sound measurement strategies require access to measures, scales, and instruments with adequate psychometric evidence (Arjoon et al., 2013; Cruz et al., 2020; Decker & McGill, 2019; Margulieux et al., 2019).

Quantitative measurement instruments, such as surveys, are commonly used tools in and important aspect of conducting discipline-based STEM education research (Arjoon et al., 2013; Decker & McGill, 2019; Knekta et al., 2019). Salmond (2008) emphasizes the importance of measures that "accurately reflect the phenomenon or construct of interest" (p.28) when designing effective student experiences. Measurement of student experiences and teaching innovations, however, is a complex task, as there can be many elements and constructs within a single experience or innovation, which continues to challenge STEM researchers and educators (Kimberlin & Winterstein, 2008). An even larger issue is the lack of measurement instruments grounded within holistic STEM theoretical frameworks to guide empirical research that specifically targets STEM students (Wang & Lee, 2019). Similarly, researchers need to be aware of measurement instruments available to them and be able to make choices appropriate for their research needs, which can be challenging due to the inability to locate validated and reliable measures (Decker & McGill, 2019).

There is a gap in post-secondary STEM education research when it comes to measurement and psychometric evidence. For example, a review of the 20 top mathematics, science, and STEM education journals found that less than 2% of empirically based publications in the last 5 years were instrument validation studies (Sondergelt, 2020).

Appianing and Van Eck (2018) share that whereas researchers have developed valid and reliable instruments to measure students' experiences in STEM, most of them focus on middle and high school students rather than on college students. Moreover, Hobson et al. (2014) have suggested that construction and usage of rubrics that effectively assess specific skill development such as collaboration, critical thinking, and communication in research continues to be a problem in STEM education. In summary, without effective measurement instruments and strategies that assess the efficacy of interventions and students' experiences, it is difficult to trace and document students' progress. Given that much of past research efforts in education have been focused on the K-12 level, and the fact that there has been increase in research interest in post-secondary STEM education research and publications (Li & Xiao, 2022; Li et al., 2022), this is especially needed at the post-secondary level.

A systematic review that compiles relevant scales in STEM post-secondary education and assesses available psychometric evidence is one strategy to mitigate the challenges above.

Scholars have also emphasized the importance of having a holistic understanding of measurement. This includes the statistical and theoretical underpinnings of validity, as well as the psychometric measures and dimensions that represent what is measured and evaluated, which are critical in the development, selection, and implementation of measurement instruments (Baker & Salas, 1992; Knekta et al., 2019). Likewise, there is a call to bring the design, testing, and dissemination of measurement instruments in STEM education to the forefront if researchers wish to have their quantitative results viewed as scientific by broader audiences (Sondergelt, 2020). Thus, in this study we provide insight into the measurement trends in survey instruments utilized in STEM education research and establish a resource that can be used by STEM education researchers to make informed decisions about measurement selections.

Some work has already been done to this end, with similar studies examining psychometric evidence for scales, measures, or instruments in chemistry (Arjoon et al., 2013), engineering (Cruz et al., 2020), and computer science (Decker & McGill, 2019; Margulieux et al., 2019) education research. These studies have examined and reported on psychometric evidence and the various constructs being measured in their respective fields as well as suggest professional development in measurement training for educators and researchers. For example, Arjoon et al. (2013) asserts that, to bridge the gap between what is known about measurement and what the actual accepted standards are, there is a need for measurement education within the chemistry education community. However, to our knowledge, no such study has been conducted across all of STEM education research. Thus, in the present study we build upon past work by conducting a systematic review of scales created for STEM education research, the psychometric evidence available for them, and the constructs they are measuring.

Purpose

The purpose of this systematic literature review is twofold. First, we aim to examine the measurement trends in survey instruments utilized in STEM education research. For example, we are interested in identifying which validated and reliable surveys are currently used in STEM education contexts, as well as what each instrument is measuring. Second, we intend for this paper not to be a repository of STEM education instruments per se, but to be a tool to be used by intermediate and expert STEM education and Discipline-Based Education researchers (DBER) to make informed decisions about measurement when conducting their research and collaborating

with others. In other words, we aim to produce a systematic literature review that critically examines the current measurement and psychometric development trends in STEM education research and, in doing so, illustrates areas, where STEM education research instrumentation might be lacking, where additional psychometric evaluation may be needed (as well as what tests those should be), and what kinds of surveys still need to be created and evaluated for STEM education research purposes. In doing so, our goal is to advance the development of robust and sound measurement instruments being used in the study of STEM education, teaching, and learning by helping researchers address some of the measurement challenges they currently face. We hope that by providing such a resource and pushing for advancements in measurement, we will contribute to the overall quality and advancement of the study of education, teaching, and learning within STEM post-secondary classrooms.

Theoretical framework

Our theoretical framework was informed by the 2014 edition of the *Standards,* jointly published by the American Education Research Association (AERA), the American Psychological Association (APA), and National Council on Measurement in Education (NCME). The *Standards* defines and outlines criteria for creating and evaluating educational and psychological tests, which includes scales and inventories under their definition. The *Standards* also provides criteria for test use and applications in psychological, educational, workplace, and evaluative contexts. For the purposes of the present review, we used the *Standards*' definitions, operationalizations, and criteria for psychometric evidence (reliability and validity). An overview of the theoretical framework that guided the formation of the coding framework and decision-making is displayed in Fig. 1. The definitions and operationalizations we used in the coding framework can be found below under *psychometric evidence*.

To define the term "scale", we draw on the Standards' definition for test as it encompasses the evaluative devices of tests, inventories, and scales, thus defining a scale as "a device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process" (AERA, APA, NCME, 2014, p. 2). We conceptualize validity as "the degree to which evidence and theory support the interpretations of test scores for proposed use of tests" (AERA, APA, NCME, 2014, p. 11) and reliability as "the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported" (AERA, APA, NCME, 2014, p. 33). Finally, we corroborated the National Science Foundation's (NSF) definition of STEM with information from a previous systematic review on STEM education (Gonzalez & Kuenzi, 2012; Martín-Páez et al., 2019) to define STEM education to include: science (biology, chemistry, computer and information science, geosciences, earth science), technology, engineering, mathematics, as well as any combinations of the above fields. Finally, we defined STEM education research as the multiple methodologies for exploring cross-disciplinary teaching, learning, and management strategies that increase student/public interest in STEM-related fields to enhance critical thinking and problem-solving abilities (Bybee, 2010).

Methods

We utilized the method articulated by Hess and Fore (2018), which itself was an extension of the method detailed by Borrego et al. (2014) for the present review. Following Hess and Fore (2018), we began by identifying the need for a systematic review such as this, and then proceeded to *define* the scope and focus of the study with three research questions (see next section). Then,

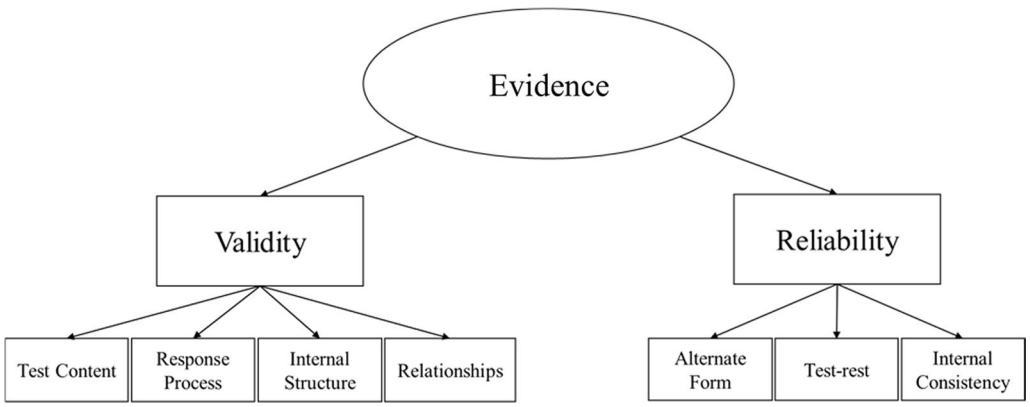

Fig. 1 Theoretical frame work of psychometric evidence

we initiated *scoping*, which is concerned with identifying the ways in which we were going to search for relevant literature.

We then proceeded to what we term *abstract screening* and then *full-text screening*, referred to as "cataloguing" by Hess and Fore (2018). Given the context of our systematic literature review and its demands, we determined that a slight shift was needed. In the context of the Hess and Fore study (i.e., engineering ethics education), the "cataloguing" step was concerned with the creation of the initial article database and the criteria by which inclusion and exclusion were determined. We also created an article database, but we changed it to a two-step screening process (*abstract screening* and *full-text screening*) in the present review. During these two phases, articles were screened against inclusion and exclusion criteria that were determined a priori, which are discussed in depth in *screening* below.

In Hess and Fore (2018), the "exploring" step was focused on crafting a coding structure and whittling down the article database further based upon the evolving coding parameters of the study in a deductive manner. However, our systematic review required a different approach.

Rather than using a deductive coding process, we created an a priori coding structure based on information outlined in the *Standards* (AERA, APA, and NCME, 2014). This approach allowed our article database to become specialized according to our study parameters. We only used emergent codes for categorizing the constructs being measured (see below under *coding*). Unlike Hess and Fore (2018), we did not further whittle down the article database in this phase, articles were simply coded for psychometric evidence present based on the coding structure. For this reason, we simply call this phase *coding*.

Next, during the *checking* phase, just as in the Hess and Fore (2018) review, we engaged in examining interrater reliability. Authors one, two, and three familiarized themselves with the coding structure and then we performed interrater reliability testing. Hess and Fore (2018) named their subsequent step "quantizing"; however, we decided to rename that step *results*, as we felt that this title better communicated the step's purpose to report the descriptive statistics related to our coding efforts. The two final steps identified by Hess and Fore (i.e., "interpreting" and "narrating") were merged into a step we simply titled *discussion*. In this step, we identified and interpreted our results before crafting an overview of all results.

Defining

As argued in the "Purpose" section above, there is a need for a broad systematic review of the literature on survey instruments across STEM fields. Seeking out and identifying a valid and reliable instrument for one's project may be laborious and time consuming, especially for those who may be starting out as STEM education researchers or discipline-based education researchers. This study seeks to introduce current instrumentation trends across STEM fields to provide researchers with a tool to identify rigorously developed scales which may foster insight into where psychometric work still needs to be done. To accomplish this, we seek to address three research questions:

> **RQ1:** *What are the valid and reliable measures being reported for use in post-secondary STEM education research in the United States between the years 2000 and 2021?*
> **RQ2:** *What are the common categories within which the measures can be organized?*
> **RQ3:** *What is the psychometric evidence that has been reported for these STEM education measures?*

Scoping

We started with an initial list compiled by the first author as an internal resource for STEM education research in our institute. Building upon this, a literature search was conducted using both quality-controlled and secondary sources (see Cooper, 2010). Quality-controlled sources included one book (Catalano & Marino, 2020), similar systematic reviews (Arjoon et al., 2013; Cruz et al., 2020; Gao et al., 2020; Margulieux et al., 2019), conference proceedings (Decker & McGill, 2019; Verdugo-Castro et al., 2019), and the Physics Education Research Center (PERC) online repository of measures. Secondary sources included the Web of Science, Education Resources Information Center (ERIC), SCOPUS, and PsycINFO databases, which index the major STEM education research journals. These journals were identified in a previous systematic review examining STEM education research publication trends (Li et al., 2020).

Constraints and limiters

We used several constraints and limiters to narrow down the number of papers obtained in the literature search. First, given that reliability and validity are complimentary, and that high reliability is needed for high validity (Knekta et al., 2019), we only included papers that reported on both validity and reliability. Second, because our work and expertise primarily revolve around STEM education in the United States, we were interested in measures used in STEM education research in the US. This decision was further informed by the fact that sometimes scores can differ between groups due to scale characteristics unrelated to the actual construct being

measured, thus introducing measurement bias (AERA, APA, & NCME, 2014). Likewise, population differences in factors such as culture and language necessitate an examination of the degree to which a scale measures the same construct across groups (Schmitt & Kuljanin, 2008), which can be especially important when comparing constructs, such as values, attitudes, and beliefs across groups (Milfont & Fischer, 2010; Van De Schoot, 2015). Thus, for the sake of simplicity and brevity, we only included studies conducted in the US and written in English. There were some exceptions, where samples from non-US countries were included with US samples; however, these papers reported examining measurement invariance between the US and non-US samples.

We further constrained the search to papers published between the years 2000 and 2021 (the present time of the search). Similar systematic reviews in chemistry, engineering, and interdisciplinary STEM education began their searches in the early 2000s (Arjoon et al., 2013; Cruz et al., 2020; Gao et al., 2020). Likewise, in the early 2000s there was an increased emphasis on institutional assessment of learning as well as the need for better assessment tools that reflect the goals and emphases of new courses and curricula being developed in STEM education (Hixson, 2013). Finally, the early 2000s is said to be when the term "STEM" was first used (Mohr-Schroeder et al., 2015; Li et al., 2020), which symbolically helps focus attention to STEM education efforts (Li et al., 2020, 2022). Taken together, we decided that the year 2000 would be a reasonable starting point for the present review.

We finally constrained the search to papers published in peer-reviewed journals or conference proceedings by clicking 'peer-reviewed only' when searching databases. We also limited the search to studies that included sampled college/university students that were 18 years or older and research settings, that are in post-secondary institutions (2-year college, 4-year college, or university), and STEM courses (based on our conceptualization and operationalization of STEM education above).

Search terms

The first author derived the search terms for the literature search using the thesaurus in the ERIC database and in consultation with a university librarian. These search terms were created based upon the research questions and constraints and limiters outlined above. Terms were derived from four main constructs of interest—STEM education, higher education, measures, and psychometrics—although specific Boolean operators and searching strategies varied slightly depending on the database. For full search terms, limiters, and operators used, please see Tables S1, S2, S3, and S4 in Additional file 1.

Screening

After the first author obtained the initial 603 studies from all sources and stored and managed them using an EndNote™ citation database, duplicates were deleted, and two rounds of screening and reduction against screening questions based on pre-determined inclusion and exclusion criteria were conducted. All screening questions could be answered with *yes, no,* or *unsure.* The *unsure* option was only used when enough information to answer the screening questions could not be obtained from the abstracts and thus had to undergo full-text review.

Besides the constraints and limiters outlined above, we had some extra considerations when developing screening questions. We did not consider observation protocols, interview protocols, rubrics, or similar instruments, because their development adheres to a set of standards distinct from surveys and scales and would be outside of the scope of the present study. We also excluded scales testing content knowledge, because they have limited opportunity for cross-disciplinary use due to their specificity. Finally, we omitted studies that included scales or subscales with less than three items, because using one- or two-item scales has been recognized as problematic (Eisinga et al., 2013). For example, in factor analysis, factors defined by one or two variables are considered unstable (Tabachnick & Fidell, 2014) and it has been argued that more items increase the likelihood of identifying the construct of interest (Eisinga et al., 2013).

Abstract screening

In the first round of screening, the first author reviewed just the abstracts and screened them for inclusion against the following five screening questions:

1. Does the study report the process of examining psychometric properties (i.e., evidence of validity/reliability) of a measure? (yes/no/unsure)
2. Does the study examine a measure meant to be used in a post-secondary setting (i.e., 4-year college, university, 2-year college)? (yes/no/unsure)
3. Does the study examine a quantitative measure (i.e., closed-response options such as Likert items)? (yes/no/unsure)
4. Are the participants in the study college/university students? (yes/no/unsure)
5. Has the measure been developed for a STEM education context? (yes/no/unsure)

The second round of screening included papers that had all screening questions marked with *yes*, and papers that had one or more screening questions marked with *unsure*. The most common reasons for exclusion in

this round included studies not reporting the process of collecting psychometric evidence (e.g., Lock et al., 2013; Romine et al., 2017) and the measure not being developed for a STEM education context (e.g., Dixon, 2015; Zheng & Cook, 2012). A total of 114 papers were marked for inclusion and seven were marked as unsure.

Full-text screening

In the second and final round of screening, the first author obtained full-text manuscripts and any corresponding supplementary materials for the 121 studies left after abstract screening. During the full-text screening process, four additional papers were found in the references of other papers, increasing the total number to 125. These studies were screened against the following three screening questions:

1. Does the measure under study ask questions about specific content knowledge? (yes/no)
2. Is there evidence for both reliability and validity for the measure? (yes/no)
3. Do scales/subscales have at least three items in them? (yes/no)

To be included in coding, the first question had to be marked *no*, and the second and third questions marked *yes*. After coding, the first author further organized the sample by scale, because some papers reported developing several scales and some scales were developed across multiple papers. See the PRISMA diagram in Fig. 2 for further information on the screening and reduction process as well as final sample sizes.

Many of the articles that were excluded in this round appeared to meet our criteria at first glance, but ultimately did not. The most common reason for this was because scales or subscales had less than three items each (e.g., Brunhaver et al., 2018; Jackson, 2018). Furthermore, although limiters were used when conducting the literature search, it was not obvious that several studies were non-US studies and these were ultimately excluded upon full-text review (e.g., Brodeur et al., 2015; Ibrahim et al., 2017). A few articles were also excluded, because the authors did not report both reliability and validity evidence (e.g., Godwin et al., 2013; Hess et al., 2018). It is important to note that just because a paper was excluded at any stage of the screening process does not mean that paper is of low quality. These papers simply did not meet our specific parameters. The final list of references of the articles included in the review can be found in Additional file 2 and further information on the scales can be found in Additional file 3 and Additional file 4.

Coding

Per recommendations for conducting systematic reviews and meta-analyses (Cooper, 2010), the first author created a coding framework to pull out sample information, descriptive information, and psychometric evidence for each scale. This was compiled into a codebook, which was shared with authors two and three, who served as the second and third coders, respectively. Each section of the coding framework is described below.

Sample information

The first author extracted sample sizes and characteristics of each sample used in scale development in each study. If several studies were reported in a single publication, the first author extracted sample characteristics for each study, when available. Specifically, for each scale, the first author coded the sample age (either the mean or age range), racial distribution (by percentile), and gender distribution (by percentile).

Descriptive information

The first author extracted the following descriptive information for each scale included in the review:

1. The number of items in the final scale.
2. The number of items in each subscale.
3. Whether the scale is a short form of a longer, previously developed scale.
4. The disciplinary context of the scale.
5. The construct or constructs the scale is measuring.
6. Scale response anchors.
7. The education level the scale is intended for.

The disciplinary context was coded in accordance with the predetermined definition of STEM education as outlined above in the theoretical framework section. The scale constructs were coded based upon the main constructs that were operationalized and defined by the authors of the scales. Thus, if a scale author stated that the scale was designed to measure chemistry attitudes, for example, then the construct was coded as "attitudes towards chemistry." The scale constructs were further developed into broader categories through emergent codes, which is described below under *checking*.

Psychometric evidence

Following similar systematic reviews (e.g., Arjoon et al., 2013), we created a coding structure based upon the psychometric evidence outlined in the *Standards* (AERA, APA, & NCME, 2014) for validity and reliability. Specifically, we pulled out the types of validity

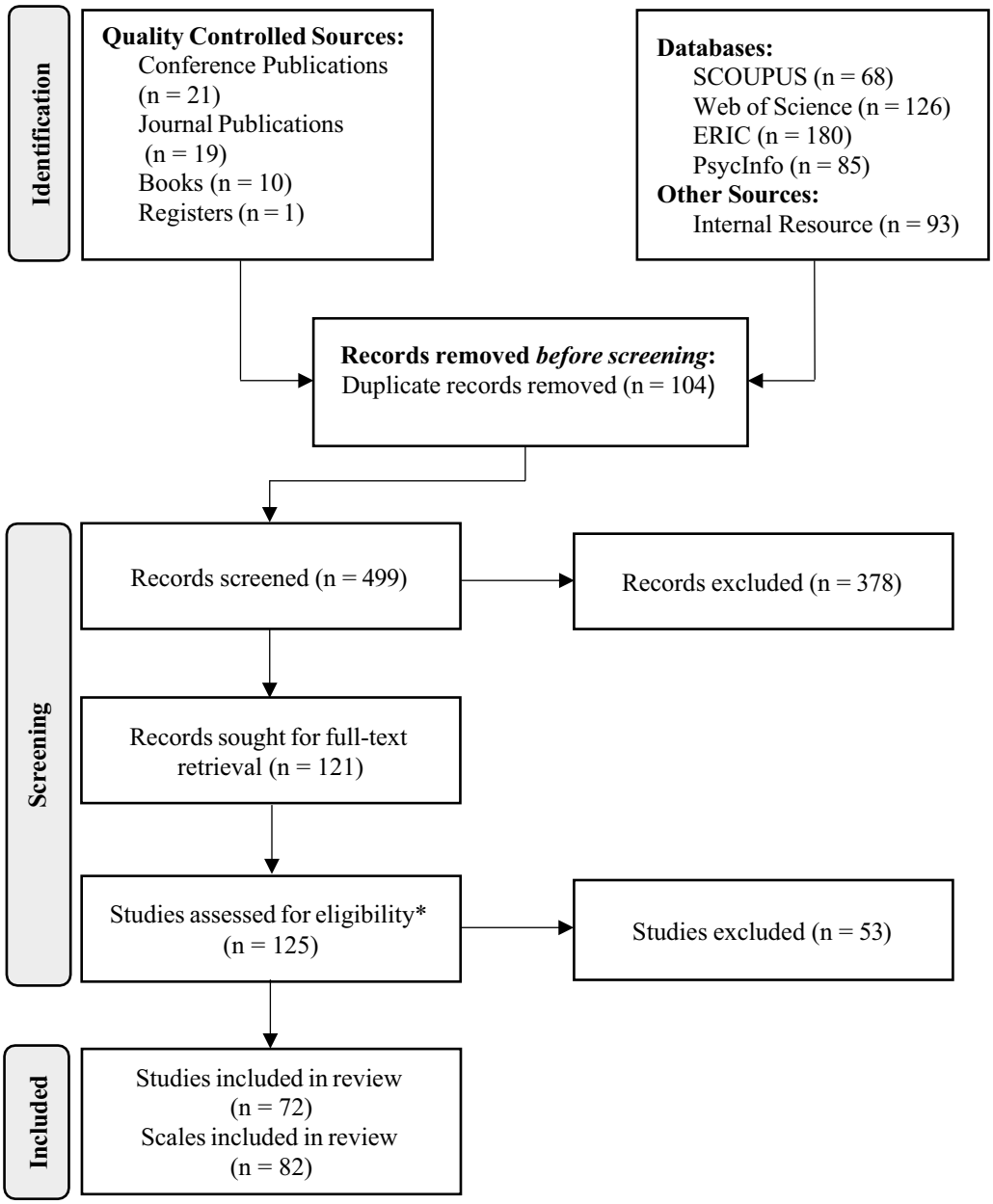

Fig. 2 PRISMA diagram. Four additional studies were found in the second round of screening and included in the full article review

and reliability evidence and used binary codes (yes/no) to mark them as either present or not present for each scale. Extra definitions for statistical techniques discussed below can be found in Additional file 5.

Validity

Per the *Standards* (AERA, APA, & NCME, 2014), validity evidence coded in this review included test content validity, response process validity, internal structure validity, and relationships with other variables. Test content validity was defined as evaluations from expert judges. Response process evidence was defined as evaluating cognitive processes engaged in by subjects through cognitive interviews, documenting response times, or tracking eye movements. Internal structure evidence was defined as the extent to which the relationships among test items and components conform to the construct on which the proposed test score interpretations are based. This included exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and Differential Item

Functioning (DIF). We also considered other statistical techniques not listed in *The Standards* such as Rasch Analysis, Q-sort methodology, Item Response Theory (IRT), and Multidimensional IRT.

Evidence based on relationships with other variables was defined as analyses of the relationship of test scores to variables external to the test. This included convergent validity, which examines whether scales are highly correlated with similar constructs, and discriminant validity, which examines whether scales are not correlated with dissimilar constructs. Test-criterion validity, which examines how accurately scores predict criterion (some attribute or outcome that is operationally distinct from the scale) performance is also included. Test criterion validity can come in the form of predictive validity, in which criterion scores are collected a later time, or concurrent validity, in which criterion scores are collected at the same time as scale scores. Finally, under evidence based on relationships, we also considered validity generalization via meta-analysis.

Reliability

Per the Standards (AERA, APA, NCME, 2014), reliability evidence considered in this review includes alternate-form reliability, test–retest reliability, and internal consistency. Alternate-form reliability was defined as the examination of the relationship between the measure and a different but interchangeable form of the measure, usually in the form of a correlation. Test–retest reliability is the examination of the relationship between two or more administrations of the same measure, also typically reported as a correlation. Finally, internal consistency includes the observed extent of agreement between different parts of one test that is used to estimate the reliability of form-to-form variability, which encompasses Cronbach's alpha and split-half coefficients. We also considered coefficients not listed in the standards, such as ordinal alpha and McDonald's Omega.

Checking

To check the trustworthiness of our data, we engaged in interrater reliability and data categorization between the first three authors.

Interrater reliability

We interpreted interrater reliability as a measure of the percentage of identically rated constructs between the three raters to ensure accuracy, precision, and reliability of coding behavior (Belur et al., 2021). O'Connor & Joffe (2020) suggest that it is prudent to double-code a randomly selected small amount of data (e.g., 10% of sample) rather than double-code a sizable quantity of data. Thus, due to the low number of samples (i.e., 72), we randomly double-coded 10% (seven) samples to estimate intercoder reliability. Though rules of thumb typically fall within 10% sub-samples (O'Connor & Joffe, 2020), in relation to Armstrong et al. (2020), our intercoder reliability value increased as we double-coded more samples but reached saturation by the fifth to seventh sample.

As described earlier, the first author created a priori rating scheme by compiling psychometric evidence for each scale. The first, second, and third authors applied the a priori coding structure to the same seven (9.72% of the total sample) articles and examined interrater reliability. The interrater agreement between the first author and second author was 93.40% and 73.40% between the first and third author. The agreement between the second author and third author was 74.27%. These values are higher than Belur et al. (2018) accepted range of 70.00% for systematic reviews. Again, the first three authors discussed and resolved all disagreements until a 100.00% agreement was achieved.

Categorization

We categorized constructs from each of the articles through emergent coding (see Drishko & Maschi, 2016). The first author organized the articles by scales to create initial categories of the 82 scales. The first three authors then engaged in an emergent coding process using the 72 articles and categorized the codes to create 18 primary and 11 secondary and one tertiary categories for the scales. This allowed for multiple or triple categorization of articles. Primary categories are the main overall construct the scale is measuring. When a second (or third in one case) clear but less prominent construct was evident, these were coded into the secondary and tertiary categories, respectively. For example, the article "Examining Science and Engineering Students' Attitudes Toward Computer Science" which measures both student interest and attitudes was multiple coded under attitudes as a primary category, and under interest as a secondary category. We provide a list of the codes and their definitions in Table 1.

Results

Descriptive information

Out of the 82 scales in our sample, only 12 were short forms of longer scales. The average scale length was $M = 29.84$ ($SD = 29.86$), although there was a wide range with the shortest being four items and the longest being 216 items. Out of the 82 scales in the sample, 62 reported containing subscales, with the median number of subscales per scale being three. The smallest number of items in a subscale was three items, while the largest number of items in a subscale was 30 items. Full information on the number of items in each subscale within each scale can

Table 1 Construct category definitions

Category	Definition
Affective outcomes	Students' positive and negative emotional activations/deactivations in relation to a STEM course or activities
Anxiety	Students' self-reported physiological reactivity, negative cognitions, and avoidance behaviors related to STEM
Attitudes	A psychological tendency that is expressed by evaluating a particular entity in STEM with some degree of favor or disfavor
Belonging and Integration	Students' sense that they are a part of a STEM course, program, or community, receive support, and have the skills for success
Cognitive Outcomes	Students' perceptions of their conscious reasoning related to thinking about a concept and constructing knowledge in STEM
Community Engagement	Students' perceptions of their skills and knowledge constructed through community relationships and practices
Course Perceptions	Students' assessments and expectations of, and their experiences in a STEM course
Diversity	Outcomes related to prejudice, inclusivity, stereotyping, equity, and inclusion in STEM contexts
Engagement	Students' behavioral, cognitive, and emotional involvement and investment in the learning process of a STEM course
External Climate	Students' perceptions and evaluations of the external context and/or environment that can shape their experiences in STEM
Identity	Students' interest in STEM and their ability to recognize themselves as someone that is aligned with STEM skills and values
Interest	Students' positive feelings towards and sense of value of STEM courses, fields, or activities
Learning Gains	Students' perceptions of the preparation and training they are receiving in a STEM course or program
Literacy	Students' perceptions, awareness, and exposure to STEM concepts knowledge and skills
Long-term Outcomes	Outcomes related to students persisting in and pursuing STEM majors and fields, pursuing further education in STEM, and/or pursuing employment in STEM
Motivation	Students' desire to participate in an STEM-related task that is influenced by their expectations and values
Non-technical Skills	Students' perceptions of their transferable skills including interpersonal/societal engagement, ethics, and management
Self-efficacy	Students' belief in their ability to achieve an STEM-related task
Social Support	Students' willingness to work with others and seek help in STEM courses

Table 2 Intended education level of the scales

	N	%
2-Year	3	3.66
Undergraduate	57	68.29
Graduate	2	2.44
Mixed	8	9.76
N/A	13	15.85

2-year includes community college programs. Mixed refers to any scales that were developed with a mix of education levels (i.e., undergraduate and graduate students)

be found in Additional file 4. We found that the majority of the scales in our sample were intended for use on an undergraduate sample or developed with undergraduate students (68.29%). Full information on education level can be found in Table 2. Very few scales were created for use with students in a 2-year setting (i.e., community college) and for graduate-level students.

The scales in our sample used response anchors that ranged from 3-point to 101-point response anchors, although the majority (42.90%) used 5-point Likert type response anchors, followed by 6-point (21.40), and 7-point (14.30%). Two instruments had response anchors that varied within subscales. The Full Participation Science and Engineering Accessibility (Jeannis et al., 2019) instrument used response anchors that ranged from 1 (strongly disagree) to 5 (strongly agree) for some sets of items, and response anchors that ranged from 1 (yes) to 3 (not present) for another set of items. Likewise, the Sustainable Engineering Survey (McCormick et al., 2015) used response anchors that ranged from 0 (no confidence) to 100 (very confident) for one subscale and anchors ranging from 0 (strongly disagree) to 5 (strongly agree) for other subscales.

The sample size used for the main statistical analyses (i.e., EFA, CFA, Cronbach's alpha, etc.) amongst the studies varied. The largest sample size used was $N=15{,}847$, while the smallest was $N=20$. Sample sizes most frequently fell between the range of 100–300 participants (34%), followed by 301 to 500 participants (24%). Full information on sample size ranges is displayed in Fig. 3.

When available, we noted participant demographic information. Of those that reported participant age, most of the scales reported age means and ranges between 18 and 25, which is not surprising given our post-secondary focus. Only one scale reported a range between 18 and 63 and another reported a range between 20 and 31. A total of 34 scales out of the 54 that reported a gender

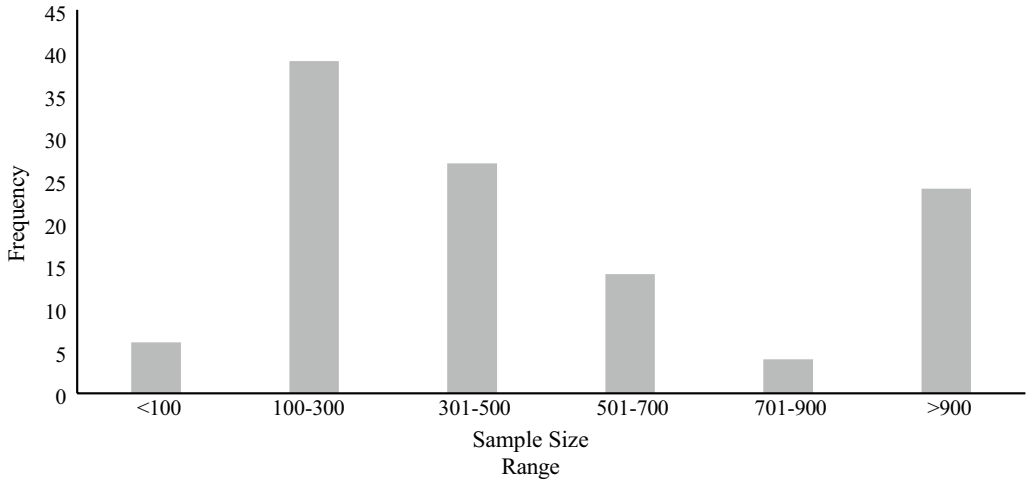

Fig. 3 Frequencies of sample size ranges. When papers reported on several studies (e.g., pilot study, main study) the sample sizes reported were counted separately. When papers reported samples from several populations (i.e., different universities), the samples were summed up and counted as one. Many papers reported sample sizes for expert judge evaluations and cognitive interviews, but we did not count those in this analysis

distribution, had a majority female sample. Of the 34 scales that have participant race and ethnicity available, 32 report White as either the majority or the largest group in their sample. Two scales reported an African American or non-White majority sample, respectively, and one scale reported a Hispanic/Latinx majority in one of their samples.

Given the large timespan we were working with, we checked whether psychometric trends differed between scales created before 2010 and after 2010 to check whether a larger analysis across time is needed. However, we could not do such an analysis for discriminant validity, predictive validity, other internal structure validity, alternate form reliability, and other internal consistencies, because there were so few datapoints. A series of Chi-square analyses showed that there were no statistically significant differences in occurrences of psychometric tests before 2010 compared to after 2010, except for the frequency of CFA. CFA was conducted more frequently post-2010 (56.9%), compared to pre-2010 (29.4%), X^2 (1, 82) = 4.08, $p = 0.04$. We further conducted a more granular examination by comparing pre-2010, 2010–2015, and post-2015, which did not yield any statistically significant results. Given that we only had one statistically significant result, we determined further analyses across the full 20 years were not necessary.

Scale disciplines

In our descriptive analysis, we identified several scales for each of the traditional STEM disciplines, except for geosciences which includes environmental and space science. Among the discipline-specific scales, most were designed for engineering, with biology being the least represented discipline. However, the largest proportion of scales (31.30%) were classified as unspecified STEM or science without specification to any discipline, and 7.20% of the scales were described as multidisciplinary (that is, more than one STEM discipline is specified). All scale disciplines can be found in Table 3.

Scale categories

In conversation with each other, the authors of this paper collaboratively assigned each instrument to a category denoting what the instrument was intended to measure, as it was not always obvious how a given measure should be categorized. The agreed upon categories of the instruments and the percentage of the overall sample that each represents are listed in Tables 4 and 5. We settled on 18 primary categories. When instruments were too complex to confine into one category, we identified secondary and/or tertiary categories.

Table 3 Scale disciplines

	N	%
Biology	1	1.20
Chemistry	9	11.00
Computer science	2	2.40
Engineering	21	25.60
Mathematics	9	11.00
Physics	5	6.10
Technology	4	4.90
Unspecified STEM or science	26	31.70
Multiple disciplines	5	6.10

Table 4 Primary construct categories

	N	%
Affective outcomes	1	1.20
Anxiety	3	3.70
Attitudes	14	17.10
Belonging and integration	4	4.90
Cognitive outcomes	4	4.90
Course perceptions	4	4.90
Diversity	7	8.50
Engagement	1	1.20
External climate	1	1.20
Identity	4	4.90
Interest	6	7.30
Learning gains	1	1.20
Literacy	3	3.70
Long-term outcomes	3	3.70
Motivation	9	11.00
Non-technical skills	11	13.40
Self-efficacy	5	6.10
Social support	1	1.20

Table 5 Secondary construct categories

	Category level	*N*	%
Attitudes	Secondary	1	1.20
Cognitive outcomes	Secondary	2	2.40
Community engagement	Secondary	4	4.90
Diversity	Secondary	1	1.20
Engagement	Secondary	1	1.20
External climate	Secondary	1	1.20
Interest	Secondary	1	1.20
Literacy	Secondary	2	2.40
Long-term outcomes	Secondary	1	1.20
Motivation	Secondary	2	2.40
Non-technical skills	Secondary	1	1.20
Self-efficacy	Tertiary	1	1.20

All primary categories can be found in Table 4. The most common kinds of instruments found in our literature sample were concerned with measuring attitudes (17.10%), various non-technical skills (13.40%), motivation (11.00%), diversity (8.50%), and interest (7.30%). That said, there was a great deal of variation with no category taking on a clear and overwhelming majority. This is unsurprising considering the breadth of our sampling. The categories that occurred the least included affective outcomes, engagement, external climate, learning gains, and social support, all of which occurred only once.

Table 6 Frequencies for validity evidence

	N	%
Judge evaluation	36	43.90
Response process	13	15.90
EFA	57	69.50
CFA	42	51.20
DIF	5	6.10
Other internal structure	9	11.00
Convergent	23	28.00
Discriminant	2	2.40
Concurrent	28	34.10
Predictive	3	3.70

Table 7 Frequencies for other types of internal consistency

	N	%
IRT	3	3.70
Q-sort methodology	1	1.20
Rasch analysis	2	2.40
Structural equation modeling	1	1.20
Multi-dimensional IRT	1	1.20
Reduced basis factor analysis	1	1.20

Of the sampled instruments, 17 were assigned a secondary category and one was assigned both a secondary and tertiary category (see Table 5). We found great variation here too, with most occurring once. There was only one instance ("community engagement") from the secondary and tertiary categories, where the authors felt the need to add a category that was not also included as a primary category. There was no instrument in our sample that focused primarily on community engagement; however, there were four surveys that focused on community engagement as a context for the primary category of non-technical skills. These all came from the same paper reporting on the Engineering Projects in Community Service (EPICS) program (Tracy et al., 2005). Community engagement was also the most frequently occurring secondary category (4.90%), followed by cognitive outcomes, literacy, and motivation, all of which occurred twice.

Overall validity evidence

In our sample, the median total number of validity evidence reported per scale was three, with a range of one to five types of validity reported per scale. Table 6 displays frequencies for validity evidence. The majority of the validity evidence reported in our sample was EFA and CFA for internal structure evidence, with 26 scales reporting evidence for both. A few scales reported other

types of internal structure evidence, such as IRT and Rasch Analysis (see Table 7). The next most frequent type of validity evidence in our sample was test content validity in the form of expert judge evaluations, then followed by concurrent test criterion validity. Validity evidence used to build nomological networks, such as convergent and discriminant validity, was reported less frequently, with convergent validity being the more prominent of the two. Although concurrent validity was used over a third of the time, evidence for test-criterion validity in the form of predicative validity was also scarce. We did not see generalization through meta-analysis reported in our sample.

We also examined the number of scales that contain joint evidence for the different combinations of the major four types of validity that grounded our theoretical framework (test content, response process, internal structure, and examining relationships). Analyses were completed using crosstabulations in IBM SPSS (version 28). Full information can be found in Table 8. The most frequent combination (45.12%) was reporting on internal structure validity along with the examination of relationships. The second most frequent combination was examining test content validity along with internal structure. All other combinations were less frequent with the combination of relationships and response process being the scarcest.

Validity evidence by construct

We further examined validity evidence by the categories representing the commonly measured constructs in our sample. For brevity, we only discuss the most frequently reported types of evidence, although all analyses are reported in Table 9. The one scale measuring affective outcomes, all the scales measuring engagement, learning gains, literacy, as well as the majority (over 50%) of the scales measuring motivation, course perceptions, long-term outcomes, and non-technical skills reported obtaining judge evaluations.

Scales from all categories except for the one scale on affective outcomes reported an EFA. This includes all scales measuring anxiety, engagement, external climate, learning gains, literacy, long-term outcomes, self-efficacy, and social support and at least 75% of those examining belonging and integration, course perceptions, and motivation. At least half of scales from all other categories reported on an EFA, except those measuring interest, one-third of which contained an EFA. Similarly, CFA was reported for scales in all categories except for those measuring course perceptions and literacy. We found that all the scales measuring affective outcomes, engagement, external climate, learning gains, and social support, as well as most of the anxiety, diversity, interest, long-term outcomes, motivation, and self-efficacy scales contained evidence for CFA. We observed that anywhere between one-quarter and half of scales in all other categories displayed evidence from CFA. Comparatively, DIF and other types of internal structure validity were scarcely reported.

Few scales reported response process, while reports on relationships between variables varied. Convergent validity evidence was found among all scales measuring affective outcomes, anxiety, and learning gains. We also observed convergent validity evidence among most of the self-efficacy scales, half of the course perceptions scales, over a third of the attitude scales, and for a third of the scales measuring interest. It was observed for at least a quarter of the scales measuring identity, belonging and integration, and a few of the scales for non-technical skills and motivation. Test-criterion validity, mostly in the form concurrent validity was reported for all scales measuring self-efficacy and social support, and most of the scales measuring anxiety, interest, and motivation. Half of the identity and belonging and integration scales, and one-third of the literacy scales reported concurrent validity. Concurrent validity evidence was found for one-quarter of those measuring cognitive outcomes and course perceptions, and less than a quarter of the time for all other categories. Conversely, predictive and discriminant validity were found in very few categories.

Overall reliability evidence

In our sample, the median total number of types of reliability evidence reported per scale was one, with a range of one to three types of reliability evidence per scale. Table 10 displays frequencies for reliability evidence. The most frequently reported reliability evidence in our sample was internal consistency, with a large skew towards

Table 8 Joint evidence reported for validity

	Test content	Response process	Internal structure	Relationships
Test content	–			
Response process	10 (12.20%)	–		
Internal structure	32 (39.02%)	11 (13.41%)	–	
Relationships	16 (19.51%)	4 (4.88%)	37 (45.12%)	–

Table 9 Percentage of articles reporting types of validity evidence broken down by category

	Judge evaluation %	Response process %	EFA %	CFA %	DIF %	Other internal structure %	Convergent %	Discriminant %	Concurrent %	Predictive %
Affective outcomes	100			100.00			100.00			
Anxiety			100.00	66.70			100.00	33.30	66.70	
Attitudes	21.40	28.60	71.40	50.00	7.10	7.10	35.70		14.30	14.30
Belonging and integration	25.00		75.00	25.00			25.00		50.00	
Cognitive outcomes	50.00	25.00	50.00	50.00					25.00	
Course perceptions	75.00	25.00	75.00				50.00		25.00	
Diversity	28.60	14.30	57.10	57.10						
Engagement	100.00		100.00	100.00						
External climate			100.00	100.00						
Identity	50.00	25.00	50.00	50.00			25.00		50.00	
Interest			33.30	66.70	50.00	50.00	33.30		66.70	
Learning gains	100.00		100.00	100.00			100.00			
Literacy	100.00		100.00		66.70	66.70			33.30	
Long-term outcomes	66.70	66.70	100.00	66.70		3.30				
Motivation	55.60	11.10	77.80	66.70			11.10		55.60	
Non-technical skills	72.70	9.10	54.50	27.30		18.25	18.20		18.20	
Self-efficacy	40.00		100.00	80.00			80.00	20.00	100.00	
Social support			100.00	100.00					100.00	

Table 10 Frequencies for reliability evidence

	N	%
Alternate Form	1	1.20
Test–retest	10	12.20
Cronbach's Alpha	77	93.90
Other internal consistency	9	11.00

Table 11 Frequencies for other types of internal consistency

	N	%
McDonald's Omega	7	8.40
Ordinal Alpha	1	1.20
Person separation variable	1	1.20

Table 12 Joint evidence reported for reliability

	Alternate form	Cronbach's Alpha	Test–retest
Alternate Form	–		
Cronbach's Alpha	1 (1.22%)	–	
Test–retest	1 (1.22%)	8 (9.76%)	–

from internal consistency, the second most frequent reliability evidence in our sample was test–retest reliability, although there is a considerable drop in occurrence compared to Cronbach's alpha. Likewise, only one study reported alternate form reliability in our sample and no studies in our sample reported split-half reliability.

We also examined the frequencies of joint evidence reported for the different combinations of reliability evidence using crosstabulations in IBM SPSS (version 28). Split-half reliability was not included in this analysis as it was not present in our sample. Full information can be found in Table 12. Given how much Cronbach's alpha dominated the reliability evidence found in our sample and how infrequent other sources were, it is unsurprising that there was not much joint evidence found. The most frequent combination was Cronbach's alpha reported with test–retest reliability. Other combinations only occurred once.

Reliability evidence by construct

Just as with validity evidence, we further broke down analyses for reliability by categories (see Table 13). Given the frequency of Cronbach's alpha in our sample, it is unsurprising that most of the scales across categories reported conducting Cronbach's alpha for internal consistency. It is reported for all the scales measuring

Table 13 Percentage of articles reporting types of reliability evidence broken down by category

	Alternate form %	Test–retest %	Cronbach's Alpha %	Other internal consistency %
Affective outcomes			100.00	
Anxiety	33.30	66.70	100.00	
Attitudes		14.30	92.90	
Belonging and integration			100.00	25.00
Cognitive outcomes		25.00	100.00	
Course perceptions			100.00	
Diversity		14.30	85.70	
Engagement			100.00	
External climate			100.00	
Identity		25.00	75.00	25.00
Interest			83.30	16.70
Learning gains			100.00	
Literacy			100.00	33.30
Long-term outcomes			100.00	
Motivation		22.20	100.00	11.10
Non-technical skills			90.90	9.10
Self-efficacy		20.00	100.00	40.00

Cronbach's alpha. Out of the other types of internal consistency reported, the most popular was McDonald's omega, with ordinal alpha and the person separation variable being reported only once each (see Table 11). Apart

affective outcomes, anxiety, belonging and integration, cognitive outcomes, course perceptions, external climate, engagement, learning gains, literacy, long-term outcomes, motivation, and self-efficacy, respectively.

Likewise, the majority of the scales in all other categories reported Cronbach's alpha for internal consistency. Other evidence for internal consistency was reported by almost half of the scales measuring self-efficacy, one third of the scales measuring literacy, and a quarter of the scales measuring belonging and integration as well identity. Beyond a few of the scales measuring interest, motivation, and non-technical skills, no other scales in any other categories reported examining other types of internal consistency.

Test–retest reliability was reported for most of the scales measuring anxiety and one-quarter of the scales measuring cognitive outcomes as well as identity. Beyond that, test–retest reliability was reported for a few of the scales measuring attitudes, diversity, motivation, and self-efficacy. No other scales in any other categories contained test–retest reliability evidence. Only scales measuring anxiety contained alternate form reliability evidence.

Discussion

The most frequently reported validity evidence in our sample was test content validity and internal structure validity. Specifically, evaluations from expert judges were reported for test content validity for nearly half of the scales and were present in most of the categories in our sample. Previous systematic reviews have similarly observed test content as a commonly reported type of validity evidence (Arjoon et al., 2013; Cruz et al., 2020; Decker & McGill, 2019). Although informative, scholars (e.g., Reeves et al., 2016) have argued that this type of validity evidence alone is not sufficient. We only had two scales in our sample that only have evaluations from expert judges as validity evidence.

For internal structure evidence, EFA and CFA were reported for over half and nearly half of the scales in our sample, respectively, and were well-represented in all but a few of the categories. However, other forms of internal structure validity, such as DIF, were much less prominent. Comparatively, a systematic review of chemistry education measures found internal structure validity evidence was reported in about half of the sample, with EFA being the most common and DIF completely lacking (Arjoon et al., 2013). While CFA is underutilized in chemistry education measures (Arjoon et al., 2013), it is reported more frequently across all STEM education research here, although our sample follows the trend of DIF being underreported. Similarly, other forms of internal structure validity were rare. While EFA and CFA can provide essential information about a scale's internal structure, other types of internal structure validity evidence can be valuable or even more appropriate.

Compared to test content and internal structure validity, we found that response process evidence was much less present, with only 13 scales reporting cognitive interviews. This aligns with similar work, which found a dearth of response process validity (Arjoon et al., 2013; Cruz et al., 2020), with cognitive interviews reported for only four out of 20 scales in one review (Arjoon et al., 2013). We also observed that evidence for relationships between variables—convergent, discriminant, and test-criterion validity—were far less present, with convergent and concurrent validity being reported on much more frequently than their counterparts. In contrast, previous work finds that all but one of the chemistry education scales in their sample reported some form of relationship with other variables (Arjoon et al., 2013). Looking across all STEM education disciplines, there may need to be more work to collect evidence based on relationships and build nomological networks around the constructs being measured.

Internal consistency, namely, Cronbach's alpha, was the most dominant reliability evidence and was prominent in all categories. All other forms of reliability evidence were reported far less frequently and were less represented across categories. This is unsurprising as others have reported similar observations in their reviews (Arjoon et al., 2013; Cruz et al., 2020; Decker & McGill, 2019), and as Cronbach's alpha is mistakenly provided as the only evidence for validity in many biology education research papers (Knekta et al., 2019). In comparison, test–retest reliability was reported for less than ten percent of our sample, alternate form only once, and split-half reliability was not observed at all, aligning with previous work (Arjoon et al., 2013; Cruz et al., 2020; Decker & McGill, 2019).

Although several gaps were observed, several surveys in the sample contained more comprehensive evidence and drew from several sources, which gave them a higher chance of being robust when used in a research setting. For example, the Engineering Professional Responsibility Assessment Tool (Canney & Bielefeldt, 2016) reported the highest amount of validity sources (five) and reported at least one piece of evidence from each of the four main categories in our theoretical framework. That said, this survey only reported one type of reliability evidence—ordinal alpha. The Engineering Skills Self-Efficacy Scale (Mamaril et al., 2016) also provided more comprehensive evidence with five sources reported and three of the categories in the theoretical framework represented (test content, internal structure, and relationships. This scale also reported two forms of internal constancy—Cronbach's alpha and McDonald's omega.

Surveys that reported more comprehensive reliability evidence were rare, although the Abbreviated Math

Anxiety Scale (Hopko et al., 2003) drew from the most sources in our sample (three)—alternate form, test–retest, and internal consistency. This scale also reported four sources of validity evidence from two categories (internal structure and relationships to other variables).

Implications for STEM education research

Psychometric development

Although a full discussion on psychometric evidence is beyond the scope of this review, validity is considered a unitary concept in contemporary theory (APA, AERA, & NCME, 2014; Reeves et al., 2016). This can be viewed as a continuum in which one's confidence in a measure grows as accumulating evidence increases support for its intended interpretations. There were several scales that did not go beyond reporting on EFA for validity evidence, and while this is a good starting point, EFA alone is not enough evidence and should ideally be corroborated with other sources (Knekta et al., 2019) For example, following up with a CFA can expand upon EFA by confirming the underlying factor structure found in the previous analytical results (DeVellis, 2017). We also echo past work (Arjoon et al., 2013) in encouraging scale developers in STEM education to examine DIF as it can provide valuable information about a scale's multidimensionality and whether items function differently among distinct groups (APA, AERA, NCME, 2014; Arjoon et al., 2013). Likewise, we suggest considering other forms of internal structure validity when appropriate, such as Rasch Analysis, IRT, or Q-sort methodology, to name a few. For example, IRT can allow researchers to examine item-level characteristics, such as item difficulty and item discrimination, as compared to factor analyses.

Beyond internal structure, other forms of validity provide valuable information.

Specifically, response process evidence, such as cognitive interviews, can provide insight as to how participants are interpreting and reasoning through questions (APA, AERA, NCME, 2014; Arjoon et al., 2013), which can provide important qualitative data missing in other forms of validity. Likewise, building a nomological network by examining a scale's relationships (or lack thereof) to other variables can illuminate how the scale fits in with a broader theoretical framework (Arjoon et al., 2013).

However, the median total of validity evidence sources amongst the scales in our sample was three. Furthermore, the majority of joint evidence reported was between internal structure and relationships and between internal structure and test content validity. Taken together, there was not a lot of breadth when it comes to the validity evidence that was examined. Although there is no "optimal number" of sources, drawing from multiple sources of evidence typically creates a more robust measure (APA, AERA, NCME, 2014). We recommend researchers carefully consider the goals of a measure and seek to examine a breadth of validity evidence and accumulate as much evidence as is needed and is feasible within their specific research contexts.

Reliability is also a fundamental issue in measurement that takes several different forms (DeVellis, 2017). However, we mostly observed evidence for internal consistency, which only provides evidence on the relationships between individual items. Alternate form evidence can demonstrate reliability by examining the relationship between the scale and an alternate scale, essentially replicating the scale (APA, AERA, NCME, 2014). Split-half reliability follows a similar logic to alternate form reliability by examining how two halves of a scale relate to each other (DeVellis, 2017). Test–retest reliability provides insight into a scale's consistency over time (DeVellis, 2017). Put simply, distinct types of reliability evidence provide different information, have various sources of error, and certain sources of evidence may be preferable depending on the context and needs of the research (APA, AERA, NCME, 2014). Despite this, very few scales in our sample examined some combination of reliability evidence and the median total of reliability sources was one. Given that each of these techniques has strengths and weaknesses, we encourage researchers to diversify reliability evidence in STEM education research, so that different sources of evidence can complement each other.

Prominence of Cronbach's alpha

Not only was Cronbach's alpha the most prominent form of internal consistency, but it was also the only type of reliability evidence observed for 64 of 82 scales in our sample. This is no surprise as Cronbach's alpha is commonly associated with instrument reliability in science education research (Taber, 2018). Although a full discussion around Cronbach's alpha is beyond the scope of the present review, it has been argued that Cronbach's alpha is not an ideal estimate of internal consistency, because it is typically a lower bound for the actual reliability of a set of items (DeVellis, 2017; Sijtsma, 2009; Taber, 2018). Beyond that, Cronbach's alpha relies on assumptions that are rarely met; and these assumption violations can lead to internal consistency estimate inflation (see DeVellis, 2017 and Dunn et al., 2014). It has also been critiqued, because the cutoffs (i.e., $\alpha = 0.70$) for what constitutes good or acceptable internal consistency are arbitrary (Taber, 2018).

Cronbach's alpha is also designed for continuous data, and it is argued that social science measures may not be continuous, thus making Cronbach's alpha inappropriate to use. The majority of the scales in our sample were on a 5-point scale and most scales used either Likert or

semantic differential response formats (see DeVellis, 2017 for discussion on response formats). Although the exact number of response options to include depends on a myriad of factors, it is argued that these types of response formats are ordinal rather than continuous in a strict sense, because one cannot assume that the intervals between response options are equal (DeVellis, 2017). Thus, scholars argue that this can lead to inaccuracies in Cronbach's alpha and suggest ordinal alpha as an alternative (DeVellis, 2017).

We recommend that researchers critically engage with the use of Cronbach's alpha and not to solely rely on it as evidence for internal consistency or overall reliability. Researchers have suggested additions and alternatives such as using bootstrapping to find the confidence interval around Cronbach's alpha to obtain a range of values for internal consistency, or using McDonald's Omega, to name a few (see DeVellis, 2017 and Dunn et al., 2014 for a full review). We suggest examining the individual advantages and disadvantages of each of these methods and using what is the most appropriate.

Disciplinary trends

As identified above, there were several disciplines represented in our sample with the most common disciplines being Unspecified STEM (31.3%), Engineering (25.3%), Chemistry (10.8%), and Mathematics (10.8%). Due to the federal push for advancing and investing in STEM education in the US (Holdren et al., 2010; Olson & Riordan, 2012), it is unsurprising to see unspecified STEM education instruments being the most popular scale discipline. Engineering lagging only slightly behind the unspecified STEM category was also foreseeable due to engineering education being a well-established discipline focused on quality discipline-based education research. However, we observed a distinct lack of scales in geosciences, as well as very few scales coming out of biology, computer science, and technology. As other disciplinary professionals further establish and/or expand their discipline-based education research efforts, we anticipate seeing more validated instruments arising therefrom.

Categorical trends

We observed a breadth of constructs being measured by the scales in our sample. Interestingly, we found that several constructs were seldom measured, which includes but not limited to engagement, belonging and integration, anxiety, and self-efficacy. A review in computer science education found that a sizable portion of their sample measured what they deemed non-cognitive processes, including self-efficacy, anxiety, and sense of belonging, among others (Decker & McGill, 2019). Another similar review found 76 measures out of 197 papers in their sample examined what they called experience measures, including motivation, self-efficacy, and engagement (Marguilieux et al., 2019). Finally, a review on assessment in interdisciplinary STEM education found that "the affective domain", which includes awareness, attitudes, beliefs, motivation, interest, and perceptions of STEM careers was the most frequent assessment target in their sample of papers (Gao et al., 2020). Aside from motivation, which was our second largest group in the primary categories, we noted many of these constructs only a few times in our sample. Thus, we recommend doing more work to develop scales measuring these constructs across the entirety of STEM education research.

Constructs that were observed more frequently included non-technical skills, constructs related to diversity, self-efficacy, and interest. We found that the majority of measures for non-technical skills came out of engineering. This is likely due, in part, to engineering education's significant focus on training, workforce development, and the need for professionalism in industry. Given such foci, as well as the extent to which engineering education has been embraced as its own disciplinary field, it is unsurprising to encounter extensive work in engineering ethics (Hess & Fore, 2018), professionalism (Felder & Brent, 2003; Layton, 1986; Shuman, et al., 2005), and interpersonal/societal engagement (Hess et al., 2018, 2021). With this in mind, we recommend other STEM disciplines consider examining these important professional skills.

Scales related to diversity, such as scales measuring racial/sex bias in STEM or stereotype threat susceptibility, were also a larger group in our sample. Given that there are significant disparities that exist in STEM education as well as calls to action to address these disparities, close achievement gaps, and diversify the STEM workforce (Jones et al., 2018), this was foreseeable. This trend also aligns with a review of high-impact empirical studies in STEM education, which found that the most frequently published topic pertained to cultural, social, and gender issues in STEM education (Li et al., 2022). Although these scales compromise one of the larger groups in our sample, because of how dispersed our categories were, there are only seven total. Given that diversity issues affect all STEM fields and that being able to assess diverse students' experiences is an important aspect in addressing disparities and gaps, there is more work to be done in the development of these scales.

Similarly, self-efficacy and interest were observed five and six times, respectively. Although these were among the larger groups in the sample, these are objectively not large numbers. Interest and self-efficacy work with each other as well with other factors to play an integral role

in student motivation, which affects students' academic behaviors, achievements, and choices (Knekta et al., 2020; Mamaril et al., 2016). Given these far-reaching effects, more measures examining these constructs across all domains of STEM education are needed.

Although no one category took on a clear majority, the construct of attitudes constituted the largest group. This is unsurprising as attitudes (among other non-cognitive constructs) are emphasized by many science educators as important for scientific literacy (Xu & Lewis, 2011). In our sample, attitudes encompassed a range of constructs. Some scales asked about students' beliefs on certain topics (e.g., Adams et al., 2006), others were more evaluative (e.g., Hoegh & Moskal, 2009), many reported generally examining attitudes (e.g., Cashin & Elmore, 2005), others assessed students' epistemologies and expectations (e.g., Wilcox & Lewandowski, 2016), and some focused on the cognitive affective components of attitudes (e.g., Xu & Lewis, 2011). In social psychology, which has a rich history in attitudes research, an attitude is defined as "a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor" (Eagly & Chaiken, 1993). Attitudes are comprised of cognitive (thoughts), affective (feelings), and behavioral (actions) responses to an object or event. Attitudes can be formed primarily or exclusively based on any combination of these three processes and any of these can serve as indicators for attitudes in measurement. Given the range we observed, when measuring attitudes, we suggest that researchers ground these scales in attitude theory and to specifically define which aspects are being measured.

Practical implications for researchers

Our goal for the present systematic review is to serve STEM education researchers, DBER professionals, and professionals that work in STEM education centers and institutions across the United States. Whether one is conducting research themselves or collaborating with STEM professionals to help them conduct STEM education research, we hope that researchers may use this review as a foundation when making decisions about measurement. Several practical implications are discussed below.

First, researchers and professionals may use this review when deciding whether to create a new scale, to use a pre-existing scale as is, or to adapt. In general, it is ideal to use a scale that already exists (DeVellis, 2017). The present review gives an overview of what is available, allowing researchers to determine whether they can use scales from our sample or adapt them. When multiple adequate pre-existing measures are available, it is important to consider the amount, variety, and quality of reported psychometric evidence. As psychometric development is an ongoing process, scales with a greater variety of quality psychometric evidence will be more robust and trustworthy. Although we appreciate that there will be occasions when only one scale is available, our goal is that this systematic review can inform research decisions about what scales are most trustworthy and robust. Finally, it is important to remember that one should use the full scale, and not just 'cherry-pick' questions when using pre-existing measures. One reason for this is because factor analysis is designed to examine the relationships between sets of survey items and whether subsets of items relate more to each other rather than other subsets (Knekta et al., 2019). Thus, if one uses a select few items from a set, the validity evidence previously collected is no longer relevant, and re-validating the scale is recommended before its use.

When using pre-existing measures, one must also consider the sample that was used to develop the measure and the context within which it was developed. Sample demographics (i.e., race, gender, etc.) are an important factor to consider due to the potential for measurement bias. It is rare for measurement parameters to be perfectly equal across groups and measurement occasions (Van de Schoot, 2015) and thus there is always potential for measurement bias. For this reason, it is important to report sample demographics when developing a scale and to also consider demographics when using a pre-existing scale. If the population one is sampling from is quite different from the one that was used to develop the scale, it may not be appropriate to use that measure without examining measurement invariance before its use.

Due to the potential for measurement bias across different groups, it also may not always be appropriate to use pre-existing measures developed for other disciplines or even measures developed for an unspecified STEM context, especially if one's research questions are discipline specific. However, one can always adapt pre-existing measures if they do not wish to create a new one. Several scales in our sample were adapted and/or revalidated from scales designed for a different discipline or for all college students, such as the Academic Motivation Scale—Chemistry (Liu et al., 2017), the Civic-Minded Graduate Scale in Science and Engineering (Hess et al., 2021), and the Colorado Learning Attitudes about Science Survey (Adams et al., 2006). Researchers can look to these as examples when adapting or revalidating scales within their own disciplines. Depending on what is being measured, significant adaptations to measures from other disciplines may not be needed. To illustrate, when adapting an academic motivation scale, one may only need to change the discipline referenced in scale items. However, if one seeks to examine non-technical skills (often also referred to as soft skills) specific to a single discipline,

then significant changes may be necessary, thus requiring more involved adaptations and psychometric evaluations.

Similarly, older scales may sometimes need to be re-examined for use in today's context and society. One should carefully examine scale items and consider whether they fit in the context in which they are trying to use them. Sometimes items or words pertain to something that is no longer relevant or is obsolete in today's context. However, this does not mean the scale or measure itself is of poor quality. One can change items or make updates and re-examine psychometric evidence accordingly.

Finally, the present review aims to give researchers and professionals a sense of where there are gaps and allow them to make more informed decisions about when to create a new scale in a developing field, such as STEM education research. As we have emphasized the complex and ongoing process that is psychometric development, researchers may look to this review to get a broad sense of what kind of psychometric evidence can be examined and the purposes they serve. However, we encourage corroborating this review with the resources we cite such as the Standards (APA, AERA, NCME, 2014), DeVellis (2017) and Knekta et al. (2019) and other quality resources available.

Limitations and future directions

Several significant limitations of this study are inherent in the inclusion criteria and sampling strategy, which examined only higher education STEM research in the United States. Although sampled literature consisted of diverse STEM fields, a single country and a concentration on only undergraduate and graduate education limits the possibilities of generalization to a broader population. We also excluded literature from dissertations, thesis, and non-peer-reviewed articles which have the possibility to limit our findings. We suggest further studies should examine the measurement trends in survey instruments utilized in STEM education research among a wider population of samples from other countries and education levels, such as K-12, and to include dissertations and theses, and non-peer-reviewed papers to extend our findings. Although we did not find it necessary to conduct analyses across time in our sample, as STEM education research grows and measurement further develops, especially as more disciplines get involved and more constructs are added, future research may examine trends across time once more datapoints exist. Finally, our samples contained uneven group sizes in categories and disciplines which made comparative analysis of the samples difficult for the researchers.

Summary of recommendations

The following recommendations were developed through a synthesis of the patterns we observed in our analyses as well as information from the *Standards*—the theoretical framework that informed this review. Although we hope that these recommendations may serve our readers well in their own pursuits, it is important to note that they do not cover the full scope of psychometric development. Discussing the full nuance of the process is beyond the scope of this work and we strongly encourage readers to engage with the *Standards* and other resources we cite (i.e., DeVellis, 2017), which have the space to provide more detailed discussion on these topics.

1. Measurement is fully dependent on the context of one's research and decisions will be unique to each researcher. Before making any decisions, carefully consider the research context, questions, goals, and population.
2. It is typically preferable to use a pre-existing measure whenever possible (DeVellis, 2017). If one has found a scale that might be a good fit for their research, we recommend:
    a. Comparing the population and context that the scale was developed with and within to one's own. Are they similar? This will help determine how suitable the scale may be or if adaptations will be needed.
    b. Using scales in full, the way that they were intended to be used when they were developed. You cannot 'cherry-pick' items. If items need to be removed, because they are not relevant, then collecting psychometric evidence again would be needed.
3. If one has determined they need to create their own scale, there are many ways to begin.

We recommend looking at the relevant theories, past research, similar scales, or using qualitative data to create scales. All of these are good starting points, but it is up to the researcher to decide which one is the most appropriate.

4. When collecting psychometric evidence, consider what is needed as well as what is most feasible. This includes considering the size of the sample one is working with, timeframe, the structure of the scale itself, and its intended use. Given the relationship between validity and reliability, we recommend collecting some form of validity evidence and some form of reliability evidence.

5. Because measurement invariance can arise with many group variables, we recommend collecting demographic data from the sample the scale was developed on. This includes (but is not limited to) variables, such as race, gender, class standing, and age.
6. Many of the scales in our sample rely solely on exploratory factor analysis and/or confirmatory factor analysis for validity evidence. Validity is the degree to which accumulating evidence supports the interpretation of a scale for an intended use (AERA, APA, and NCME, 2014). Adequate support typically involves multiple sources of evidence but does not always require all sources of evidence outlined in this review. We recommend considering what interpretations and intended uses one has for a scale and then deciding which sources will be most appropriate. For example, if one wishes to use a scale to predict student GPA, then they would need to collect predictive validity evidence. If one wishes to propose that a scale is suitable for a specific content domain, it would be prudent to collect test content validity evidence. It is prudent to collect evidence that supports all intended propositions that underly a scale's intended interpretations.
7. When collecting reliability evidence, consider what kinds of decisions will be informed by the scale's use. Consider how reversible these decisions would be and whether they can be corroborated with other information sources (AERA, APA, and NCME, 2014). Although reliable and precise measurement is always important, these considerations will inform how modest or high that degree of precision should be.
8. Most scales in our sample collected information on internal consistency only, even though it typically takes multiple sources of evidence. Just as with validity evidence, we recommend collecting from as many sources of evidence as possible while taking into consideration the intended purposes of the scale. For example, if one is proposing that a scale's items are interrelated, then internal consistency is important to measure. If one is measuring an attribute that is not expected to change across an extended time period, test–retest reliability with scores collected across 2 days would be appropriate. Which sources of evidence and how many one wishes to draw upon will look different for each researcher.
9. When choosing statistical tests for validity and reliability evidence collection, we recommend not simply relying on what is most typically used. For example, there are many options that exist for examining internal structure validity and internal consistency, yet many of the scales in this review rely on exploratory factor analysis and Cronbach's alpha, respectively. We recommend considering a broader spectrum of statistical analyses when collecting psychometric evidence.

Conclusion

Through this systematic literature review, we have found that there is a great deal of quantitative instrumentation being used by STEM education researchers and evaluators to measure the outcomes of STEM education activities and courses. That said, there are many published instruments that lack thorough assessments of their validity and reliability. Of those instruments that have been held up to rigorous testing of validity and reliability, there is still more work that can be done, particularly regarding the use of different approaches—other than Cronbach's alpha—to examine reliability. As STEM education researchers build up a canon of validated and reliable instruments measuring a variety of different learning outcomes, we approach the potential for creating a repository for STEM education surveys.

Moving forward, there is a need for more instruments to be created for a greater diversity of learning outcomes and STEM fields, as well as a need for more rigorous and diversified psychometric development of these instruments. STEM education researchers, as mentioned above, may benefit from having more scales that measure engagement, sense of belonging, perceived fit, anxiety, and self-efficacy. It may also be worthwhile for STEM education researchers to examine the education psychology literature (and related fields) to identify additional instruments that may have never been or that have rarely been used in STEM settings. Such an approach could open STEM education researchers up to a variety of new kinds of validated and reliable instruments, allowing for more complex, sophisticated, and insightful studies, and analyses.

Abbreviations

STEM	Science, technology, engineering, and mathematics
DBER	Discipline-based education research
AERA	American Education Research Association
APA	American Psychological Association
NCME	National council on measurement in education
NSF	National science foundation
PERC	Physics education research center
ERIC	Education resources information center
EFA	Exploratory factor analysis
CFA	Confirmatory factor analysis
DIF	Differential item functioning
IRT	Item response theory
EPICS	Engineering projects in community service

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40594-023-00430-x.

Additional file 1. Search terms, boolean phrases, and limiters used in databases.

Additional file 2. Articles reviewed.

Additional file 3. Scale names, publications, year published, and journal/conference.

Additional file 4. Scale descriptive information.

Additional file 5. Definitions of statistical techniques.

Additional file 6. Primary dataset.

Additional file 7. Categories dataset.

Acknowledgements
We would like to acknowledge and thank Precious-Favour Nguemuto Taniform for her work in coding the education level of each scale, which was an important contribution in putting together the table in Additional file 4.

Author contributions
The first author was responsible for organizing the project team, creating the protocol and coding framework, conducting literature searches and managing collected studies, screening, coding, data analysis, and writing a substantial amount of the manuscript. The second author was responsible for providing feedback during protocol and codebook development, coding, and writing a substantial amount of the manuscript. The third author was responsible for coding and writing a substantial amount of the manuscript. The fourth author was responsible for helping to organize the project team, overseeing the progress of the project, and providing continuous consultation, feedback, and edits during manuscript preparation.

Funding
This study was not supported by any internal or external funding sources.

Availability of data and materials
Supporting data is available in Additional file 6 and Additional file 7.

Declarations

Ethics approval and consent to participate
Because our study does not involve any animals, humans, human data, human tissue or plants, this section is not applicable.

Consent for publication
Because our manuscript does not contain any individual person's data, this section is not applicable.

Competing interests
We affirm no competing interests financially or otherwise with the outcomes of this research.

Received: 17 October 2022 Accepted: 19 May 2023
Published online: 02 June 2023

References

Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics-Physics Education Research, 2*(1), 010101.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Appianing, J., & Van Eck, R. N. (2018). Development and validation of the Value-Expectancy STEM Assessment Scale for students in higher education. *International Journal of STEM Education, 5*(1), 1–16.

Arjoon, J. A., Xu, X., & Lewis, J. E. (2013). Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence. *Journal of Chemical Education, 90*(5), 536–545.

Baker, D. P., & Salas, E. (1992). Principles for measuring teamwork skills. *Human Factors, 34*(4), 469–475.

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods & Research, 50*(2), 837–865.

Borrego, M., Foster, M. J., & Froyd, J. E. (2014). Systematic literature reviews in engineering education and other developing interdisciplinary fields. *Journal of Engineering Education, 103*(1), 45–76.

Brodeur, P., Larose, S., Tarabulsy, G., Feng, B., & Forget-Dubois, N. (2015). Development and construct validation of the mentor behavior scale. *Mentoring & Tutoring: Partnership in Learning, 23*(1), 54–75.

Brunhaver, S. R., Bekki, J. M., Carberry, A. R., London, J. S., & McKenna, A. F. (2018). Development of the Engineering Student Entrepreneurial Mindset Assessment (ESEMA). *Advances in Engineering Education, 7*(1), n1.

Bybee, R. W. (2010). What is STEM education? *Science, 329*(5995), 996–996.

Canney, N. E., & Bielefeldt, A. R. (2016). Validity and reliability evidence of the engineering professional responsibility assessment tool. *Journal of Engineering Education, 105*(3), 452–477.

Cashin, S. E., & Elmore, P. B. (2005). The Survey of Attitudes Toward Statistics scale: A construct validity study. *Educational and Psychological Measurement, 65*(3), 509–524.

Catalano, A. J., & Marino, M. A. (2020). Measurements in evaluating science education: *A compendium of instruments, scales, and tests.* ProQuest Ebook Central https://ebookcentral-proquest-com.proxy.ulib.uits.iu.edu

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Sage Publications Inc.

Cruz, M. L., Saunders-Smits, G. N., & Groen, P. (2020). Evaluation of competency methods in engineering education: A systematic review. *European Journal of Engineering Education, 45*(5), 729–757.

Decker, A., & McGill, M. M. (2019, February). A topical review of evaluation instruments for computing education. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (pp. 558–564).

DeVellis, R. F. (2017). *Scale development: Theory and applications*. Sage publications.

Dixson, M. D. (2015). Measuring student engagement in the online course: The Online Student Engagement scale (OSE). *Online Learning, 19*(4), n4.

Drishko, J. W., & Maschi, T. (2016). *Content analysis*. Oxford University Press.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412.

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Eisinga, R., Grotenhuis, M. T., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health, 58*(4), 637–642.

Felder, R. M., & Brent, R. (2003). Designing and teaching courses to satisfy the ABET engineering criteria. *Journal of Engineering Education, 92*(1), 7–25.

Gao, X., Li, P., Shen, J., & Sun, H. (2020). Reviewing assessment of student learning in interdisciplinary STEM education. *International Journal of STEM Education*. https://doi.org/10.1186/s40594-020-00225-4

Godwin, A., Potvin, G., & Hazari, Z. (2013). The development of critical engineering agency, identity, and the impact on engineering career choices. In *2013 ASEE Annual Conference & Exposition* (pp. 23–1184).

Gonzalez, H. B., & Kuenzi, J. J. (2012). Science, technology, engineering, and mathematics (STEM) education: A primer. Congressional Research Service, Library of Congress.

Hess, J. L., Chase, A., Fore, G. A., & Sorge, B. (2018). Quantifying interpersonal tendencies of engineering and science students: A validation study. *The International Journal of Engineering Education, 34*(6), 1754–1767.

Hess, J. L., & Fore, G. (2018). A systematic literature review of US engineering ethics interventions. *Science and Engineering Ethics, 24*(2), 551–583.

Hess, J. L., Lin, A., Fore, G. A., Hahn, T., & Sorge, B. (2021). Testing the Civic-Minded Graduate Scale in science and engineering. *International Journal of Engineering Education, 37*(1), 44–64.
Hixson, S. H. (2013). Trends in NSF-Supported Undergraduate Chemistry Education, 1992-2012. In *Trajectories of Chemistry Education Innovation and Reform* (pp. 11–27). American Chemical Society.
Hobson, C. J., Strupeck, D., Griffin, A., Szostek, J., & Rominger, A. S. (2014). Teaching MBA students teamwork and team leadership skills: An empirical evaluation of a classroom educational program. *American Journal of Business Education (AJBE), 7*(3), 191–212.
Hoegh, A., & Moskal, B. M. (2009). Examining science and engineering students' attitudes toward computer science. In *2009 39th IEEE Frontiers in Education Conference* (pp. 1–6). IEEE.
Holdren, J., Lander, E., & Varmus, H. (2010). *Prepare and inspire: K-12 science. technology, engineering and math (STEM) education for America's Future*. Executive Office of the President of the United States.
Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The abbreviated math anxiety scale (AMAS) construction, validity, and reliability. *Assessment, 10*(2), 178–182.
Ibrahim, A., Aulls, M. W., & Shore, B. M. (2017). Teachers' roles, students' personalities, inquiry learning outcomes, and practices of science and engineering: The development and validation of the McGill attainment value for inquiry engagement survey in STEM disciplines. *International Journal of Science and Mathematics Education, 15*(7), 1195–1215.
Jackson, C. R. (2018). Validating and adapting the motivated strategies for learning questionnaire (MSLQ) for STEM courses at an HBCU. *Aera Open, 4*(4), 2332858418809346.
Jeannis, H., Goldberg, M., Seelman, K., Schmeler, M., & Cooper, R. A. (2019). Participation in science and engineering laboratories for students with physical disabilities: Survey development and psychometrics. *Disability and Rehabilitation: Assistive Technology*.
Jones, J., Williams, A., Whitaker, S., Yingling, S., Inkelas, K., & Gates, J. (2018). Call to action: Data, diversity, and STEM education. *Change the Magazine of Higher Learning, 50*(2), 40–47.
Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*(23), 2276–2284.
Knekta, E., Runyon, C., & Eddy, S. (2019). One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE Life Sciences Education, 18*(1), 1–17.
Knekta, E., Rowland, A. A., Corwin, L. A., & Eddy, S. (2020). Measuring university students' interest in biology: Evaluation of an instrument targeting Hidi and Renninger's individual interest. *International Journal of STEM Education, 7*(1), 1–16.
Layton, E. T., Jr. (1986). *The Revolt of the Engineers*. Johns Hopkins University Press.
Li, Y., Wang, K., Xiao, Y., & Froyd, J. E. (2020). Research and trends in STEM education: A systematic review of journal publications. *International Journal of STEM Education, 7*(1), 1–16.
Li, Y., & Xiao, Y. (2022). Authorship and topic trends in STEM education research. *International Journal of STEM Education, 9*(1), 1–7.
Li, Y., Xiao, Y., Wang, K., Zhang, N., Pang, Y., Wang, R., Qi, C., Yuan, Z., Xu, J., Nite, S. B., & Star, J. R. (2022). A systematic review of high impact empirical studies in STEM education. *International Journal of STEM Education, 9*(1), 72.
Liu, Y., Ferrell, B., Barbera, J., & Lewis, J. E. (2017). Development and evaluation of a chemistry-specific version of the academic motivation scale (AMS-Chemistry). *Chemistry Education Research and Practice, 18*(1), 191–213.
Lock, R. M., Hazari, Z., & Potvin, G. (2013). Physics career intentions: The effect of physics identity, math identity, and gender. In *AIP Conference Proceedings* (Vol. 1513, No. 1, pp. 262–265). American Institute of Physics.
Mamaril, N. A., Usher, E. L., Li, C. R., Economy, D. R., & Kennedy, M. S. (2016). Measuring undergraduate students' engineering self-efficacy: A validation study. *Journal of Engineering Education, 105*(2), 366–395.
Margulieux, L., Ketenci, T. A., & Decker, A. (2019). Review of measurements used in computing education research and suggestions for increasing standardization. *Computer Science Education, 29*(1), 49–78.
Martín-Páez, T., Aguilera, D., Perales-Palacios, F. J., & Vílchez-González, J. M. (2019). What are we talking about when we talk about STEM education? A review of literature. *Science Education, 103*(4), 799–822.
McCormick, M., Bielefeldt, A. R., Swan, C. W., & Paterson, K. G. (2015). Assessing students' motivation to engage in sustainable engineering. *International Journal of Sustainability in Higher Education, 16*(2), 136–154.
Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111–130.
Mohr-Schroeder, M. J., Cavalcanti, M., & Blyman, K. (2015). STEM education: Understanding the changing landscape. In *A practice-based model of STEM teaching* (pp. 3–14). Brill.
O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: debates and practical guidelines. *International Journal of Qualitative Methods, 19*, 1609406919899220.
Olson, S., & Riordan, D. G. (2012). *Engage to excel: producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. Report to the president.* Executive Office of the President.
Reeves, T. D., Marbach-Ad, G., Miller, K. R., Ridgway, J., Gardner, G. E., Schussler, E. E., & Wischusen, E. W. (2016). A conceptual framework for graduate teaching assistant professional development evaluation and research. *CBE Life Sciences Education, 15*(2), es2.
Romine, W. L., Walter, E. M., Bosse, E., & Todd, A. N. (2017). Understanding patterns of evolution acceptance—A new implementation of the Measure of Acceptance of the Theory of Evolution (MATE) with Midwestern university students. *Journal of Research in Science Teaching, 54*(5), 642–671.
Salmond, S. S. (2008). Evaluating the reliability and validity of measurement instruments. *Orthopaedic Nursing, 27*(1), 28–30.
Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*(4), 210–222.
Shuman, L. J., Besterfield-Sacre, M., & McGourty, J. (2005). The ABET "professional skills"—Can they be taught? Can they be assessed? *Journal of Engineering Education, 94*(1), 41–55.
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120.
Sondergelt, T. A. (2020). Shifting sights on STEM education quantitative instrumentation development: The importance of moving validity evidence to the forefront rather than a footnote. *School Science and Mathematics, 120*(5), 259–261.
Tabachnick, B. G. & Fidell, L.S. (2014). *Using multivariate statistics.* Pearson Education Limited.
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.
Tracy, S., & Immekus, J., & Maller, S., & Oakes, W. (2005), Evaluating the outcomes of a service-learning based course in an engineering education program: Preliminary results of the assessment of the engineering projects in community service epics. *In 2005 ASEE Annual Conference & Exposition.*
Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in Psychology, 6*, 1064.
Verdugo-Castro, S., García-Holgado, A., & Sánchez-Gómez, M. C. (2019). Analysis of instruments focused on gender gap in STEM education. In *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 999–1006).
Wang, X., & Lee, S. Y. (2019). Investigating the psychometric properties of a new survey instrument measuring factors related to upward transfer in STEM fields. *The Review of Higher Education, 42*(2), 339–384.
Wilcox, B. R., & Lewandowski, H. J. (2016). Students' epistemologies about experimental physics: Validating the Colorado Learning Attitudes about Science Survey for experimental physics. *Physical Review Physics Education Research, 12*(1), 010123.
Xu, X., & Lewis, J. E. (2011). Refinement of a chemistry attitude measure for college students. *Journal of Chemical Education, 88*(5), 561–568.
Zheng, R., & Cook, A. (2012). Solving complex problems: A convergent approach to cognitive load measurement. *British Journal of Educational Technology, 43*(2), 233–246.

Publisher's Note