Chemistry Education Research and Practice

Practices in instrument use and development in Chemistry Education Research and Practice 2010-2021

Journal:	*Chemistry Education Research and Practice*
Manuscript ID	RP-ART-10-2022-000275.R2
Article Type:	Paper
Date Submitted by the Author:	24-Feb-2023

Complete List of Authors:	Lazenby, Katherine; NWEA; San Diego State University College of Sciences, Chemistry & Biochemistry Tenney, Kristin; San Diego State University, Center for Research in Mathematics and Science Education Marcroft, Tina; San Diego State University, Center for Research in Mathematics and Science Education Komperda, Regis; San Diego State University College of Sciences, Chemistry & Biochemistry; San Diego State University, Center for Research in Mathematics and Science Education

ARTICLE

Practices in instrument use and development in *Chemistry Education Research and Practice* 2010-2021

Katherine Lazenby,[ac] Kristin Tenney,[b] Tina A. Marcroft[b] & Regis Komperda*[bc]

[a]*NWEA, 121 NW Everett St., Portland, OR 97209, USA*
[b]*Center for Research in Mathematics and Science Education, San Diego State University, USA.*
[c]*Department of Chemistry & Biochemistry, San Diego State University, USA.*
E-mail: rkomperda@sdsu.edu

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Assessment instruments that generate quantitative data on attributes (cognitive, affective, behavioral, etc.) of participants are commonly used in the chemistry education community to draw conclusions in research studies or inform practice. Recently, articles and editorials have stressed the importance of providing evidence for the validity and reliability of data collected with these instruments following guidance from the Standards for Educational and Psychological Testing. This study examines how quantitative instruments have been used in the journal *Chemistry Education Research and Practice* (*CERP*) from 2010–2021. Of the 369 unique researcher-developed instruments used during this time frame, the majority only appeared in a single publication (89.7%) and were rarely reused. Cognitive topics were the most common target of the instruments (56.6%). Validity and/or reliability evidence was provided in 64.4% of instances where instruments were used in *CERP* publications. The most frequently reported evidence was single administration reliability (e.g., coefficient alpha), appearing in 47.9% of instances. Only 37.2% of instances reported evidence of both validity and reliability. These results indicate that, as a field, opportunities exist to increase the amount of validity and reliability evidence available for data collected with instruments and that reusing instruments may be one method of increasing this type of data quality evidence for instruments used by the chemistry education community.

Introduction

In STEM education research, significant resources are dedicated to the development and dissemination of evidence-based instructional innovations, including curricular resources and instructional practices. A primary goal of classroom-based innovations is the creation of learning environments that support learning for all students, with the purpose of broadening participation in STEM. In any research discipline, including chemistry education research (CER), researchers must generate research questions and data that can help address those questions. To address research questions and provide evidence of the efficacy of classroom-based innovations on students' knowledge, beliefs, and experiences, particularly in large-scale studies, chemistry education researchers often generate quantitative data using assessment instruments.

In CER, assessment instruments can be used to generate data related to cognitive knowledge of chemistry topics (e.g., concept inventories, American Chemical Society exams, instructor-developed exams, or other measures of chemistry content knowledge), affect (e.g., measures of student attitudes, beliefs, etc.), and behavior (e.g., observational protocols of student [or instructor] actions). Broadly defined, assessment instruments are tools used in social science research to quantitatively measure psycho-social attributes of research participants. While the psycho-social attributes that education researchers often seek to measure, such as knowledge and beliefs, are not directly observable, assessment instruments (henceforth referred to as *instruments*) allow researchers to collect data related to non-directly-observable attributes; these attributes are described as *latent traits* or *constructs* (American Educational Research Association *et al.*, 2014; Wu *et al.*, 2016) To support the use of quantitative data generated through the use of measurement instruments designed to measure latent traits, authors should provide "evidence that supports the interpretation and use of the data" for its intended purposes (Lewis, 2022, p. 1).

Quantitative Measurement in CER

"*The ability to answer a research question is only as good as the instrument(s) used to gather the research data. High-quality instruments improve the ability to answer research questions, while low-quality instruments impede research*" (Arjoon *et al.*, 2013, p. 536).

Because of the important role of instruments in generating research data and supporting the CER community's goals for improving educational experiences for students, it is important to understand how the research community uses and evaluates the quality of instruments and instrument-generated data (Lewis, 2022; Stains, 2022). In CER, both the international and United States flagship journals (this journal and the *Journal of Chemical Education*, respectively) require studies using instruments to provide evidence

of the validity and reliability of instrument-generated data (Towns, 2013; Seery *et al.*, 2019). Additionally, a number of chemistry education scholars have advocated for field-wide adoption of best practices in measurement (Barbera and VandenPlas, 2011; Arjoon *et al.*, 2013; Komperda *et al.*, 2018; Taber, 2018; Barbera *et al.*, 2020; Rocabado *et al.*, 2020). While field- and journal-specific recommendations provide guidance for researchers wishing to provide validity evidence for instrument-generated data, there is no single correct approach to doing so. Reviews in CER have highlighted the variation in researchers' approach to gathering validity evidence and presenting it in published work (Arjoon *et al.*, 2013; Deng *et al.*, 2021). This research adds to that body of literature by examining trends in instrument development, use, and evaluation in *Chemistry Education Research and Practice* over a twelve-year period (2010 – 2021).

Conceptual framework

The analyses described in this review study were informed by a framework for the development and evaluation of tests (instruments). The Standards for Educational and Psychological Testing is generally considered to be the primary contemporary source on best practices in educational measurement. The Standards were originally developed and published by The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) in 1966 and were most recently revised in 2014. In chemistry education research, the Standards have informed authors' advocation for best practices in instrument development and evaluation. The Standards also informed a prior study on the state-of-affairs for instrument use and evaluation in CER (Arjoon *et al.*, 2013).

Before describing the details of the current study, it is necessary to describe the terminology used to describe the quality of data collected from instruments, typically categorized as validity or reliability evidence. Additionally, this section addresses item difficulty and discrimination values, which are not considered by the Standards to be sources of validity or reliability evidence but can provide important psychometric information for instrument developers and users. Our conceptual framework uses definitions and operationalizations from the Standards and Arjoon *et al.* (2013), including five types of validity evidence (evidence based on test content, response processes of respondents, internal structure, relations to other variables, and consequences of testing), three categories of reliability evidence (test-retest coefficients, single-administration coefficients, and other less frequently used estimates), and the role of item difficulty/discrimination in CER.

Validity

According to the Standards, validity "refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association *et al.*, 2014, p. 11). The Standards describe a unitary model of validity in which all of the many types of validity evidence are contributors to an overarching construct validity argument, replacing older language describing distinct types of validity (e.g., face validity, concurrent validity, etc.) (Cronbach, 1988; Messick, 1995). Importantly, validity is a property of data and interpretations from them, and instruments themselves cannot be "validated". Rather, validity evidence from multiple data sources may be used to support the argument that data may be validly interpreted. Variety in the design and purpose of instruments means that some sources of validity evidence are more appropriate and meaningful in some contexts than others, and it is not expected that all types of validity evidence are presented in all contexts. However, the Standards do dictate that validity evidence should be presented by both test developers and test users, across the entire lifetime of an instrument to support valid interpretations from instrument-generated data (American Educational Research Association *et al.*, 2014).

Evidence Based on Test Content

The Standards describe validity evidence based on test content as "analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores." (p. 14). Typically, this type of validity evidence comes from domain experts' judgment of the appropriateness of the "themes, wording, and format of the items" that comprise an instrument for sufficiently and accurately measuring the target domain(s) (Arjoon *et al.*, 2013; American Educational Research Association *et al.*, 2014, p. 14).

Evidence Based on Response Processes

The Standards describe validity evidence based on response processes as "evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (American Educational Research Association *et al.*, 2014, p. 15). Typically, evidence related to response processes is collected via think-aloud interviews where respondents detail their cognitive processes, including item interpretation, approaches to answering, and other reasoning strategies, as they engage with items; in a recent review of reported approaches to collecting evidence based on response processes, Deng *et al.* (2021) discuss alternative methods to collective response process validity evidence, including administration of open-ended versions of items, focus groups, and eye-tracking studies.

Evidence Based on Internal Structure

Validity evidence related to the internal structure of an instrument refers to evidence that the items comprising an instrument are related to one another and/or consistent with the conceptual framework for the instrument. Instruments may be designed to measure a single construct or multiple constructs that are theoretically (un)related. The investigation of item interrelationships can provide evidence that respondents interacted with items as intended. Item interrelationships may be investigated using multiple approaches, including factor analysis, principal component analysis, or item response theory (IRT) modeling (Arjoon *et al.*, 2013; American Educational Research Association *et al.*, 2014).

Evidence of measurement invariance refers to evidence that item interrelationships do not differ across subgroups (e.g., racial, ethnic, gender groups, experimental conditions). Differential Item Functioning across subgroups is known as DIF. Usually, DIF represents the measurement of an unintended dimension and is a threat to the validity of interpretations from instrument-generated data, though sometimes DIF is anticipated (American Educational Research Association *et al.*, 2014; Rocabado *et al.*, 2020). In our analysis, we code evidence of measurement invariance/DIF separately from other types of evidence based on internal structure (e.g., factor analyses).

Evidence Based on Relations to Other Variables

Evidence that scores derived from instrument data are related (or unrelated) with other variables (e.g., scores on a separate instrument) may also support an overall validity argument. Particularly when scores derived from an instrument are expected to

be correlated (or not) with other variables, "evidence based on relationships with other variables provides evidence about the degree to which these relationships are consistent" with the construct that an instrument is designed to measure (American Educational Research Association *et al.*, 2014, p. 16). Evidence based on relations to other variables encompasses several older types of validity evidence, including: convergent validity (instruments measure the same or similar targets), divergent validity (instruments measure different targets), predictive validity (instrument-generated data is useful for predicting future outcomes), and concurrent validity (agreement between two or more instruments).

Consequences of Testing

The final type of validity evidence described by the Standards is evidence that decisions that are made based on score interpretations from instrument-generated data are appropriate and warranted (American Educational Research Association *et al.*, 2014). In CER, consequence testing is rarely reported as evidence of validity, and we did not observe any instances of consequence testing in the data for this study. While our codebook was designed to include consequence testing as a type of validity evidence, based on the Standards, it will not be further discussed in the analysis.

Reliability

The Standards describe reliability of data as the "general notion of consistency of the scores across instances of the testing procedure" (American Educational Research Association *et al.*, 2014, p. 33). Reliability is also described as precision of measurement (Komperda *et al.*, 2018), and, like validity, is a property of data, not of instruments. Unreliable data are a threat to valid interpretation of data, and therefore, evidence of reliability should be reported along with validity evidence. Commonly, instrument developers and users report estimates of data reliability, termed reliability coefficients in the Standards. Our conceptual framework, adopted from the Standards and Arjoon *et al.* (2013), includes two main types of reliability coefficients: test-retest and single administration. There are other approaches to estimating the reliability of data, for instance, inter-rater reliability. While our coding structure was designed to account for multiple approaches to reliability estimation, we discuss only two here because we observed very few or zero instances of these other approaches.

Test-Retest Reliability Coefficients

The most straightforward approach to estimating reliability and providing evidence that respondents interact with items consistently is to administer the instrument to the same group of respondents multiple times. This approach to estimating reliability assumes that no treatment or other external factors will have changed respondents' interactions with items between administrations. Most typically, test-retest reliability is estimated by correlating respondents' scores on two separate administrations of the same set of items.

Single Administration Reliability Coefficients

Because it is not always practical or feasible to administer an instrument to a group of respondents multiple times, reliability may be estimated by determining the consistency of responses to items that are expected to measure the same target domain. Single-administration reliability can be estimated using a number of coefficients, including: coefficient alpha, McDonald's omega, coefficient H, Kuder-Richardson 20 and 21 (KR-20 and KR-21), split-halves correlation, and person separation. Though single administration reliability is often described as internal consistency reliability, there is concern about the accuracy of the term (Komperda *et al.*, 2018; Barbera *et al.*, 2020). We refer readers interested in the appropriate use of single administration reliability coefficients in CER to Komperda *et al.* (2018).

Other Reliability

Additional approaches to estimating the reliability of data generated using instruments are described in the Standards, for example, Generalizability Theory (G-theory); we do not discuss G-theory approaches to reliability estimation in this study because we did not observe any instances of its use. Additionally, qualitative data might be coded by multiple raters and interrater reliability estimated using percent agreement or Cohen's kappa (American Educational Research Association et al., 2014); this interpretation of "reliability" refers to the consistency of application of a scoring or coding structure, but not to the consistency of scores on an instrument. Therefore, we do not include interrater/intercoder reliability in this analysis. These approaches to estimation of reliability will not be further discussed.

Difficulty/Discrimination

Some authors present additional types of analyses that are described by the authors as evidence of validity and reliability but which are not described as such in the Standards (American Educational Research Association *et al.*, 2014). The distinction arises as the Standards discuss validity and reliability at the instrument score level while difficulty and discrimination are properties of individual items. Most commonly, CER authors reported difficulty and/or discrimination values (which can be generated using methods from either Classical Test Theory or IRT) as evidence that items comprising an instrument are functioning as intended. Because difficulty and discrimination are not identified by the Standards as evidence for validity and/or reliability, we discuss these indices separately from our discussion of validity and reliability.

Rationale for the Current Study

In 2013, Arjoon and colleagues examined the ways that researchers used and evaluated instruments in studies published in the *Journal of Chemical Education*; the authors observed that CER researchers have adopted some evaluation practices aligned with the Standards and suggested that psychometric evaluation practices in CER could be improved and diversified to include validity and reliability evidence from multiple sources. Deng *et al.* (2021) highlighted the variation in researchers' approaches to collecting and reporting data related to respondents' response processes when engaging with instruments. Others have compiled lists of instruments and/or data quality evidence in manuscripts (Barbera and VandenPlas, 2011) or websites (Bretz Research Group, 2022; PhysPort, 2022). However, to our knowledge, no studies have systematically investigated the use and evaluation of instruments in published studies in CER in the last decade; additionally, this study investigates the use and evaluation of both newly developed and adapted or adopted instruments.

Chemistry Education Research and Practice (*CERP*) is a peer-reviewed international academic journal published by the Royal Society of Chemistry and publishes articles which "inform readers about some aspect of teaching and learning chemistry" (Seery *et al.*, 2019, p. 335). *CERP* was selected for this study as it is freely accessible, all articles published in *CERP* are expected to include typical components of research articles (description of methods, results, connections to prior literature) as well as implications for the practice of teaching chemistry, and *CERP* publishes a "diverse range

of contributions from all over the world," (p. 338); therefore it is expected to broadly represent research and research practices in chemistry education for the purposes of this study. In this study, we present an investigation of the adoption of best practices in measurement and evaluation, as described by the Standards, in a census of journal articles published in *CERP* from 2010 to 2021. The study was guided by the following research questions:

1) How are measurement instruments used and/or evaluated in studies reported in *CERP*?
2) To what extent do CER researchers provide psychometric evidence for instrument data, as reported in *CERP*?

Methods

Pre-registration

This study was pre-registered. Prior to conducting this study, we submitted our methods, codebook, and data analysis plan to the Open Science Framework (OSF) repository. This pre-registration is publicly available at https://osf.io/cq43f.

Sampling

From Scopus, we generated a list of all articles published in *CERP* between January 2010 and October 2021 (N = 767); because *CERP* publishes quarterly, this list represents a census of all articles published in *CERP* in the years 2010 to 2021.

Criteria for inclusion in study

Each article published in *CERP* between 2010 and 2021 was screened for inclusion in the study. This study investigates the use of assessment instruments in CER, and therefore all articles in the sample that meaningfully use assessment instruments which can be used to generate quantitative data were included in the study (n = 296). Our criteria for meaningful use of assessment instruments include the following:

Articles considered for inclusion in the study must meet any of the following criteria and may meet multiple criteria:

1. The article reports the development of a novel instrument. [**Original**]
2. The article reports the modification of an existing instrument. [**Modification**]
3. The article reports the use of a novel or existing instrument; the data generated by the instrument is used to address research question(s). [**Use**]
4. The article reports evidence related to the quality of data generated using a novel or existing instrument. [**Evaluation**]

Additionally, instruments must generate quantitative data or qualitative data that may be scored quantitatively (e.g., assigned a numeric value according to a coding structure described by the article's authors). Articles that do not meaningfully use assessment instruments and articles using instruments that generate only qualitative data were not included in the study. Studies using qualitative data only were excluded for the purposes of limiting the scope of this study and because in the qualitative tradition, the "researcher may be considered the instrument" whereas instruments that generate quantitative data may be considered separately from the researchers conducting the study (Arjoon *et al.*, 2013, p. 536).

We recorded articles which used standardized exams as data collection instruments (e.g., ACT, SAT, American Chemical Society [ACS] Exams), but we do not include these instances in our analysis for this study as the generation of psychometric evidence for those instruments is generally conducted and reported by the publisher, not individual researchers. We have included a parallel analysis that includes these instruments in the Appendix.

Coding

For articles which met criteria for inclusion in the study, we coded each article for variables of interest. All data were entered into a Microsoft Access database that was developed by the research team. Variables were recorded on three levels:

1. Publication-level variables ($N_{publications}$ = 292)
2. Instrument-level variables ($N_{instruments}$ = 369)
3. Publication-instrument-level variables ($N_{publication\text{-}instrument}$ = 430)

Variables were recorded separately at three levels because some variables of interest apply to only publications (e.g., title of the article) or only instruments (e.g., name of the instrument); publication-instrument-level variables were used to record the relational aspects of publications and instruments (e.g., evidence of validity and reliability of instrument-generated data in a specific study).

We developed a coding protocol based on the Standards' descriptions and definitions of 1) the process of instrument development, evaluation, and use and 2) validity and reliability evidence (American Educational Research Association *et al.*, 2014); because these definitions have already been operationalized for use in CER, our coding protocol was also adopted from Arjoon *et al.* (2013). The definitions and operationalization of terms and concepts used in this study were largely unchanged between the 1999 and 2014 versions of the Standards, and therefore can be fairly applied to the analysis of publications within the study's focal timespan (2010 – 2021). The codebook definitions follow the conceptual framework described previously in this manuscript and the complete codebook is available in the OSF pre-registration materials.

Analysis

We used the R software (Version 4.2.0) for all analyses and figure generation (R Core Team); R code and data are provided with the OSF materials. Because we did not have *a priori* hypotheses regarding our research questions, we report descriptive statistics related to the use of assessment instruments and evaluation of data obtained from the instruments in CERP articles. Our results include descriptive analyses to address our research questions.

Inter-rater reliability study

Multiple researchers, including the entire author team, have coded to consensus on more than 200 publications and the instruments described in them, which has resulted in a robust coding procedure that can be consistently applied by the research team. Additionally, a subset of the publications included in this study (10%) were coded by both the first and second authors. Because our coding structure entails the application of as many codes as necessary to capture all the nature of instruments' appearance in publications and the types of evidence of validity and reliability (a one-to-many coding scheme), we calculated the Fuzzy kappa statistic as the index of inter-rater reliability (Kirilenko and Stepchenkova, 2016). Fuzzy kappa is based on Cohen's kappa and modified for use with one-to-many coding structures. The calculated value of inter-rater reliability (0.97, Fuzzy kappa) provides substantial evidence of reliability.

Results

In the 292 publications that met inclusion criteria for the study, we observed 369 unique researcher-developed instruments and 430 publication-instrument instances. Some instruments appear in multiple publications, and some publications report administration of multiple instruments.

The results presented here include analysis of data primarily at the publication-instrument level because we conceptualize publication-instrument instances as a proxy for the measurement targets of CER as a field as well as an indication of the extent to which best practices for reporting data quality evidence with each instrument administration are being followed.

Research Question 1: How are measurement instruments used and/or evaluated in studies reported in CERP?

RQ 1.1 Which researcher-developed instruments (if any) are commonly used in studies reported in CERP?

To gain a general sense of how instruments are used in CER, we investigated which instruments were administered in multiple studies. Most researcher-developed instruments which appeared in multiple publications ($n \geq 3$) were designed to measure 1) affective constructs such as: attitude, self-efficacy, interest, and motivation, and 2) students' process skills such as: information processing, scientific or logical reasoning, and visio-spatial thinking (Table 1) (Tobin and Capie, 1984; Geban *et al.*, 1992; Dalgety *et al.*, 2003; Bauer, 2008; Stamovlasis, 2010; Xu and Lewis, 2011; Cooper *et al.*, 2012; Ferrell and Barbera, 2015; Ardura and Pérez-Bitrián, 2018; Galloway and Bretz, 2015).

Other instruments not targeting affective constructs or process skills which appeared in three publications include the Reformed Teaching Observational Protocol (RTOP), the Students' Understanding of Models in Science (SUMS), and the Students' Assessment of Learning Gains (SALG) (Piburn *et al.*, 2000; Seymour *et al.*, 2000; Treagust *et al.*, 2002). All other instruments appeared in only one (n = 331) or two (n = 23) publications; most commonly, instruments observed twice appeared in two publications by the same author(s) (n=17). The observed count of all publication-instrument instances can be found in the provided data and R code.

We observed that of the instruments that appeared in three or more publications, only the Implicit Information from Lewis Structures Instrument (Cooper *et al.*, 2012) assessed chemistry-specific cognitive knowledge/skills while the Meaningful Learning in the Laboratory Instrument (Galloway and Bretz, 2015) bridges the affective and self-reported laboratory skills domain. We found this observation curious, given that the data for this study are publications in a chemistry education-specific journal, though potentially reflective of practice. It may be that those publishing articles in CER feel more competent developing instruments to measure chemistry knowledge (cognitive) than instruments that measure constructs (e.g., affective constructs) that have been more extensively studied in adjacent fields (e.g., educational psychology); therefore, researchers are more likely to reuse instruments from adjacent fields, while they might develop novel instruments for cognitive measurement goals. Additionally, researchers might use ACS exams (or other instruments developed by testing organizations) for cognitive measurement targets; while these instruments were excluded from the remainder of our analyses, we do note that standardized exams were used as instruments in studies published in CERP with some frequency. ACS exams (any version) were administered in eleven studies, and student scores on the SAT from any year were used in nine studies. No other excluded instrument was used more than once. We have included results from parallel analyses, which include these excluded instruments, in the Appendix.

Table 1: Observed count of most-used researcher-developed instruments in *CERP* publications 2010 – 2021. Affective topics are shaded to differentiate them from process skills.

Researcher-Developed Instrument	Number of uses	Topic(s)
Attitude Toward the Subject of Chemistry v2	8	Affective – Attitudes
Attitude Toward the Subject of Chemistry	4	Affective – Attitudes
College Chemistry Self-Efficacy Scale – Cognitive Skills Scale	4	Affective – Self-Efficacy
Lawson's Classroom Test of Formal Reasoning (Greek translation)	4	Process skills – scientific reasoning
Chemistry Attitudes and Experiences Questionnaire	3	Affective – Attitudes, Self-Efficacy
Initial and Maintained Interest in Chemistry	3	Affective – Interest
Spanish translation of the Science Motivation Questionnaire II	3	Affective – Motivation
Meaningful Learning in the Laboratory Instrument	3	Affective – Attitudes Cognitive – Self-assessed laboratory skills
Scientific Process Skills Test (Turkish)	3	Process skills
Implicit Information from Lewis Structures	3	Process skills – Information processing
Test of Logical Thinking	3	Process skills – logical thinking
Group Embedded Figures Test (Greek translation)	3	Visio-spatial thinking
Reformed Teaching Observational Protocol	3	Teaching strategies/behaviors
Students' Understanding of Models in Science	3	Nature of science (models)
Students' Assessment of Learning Gains	3	Self-assessed learning gains

RQ 1.2 Which topics are commonly measured for studies reported in *CERP*?

In our analysis related to RQ 1.1, we observed that the most commonly administered researcher-developed instruments are those designed to measure affective constructs and process skills, and only one (IILSI) of the most common instruments is designed to measure chemistry-specific knowledge. Based on this observation, we investigated whether/to what extent the measurement targets in CERP studies also prioritized measurement of affective constructs and process skills.

In this analysis, we observed a mismatch between the measurement targets overall (Table 2) and those for the most commonly reused researcher-developed instruments (Table 1). The target construct for more than half of all publication-instrument instances ($N_{publication\text{-}instrument} = 430$) was coded as cognitive (54.9%, n = 236), while affective measurement targets represented only 34.2% of all publication-instrument instances (n = 147). Publication-instrument instances which were coded as Behavioral (9.8%, n = 42), Metacognition (3.3%, n = 14), Evaluation (3.0%, n = 13), and Nature of Science (0.5%, n = 2) represented small proportions of measurement targets. Less than five percent of instruments were

designed to measure targets not represented in our coding scheme. The proportions of unique instruments designed to measure each target topic domain are largely very similar (Table 3). See OSF materials for additional details on the coding scheme.

Though CER has historically focused on the cognitive domain, we were interested in investigating if research (measurement targets) has diversified over the last decade to include more research on the affective domain and other domains, such as behavior and metacognition. Therefore, we disaggregated the previous analysis by year of publication to examine any trends in measurement targets by topic/domain (Figure 1). We observed moderate diversification of the topics of measurement instruments, including a small increase in the count and proportion of publication-instrument instances measuring topics in the affective domain. Generally, there were no clear trends in measurement targets over time.

Table 2: Percent (count) of publication-instrument instances by topic ($N_{publication\text{-}instrument}$ = 430)

Topic	Percent publication-instrument instances (N = 430)
Cognitive	54.9% (n = 236)
Affective	34.2% (n = 147)
Behavioral	9.8% (n = 42)
Metacognition	3.3% (n = 14)
Evaluation	3.0% (n = 13)
Nature of Science	0.5% (n = 2)
Other	4.9% (n = 21)

Table 3: Percent (count) of unique instruments by topic ($N_{instruments}$ = 369)

Topic	Percent unique instruments (N = 369)
Cognitive	56.6% (n = 209)
Affective	31.2% (n = 115)
Behavioral	9.8% (n = 36)
Metacognition	3.5% (n = 13)
Evaluation	3.0% (n = 11)
Nature of Science	0.5% (n = 2)
Other	4.9% (n = 18)

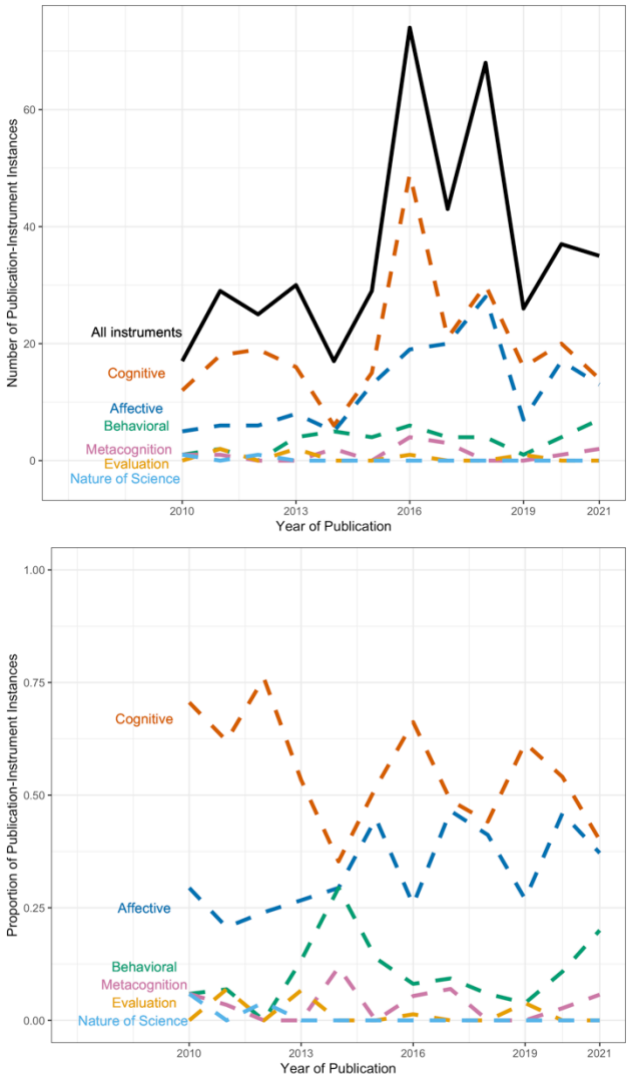

Figure 1: Researcher-developed instruments in publication-instrument instances (N = 430) by topic by year as count (top) and proportion (bottom)

RQ 1.3 What is the ratio of CERP studies using instruments previously developed by other researchers relative to the development of novel instruments for their research purposes?

Data addressing RQ 1.1 suggest that researcher-developed instruments are often developed or modified for the purposes of a specific study, rarely used again, and very infrequently administered by researchers other than the original authors. To further support this claim, we investigated the nature of the publication-instrument instances by coding for the nature of the instrument appearance in publications.

Codes and code definitions are defined by the inclusion criteria in the Methods section. All applicable codes were applied to publication-instrument instances; because meaningful use of assessment instruments, operationalized by this coding structure, was a criterion for inclusion in the study, all publication-instrument instances were assigned at least one code. The codes are not mutually exclusive categories and publication-instrument instances could receive multiple codes. Figure 2 represents the extent of the

overlap between the codes. As seen in the top of Figure 2, almost all publication-instrument instances were coded as Use. The code for Modification is not specifically visualized because it is always a subset completely contained within Original as all Modifications of one instrument also represent the Original publication for the new modified instrument.

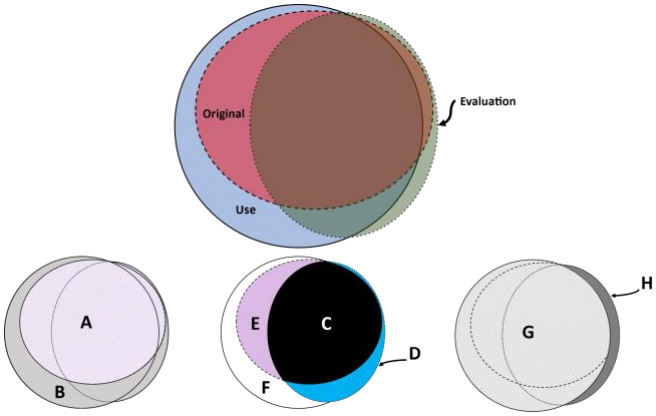

Figure 2: Visualization of overlap between publication-instrument codes. (A) Use or evaluation of original instruments; (B) Use or evaluation of pre-existing instruments; (C) Evaluation of original instruments; (D) Evaluation of preexisting instruments; (E) Original instruments used without evaluation; (F) Pre-existing instruments used without evaluation; (G) Instruments used to address research questions; (H) Instruments evaluated only

Of the 430 publication-instrument instances, 70.7% are indicated as an Original instrument (n = 304; Figure 2 – Section A); this indicates that researchers developed a novel instrument for most quantitative measurement goals in *CERP* studies. Some of these Original instruments, 41.4% (n = 126), were modified by the researchers from existing instruments for the purposes of their study (not shown in Figure 2). Of the 292 publications included in the study, 80.5% included at least one Original instrument. Together, these data may indicate that researchers are interested in studying and measuring constructs for which appropriate instruments do not already exist and/or researchers are unaware that potentially useful instruments already exist, so they develop new instruments for their research purposes.

We also investigated the co-occurrence of codes for the nature of the publication-instrument instances ($N_{publication\text{-}instrument}$ = 430). For 64.4% (n = 277) of publication-instrument instances, authors provided some evidence related to the validity and/or reliability of instrument-generated data, shown in the top of Figure 2 as Evaluation. Authors were more likely to provide evidence of data validity or reliability for Original instruments (71.1%; n = 216; Figure 2 – Section C) than pre-existing instruments (48.4%; n = 61; Figure 2 – Section D). Nearly all observed publication-instrument instances involved the use of instrument-generated data to address research question(s) (Use; n = 400; 93.0%; Figure 2 – Section G); this is unsurprising, as a primary purpose of instruments in research is the measurement of constructs through the generation of data. Of the studies that used instrument-generated data to address research questions, 37.8% (n = 151; Figure 2 – Sections E and F) provided no evidence of data validity or reliability; these instances where instruments were used to generate data without evaluation included both Original (n = 86; Figure 2 – Section E) and pre-existing (n = 65; Figure 2 – Section F) instruments.

Publication-instrument instances which were not coded as Use (Figure 2 – Section H) include the studies which investigated the psychometric properties of newly developed (n = 24) and pre-existing (n = 6) instruments but did not use instrument-generated data to address other research questions. While field- and journal-specific recommendations (Towns, 2013; Seery *et al.*, 2019) and documents like the Standards (American Educational Research Association *et al.*, 2014), provide some guidance for researchers on how to collect and report validity and reliability evidence for instrument-generated data, there is no single approach to doing so. In the above analysis, the publication-instrument instances that were coded as Evaluation varied considerably in the types and amount of evidence presented. We investigated this variation in the following analyses.

Research Question 2: To what extent do CER researchers provide psychometric evidence for instrument data, as reported in *CERP*?

RQ 2.1 To what extent is data quality evidence reported in *CERP* studies? When data quality evidence is reported, what kind of evidence is typically reported?

In our analysis for RQ 1.3, we observed that approximately two-thirds of all publication-instrument instances ($N_{publication\text{-}instrument}$ = 430) reported some evaluation of data quality evidence. There is no standard approach to collecting and reporting validity and reliability evidence, and for this analysis, we recorded the types of validity and reliability evidence reported in *CERP* studies. Our coding structure was developed based on the Standards, described in the conceptual framework; code definitions are also included in the OSF materials.

Authors most commonly reported validity evidence based on internal structure (24.9%), test content (24.4%), and relations with other variables (23.0%). Validity evidence based on response processes (12%) was less commonly reported. While evidence based on internal structure (e.g., factor analysis, principal component analysis, Rasch analysis) was reported in nearly a quarter of all studies, authors rarely reported evidence of measurement invariance or DIF (3.0%). We observed no instances of consequence testing, and therefore, consequence testing does not appear in our analysis. Nearly half of all studies (three-quarters of all evaluations) reported a single administration reliability coefficient (e.g., coefficient alpha, McDonald's Omega). Test-retest reliability was reported much less frequently (Table 4). G-theory approaches to reliability estimation were not observed and are therefore not reflected in this analysis.

The role of validity evidence is to justify the use of instruments for specified purposes. Because of the variety of instruments and circumstances in which they are used, it is "natural that some types of evidence will be especially critical in a given case, whereas other types will be less useful" (American Educational Research Association *et al.*, 2014, p. 12). However, multiple, complimentary types of validity and reliability evidence support the proposition that conclusions from instrument-generated data are trustworthy. In this study, half of instrument administrations were used without any reported evidence of validity (n = 212). Of those studies that did provide some validity evidence (n = 218), more than half reported only one source of validity evidence (n = 118). One hundred studies reported two or more complementary sources of validity evidence.

Similarly, we observed no reported evidence of reliability for more than half of the publication-instrument instances (n = 221). Because validity and reliability are complimentary constructs, the Standards suggest that evidence of both are required to support the

trustworthiness of conclusions based on instrument-generated data. In this study, just 37.2% (n = 160) of publication-instrument instances reported evidence of both validity and reliability. Some reported evidence of validity and no evidence of reliability (n = 58); others reported only reliability coefficients without evidence of validity (n = 49).

We also observed that difficulty and discrimination indices were reported in some studies as evidence of data quality. These indices are not considered evidence of either validity or reliability in our conceptual framework, but they are somewhat commonly reported in CER literature. These values can help inform test developers and users about the extent to which items are functioning as intended with the population being measured by the instrument. The Standards describe item difficulty and discrimination values as part of the item screening process that contributes to overall instrument development. Difficulty indices (based on either Classical Test Theory [CTT] or IRT/Rasch) were reported in 7.9% of studies (n = 34); discrimination indices were reported in 6.7% (n = 29) studies.

Table 4: Percent (count) of publication-instrument instances reported types of validity and reliability evidence, excluding standardized exams

Type of Psychometric Evidence	**Percent of publication-instrument Instances (N = 430)**
Evaluation (Any)	64.4% (n = 277)
Validity (n = 218)	
Internal Structure (Factor Analysis/Principal Components Analysis/Rasch Analysis)	24.9% (n = 107)
Test Content	24.4% (n = 105)
Relations with Other Variables	23.0% (n = 99)
Response Processes	12% (n = 50)
Measurement Invariance/DIF	3.0% (n = 13)
Reliability (n = 209)	
Single Administration Reliability	47.9% (n = 206)
Retest Reliability	1.4% (n = 6)

2.2 What trends exist in the data quality evidence reported?

In acknowledgement of the relative youth of CER as a discipline and a push by field leaders to improve the quality of research, including some calls for improved practice in instrument evaluation, we investigated the types of validity and reliability evidence for researcher-developed instruments presented over the 12 years of *CERP* included in this study (Figure 3). We expected that instrument evaluation practices might improve with the maturity of CER as a discipline.

Starting with validity evidence that is typically collected using qualitative methods (test content and response process), no clear trend was observed over time. We did observe trends in reported quantitative validity evidence. There was an increase in the reported use of factor analysis (and similar methods) to investigate the internal structure of instruments; similarly, measurement invariance has been reported as a source of validity evidence more frequently since 2019. Of note, in 2019, *CERP* published an editorial intended "to provide guidance on submitting manuscripts" to the journal, which formally set the expectation that authors include evidence related to validity and reliability in studies using instruments to generate quantitative data (Seery *et al.*, 2019, p. 355). We expect that practices will continue to improve to meet this standard, and we see some evidence that this is the case already (e.g., an observed uptick in measurement invariance/DIF).

The second, and more salient, observed trend is the parallel between the proportion of publication-instrument instances that are coded as Evaluation and Single Administration Reliability. Others have criticized researchers' overreliance on single administration reliability coefficients, in particular, coefficient alpha (Komperda *et al.*, 2018; Taber, 2018; Barbera *et al.*, 2020). An overwhelming majority of the publication-instrument instances ($N_{publication\text{-}instrument}$ = 430) that are coded as Single Administration Reliability (n = 206) reported coefficient alpha (n = 164). While coefficient alpha can be used to estimate reliability of data generated using instruments, we encourage readers to use alpha (and other reliability estimates) only when mathematical assumptions are met and not as a substitution for validity evidence. A plot of difficulty and discrimination indices by year can be found in the Appendix.

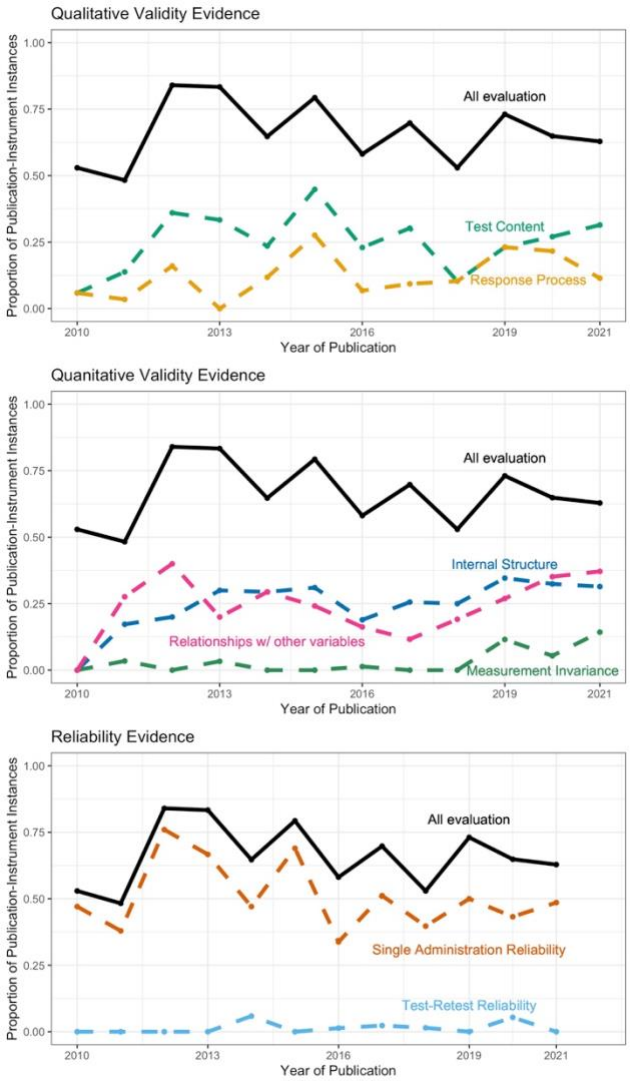

Figure 3: Proportion of publication-instrument instances (N = 430) that reported types of validity evidence, either qualitative (top) or quantitative (middle) or reliability evidence (bottom) by year.

Limitations

In this study, we report our observations from a census of articles published in a single journal, *Chemistry Education Research and Practice* (*CERP*), from 2010 to 2021. Though *CERP* publishes research articles from authors all over the world, our observations may not represent practices in CER overall. Additionally, our analysis was limited to the information that authors chose to report in their published articles. It may be the case that authors did not choose to publish all evidence related to validity and reliability of instrument-generated data.

Conclusions and Recommendations

We reviewed 292 articles published in CERP that used researcher-developed instruments to generate data; the sheer number of publications using instruments to generate data and the production of novel instruments since 2010 indicates that CER as a field is interested in quantitative measurement of latent traits. Note the substantial increase in instruments since the analysis by Arjoon et al. (2013) found 20 new instruments in 37 articles in the *Journal of Chemical Education* between 2002-2011. We observed that existing instruments published in the literature are rarely reused, and authors typically develop new instruments or modify existing instruments (thereby creating a new instrument) for their purposes. Additionally, we observed that instruments (both new/modified and existing) are often used without evaluation of data quality in the context of the study. Direct comparison with the Arjoon *et al.* (2013) study is not possible since they focused on what was reported by instrument developers, and our work included other researchers using existing instruments who may not have been as aware of the need to report validity and reliability evidence. As a result, our study has lower reported rates of psychometric evidence overall, but some trends in the type of evidence reported are similar.

We recorded the types of validity and reliability evidence that authors reported to support their inferences based on instrument-generated data. Our analyses suggest an overreliance on single administration reliability estimates (specifically, coefficient alpha) as evidence of data quality and demonstrate that researchers rarely present multiple complimentary sources of validity evidence. Our finding that roughly 50% of the publication-instrument instances report alpha is consistent with findings from Arjoon *et al.* (2013) where 75% of instruments examined reported alpha, making it much more prevalent than test-retest reliability. For validity, our analysis found roughly equivalent use of relations with other variables, test context, and internal structure evidence in approximately 25% of the publication-instrument instances. This is noticeably different from the Arjoon *et al.* (2013) results which found mostly relations with other variables (95%), followed by test content (55%), and internal structure (45%).

Though we noted the absence of validity evidence as it relates to consequences of testing, overall, practices in reporting evidence of data quality appear to be improving, perhaps in response to recommendations from field leaders in measurement (Arjoon *et al.*, 2013) and explicit expectations from both *CERP* and the *Journal of Chemical Education* (Lewis, 2022; Seery *et al.*, 2019; Stains, 2022; Towns, 2013). Based on our findings, we make the following recommendations:

Recommendation 1: Consider using instruments that have already been developed and published alongside evidence of validity and reliability.

We recorded 369 unique researcher-developed instruments in our analysis of 296 publications over twelve years (2010 – 2021) in *Chemistry Education Research and Practice*. Because of the nature of research, it is inevitable that sometimes researchers will have measurement goals that require the development of new instruments. However, we encourage researchers to consider instruments that have already been developed and evaluated for their research purposes. The use of existing instruments, where appropriate, will contribute to the body of evidence for validity and reliability of instrument-generated data and will allow for the allocation of research resources to endeavors other than development of redundant instruments. Other studies which have investigated instrument use and evaluation practices have also recommended that researchers consider extant instruments and contribute to the body of evidence supporting instruments' use across contexts (Blalock *et al.*, 2008; Arjoon *et al.*, 2013).

To support researchers (and practitioners) in choosing among the many extant instruments, this research team and our colleagues have developed an online resource, the Chemistry Instrument Review and Assessment Library (CHIRAL; Barbera *et al.*, 2022). CHIRAL can be accessed at https://chiral.chemedx.org/ and has been populated with information about instruments, the publications that instruments appear in, and published validity and reliability evidence, including for all instruments identified in this study.

Recommendation 2: Collect and publish evidence of validity and reliability in all studies that base conclusions on instrument-generated data.

In this study, we observed that researchers were more likely to present evidence of data validity and reliability when using novel instruments for data generation, compared to when using existing instruments. We encourage the field to always evaluate data for validity and reliability, which is aligned with the notion that evaluation of instruments is the responsibility of both instrument developers and users (American Educational Research Association *et al.*, 2014; Lewis, 2022; Stains, 2022). We emphasize that some approaches for evaluating the quality of data are inappropriate or impossible in some cases; for example, it would be inappropriate to perform factor analyses with very small datasets. However, researchers should consider sources of validity evidence that are appropriate for their research contexts; for example, conducting response process interviews with students from the target population is both possible and appropriate for studies with small datasets. Collecting evidence of data quality should be considered from the outset of a study and included in the study design (Stains, 2022), and multiple resources exist to support researchers in collecting and making sense of data quality evidence (Arjoon *et al.*, 2013; Komperda *et al.*, 2018; Rocabado *et al.*, 2020; Deng *et al.*, 2021). A field-wide commitment to collection and publication of data quality evidence for all studies will support the trustworthiness and the impact of research in chemistry education.

Recommendation 3: Include information on the collection of data quality evidence in a detailed methods section

One limitation of this study is that our analyses and codes were constrained by the information that researchers opted to include in their published articles. During data collection, we found that researchers adopt a range of approaches for describing their efforts

to evaluate instruments and instrument-generated data. Sometimes, our data collection was complicated by the scattering of validity and reliability evidence throughout multiple sections of an article or inclusion of this evidence only in the supplementary information, without mention in the body of the article. Sometimes, authors reference prior evaluation efforts ambiguously, and we found it difficult to distinguish between discussion of prior evaluation efforts and the authors' own efforts. If authors are relying exclusively on prior evaluation efforts to support their case that data are valid and reliable (which we do not recommend), this should be clearly stated.

It is possible that some researchers opted to not include relevant data quality evidence in their published work due to space constraints or other research, and therefore we (and future instrument users) are unaware of these efforts. We encourage researchers to explicitly and intentionally include details on their approaches to evaluating instruments and instrument-generated data in methods sections and for reviewers and editors to recommend this inclusion. If these details must be presented in supplementary materials, authors should direct the interested reader to those materials.

Additionally, the language around validity and reliability has changed over time and is often ambiguous. This complicated our efforts to code the type of data quality evidence presented in publications. Authors should consider adopting formal terms, for example from the Standards, as in this study and others (Arjoon *et al.*, 2013; American Educational Research Association *et al.*, 2014), which will support more universal understanding of methods and approaches to evaluation. Additionally, authors should include all relevant details about the target population(s) in their studies, including participant characteristics (age, course level) and context (language used in the classroom, country in which the study was conducted). The inclusion of such relevant details can support readers' interpretations and evaluation of the relevance of research relative to other contexts (Stains, 2022).

Summary

In this study, we report trends in the use and evaluation of instruments as quantitative data collection tools, based on our analysis of all articles published in *Chemistry Education Research and Practice* over more than a decade. We believe that trends point to an ever-higher standard of quality for research in chemistry education, and we have made recommendations based on our observations and analyses to support the continued improvement of the quality of research in the field.

Author Contributions

Contributions listed as informed by CRediT (2022).

KL: Conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing – original draft, and writing – review and editing.

KT: Investigation and writing – review and editing.

TAM: Investigation, writing – review and editing, and software.

RK: Conceptualization, funding acquisition, methodology, supervision, and writing – review and editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. (1914996).

Appendices

Appendix 1. Counts of instruments in *CERP* publications 2010 – 2021 including standardized exams

Table 5: Observed count of most-used instruments in CERP publications 2010 – 2021, with no exclusions from the dataset. Instruments excluded from primary analysis are shaded

Researcher-Developed Instrument	Number of uses	Topic(s)
American Chemical Society Exams – Any	11	Cognitive knowledge of chemistry topics (varies)
SAT	9	Cognitive knowledge of multiple topics
Attitude Toward the Subject of Chemistry v2	8	Affective – Attitudes
Attitude Toward the Subject of Chemistry	4	Affective – Attitudes
College Chemistry Self-Efficacy Scale – Cognitive Skills Scale	4	Affective – Self-Efficacy
Lawson's Classroom Test of Formal Reasoning (Greek translation)	4	Process skills – scientific reasoning
Chemistry Attitudes and Experiences Questionnaire	3	Affective – Attitudes, Self-Efficacy
Initial and Maintained Interest in Chemistry	3	Affective – Interest
Spanish translation of the Science Motivation Questionnaire II	3	Affective – Motivation
Meaningful Learning in the Laboratory Instrument	3	Affective – Attitudes Cognitive – Self-assessed laboratory skills
Scientific Process Skills Test (Turkish)	3	Process skills
Implicit Information from Lewis Structures	3	Process skills – Information processing
Test of Logical Thinking	3	Process skills – logical thinking
Group Embedded Figures Test (Greek translation)	3	Visio-spatial thinking
Reformed Teaching Observational Protocol	3	Teaching strategies/behaviors
Students' Understanding of Models in Science	3	Nature of science (models)
Students' Assessment of Learning Gains	3	Self-assessed learning gains

Table 6: Percent (count) of publication-instrument instances by topic including standardized exams ($N_{publication+instrument}$ = 460)

Topic	Percent publication-instrument instances (N = 460)
Cognitive	56.7% (n = 261)
Affective	32.0% (n = 147)
Behavioral	9.1% (n = 42)
Metacognition	3.0% (n = 14)
Evaluation	2.8% (n = 13)
Nature of Science	0.4% (n = 2)
Other	4.5% (n = 21)

Table 7: Percent (count) of unique instruments by topic including standardized exams ($N_{instruments}$ = 377)

Topic	Percent unique instruments (N = 377)
Cognitive	57.3% (n = 216)
Affective	30.5% (n = 115)
Behavioral	9.5% (n = 36)
Metacognition	3.4% (n = 13)
Evaluation	2,9% (n = 11)
Nature of Science	0.5% (n = 2)
Other	4.8% (n = 18)

Appendix 2. Topics of instruments in *CERP* publications 2010 – 2021 including standardized exams

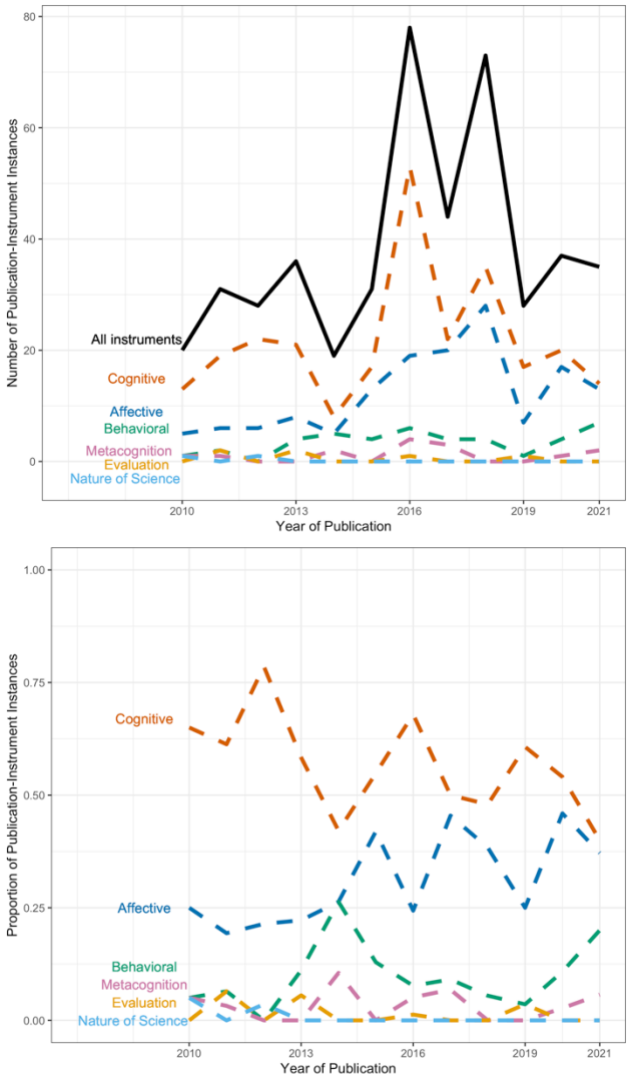

Figure 4: Researcher-developed instruments in publication-instrument instances (N = 460) including standardized exams by topic by year as count (top) and proportion (bottom).

Appendix 3. Psychometric evidence reported for instruments in *CERP* publications 2010 – 2021 including standardized exams

Table 8: Percent (count) of publication-instrument instances reported types of validity and reliability evidence including standardized exams; difficulty (n = 39) and discrimination (n =29) were not reported in any publication using a standardized exam

Type of Psychometric Evidence	**Percent of publication-instrument Instances (N = 460)**
Evaluation (Any)	60.8% (n = 280)
Validity (n = 220)	
Internal Structure (Factor Analysis/Principal Components Analysis/Rasch Analysis)	23.3% (n = 107)
Test Content	22.8% (n = 105)
Relations with Other Variables	21.9% (n = 101)
Response Processes	11% (n = 50)
Measurement Invariance/DIF	2.8% (n = 13)
Reliability (n = 211)	
Single Administration Reliability	45.2% (n = 208)
Retest Reliability	1.3% (n = 6)

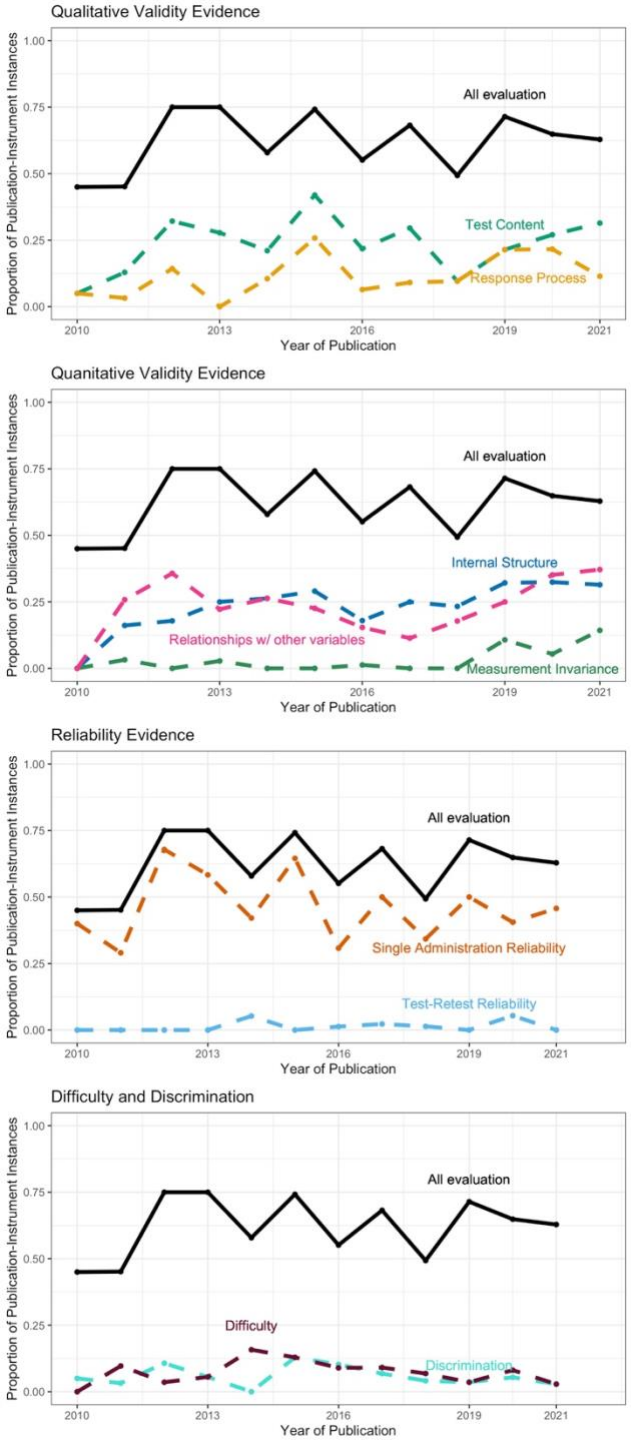

Figure 5: Proportion of publication-instrument instances (N = 460) that reported types of validity evidence, either qualitative (top) or quantitative (second), reliability evidence (third), or difficulty and discrimination (bottom) by year including standardized exams.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, (2014), *Standards for Educational and Psychological Testing*, Sage Publications, Inc.

Ardura D. and Pérez-Bitrián A., (2018), The effect of motivation on the choice of chemistry in secondary schools: adaptation and validation of the Science Motivation Questionnaire II to Spanish students. *Chemistry Education Research and Practice*, **19**(3), 905–918.

Arjoon J. A., Xu X., and 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.

Barbera J., Harshman J., and Komperda R. (eds.), (2022), *The Chemistry Instrument Review and Assessment Library* [Online]. Available at https://chiral.chemedx.org/ (Accessed 27 Sept 2022).

Barbera J., Naibert N., Komperda R., and Pentecost T. C., (2020), Clarity on Cronbach's Alpha Use. *Journal of Chemical Education*, **98**(2), 257–258.

Barbera J. and VandenPlas J. R., (2011), All assessment materials are not created equal: the myths about instrument development, validity, and reliability, in *Investigating classroom myths through research on teaching and learning*,., ACS Publications, pp. 177–193.

Bauer C. F., (2008), Attitude toward chemistry: A semantic differential instrument for assessing curriculum impacts. *Journal of chemical education*, **85**(10), 1440.

Blalock C. L., Lichtenstein M. J., Owen S., Pruski L., Marshall C., and Toepperwein M., (2008), In pursuit of validity: A comprehensive review of science attitude instruments 1935–2005. *International Journal of Science Education*, **30**(7), 961–977.

Bretz Research Group (2022). *Assessment tools* [Online]. Available at https://sites.google.com/miamioh.edu/bretzsl/research/assessment-tools (Accessed 27 Sept 2022).

Cooper M. M., Underwood S. M., and Hilley C. Z., (2012), Development and validation of the implicit information from Lewis structures instrument (IILSI): do students connect structures with properties? *Chemistry Education Research and Practice*, **13**(3), 195–200.

CRediT (2022). *Contributor Roles Taxonomy* [Online]. Available at https://credit.niso.org/ (Accessed 27 Sept 2022).

Cronbach L. J., (1988), Five perspectives on validity argument. *Test validity*, 3–17.

Dalgety J., Coll R. K., and Jones A., (2003), Development of chemistry attitudes and experiences questionnaire (CAEQ). *Journal of research in science teaching*, **40**(7), 649–668.

Deng J. M., Streja N., and Flynn A. B., (2021), Response process validity evidence in chemistry education research. *Journal of Chemical Education*, **98**(12), 3656–3666.

Ferrell B. and Barbera J., (2015), Analysis of students' self-efficacy, interest, and effort beliefs in general chemistry. *Chemistry Education Research and Practice*, **16**(2), 318–337.

Galloway K. R. and Bretz S. L., (2015), Development of an Assessment Tool To Measure Students' Meaningful Learning in the Undergraduate Chemistry Laboratory. *Journal of Chemical Education*, **92**(7), 1149–1158.

Geban Ö., Askar P., and Özkan Ï., (1992), Effects of computer simulations and problem-solving approaches on high school students. *The Journal of Educational Research*, **86**(1), 5–10.

Kirilenko A. P. and Stepchenkova S., (2016), Inter-coder agreement in one-to-many classification: Fuzzy kappa. *PloS one*, **11**(3), e0149787.

Komperda R., Pentecost T. C., and Barbera J., (2018), Moving beyond alpha: A primer on alternative sources of single-administration reliability evidence for quantitative chemistry education research. *Journal of Chemical Education*, **95**(9), 1477–1491.

Lewis S. E., (2022), Considerations on validity for studies using quantitative data in chemistry education research and practice. *Chemistry Education Research and Practice*.

Messick S., (1995), Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, **50**, 741–749.

PhysPort (2022). *Assessment* [Online]. Available at https://www.physport.org/Assessment.cfm (Accessed 27 Sept 2022).

Piburn M., Sawada D., Turley J., Falconer K., Benford R., Bloom I., and Judson E., (2000), Reformed teaching observation protocol (RTOP) reference manual. *Tempe, Arizona: Arizona Collaborative for Excellence in the Preparation of Teachers*.

R Core Team, (2022), R: A language and environment for statistical computing, [Computer software].

Rocabado G. A., Komperda R., Lewis J. E., and Barbera J., (2020), Addressing diversity and inclusion through group comparisons: a primer on measurement invariance testing. *Chemistry Education Research and Practice*, **21**(3), 969–988.

Seery M. K., Kahveci A., Lawrie G. A., and Lewis S. E., (2019), Evaluating articles submitted for publication in Chemistry Education Research and Practice. *Chemistry Education Research and Practice*, **20**(2), 335–339.

Seymour E., Wiese D., Hunter A., and Daffinrud S. M., (2000), Creating a better mousetrap: On-line student assessment of their learning gains, Pergamon Amsterdam, pp. 1–40.

Stains M., (2022), Keeping Up-to-Date with Chemical Education Research Standards. *J. Chem. Educ.*, **99**(6), 2213–2216.

Stamovlasis D., (2010), Methodological and epistemological issues on linear regression applied to psychometric variables in problem solving: Rethinking variance. *Chemistry Education Research and Practice*, **11**(1), 59–68.

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.

Tobin K. and Capie W., (1984), The Test of Logical Thinking. *Journal of Science and Mathematics Education in Southeast Asia*, **7**(1), 5–9.

Towns M. H., (2013), New guidelines for chemistry education research manuscripts and future directions of the field. *Journal of Chemical Education*, **90**(9), 1107–1108.

Treagust D. F., Chittleborough G., and Mamiala T. L., (2002), Students' understanding of the role of scientific models in learning science. *International Journal of Science Education*, **24**(4), 357–368.

Wu M., Tam H. P., and Jen T.-H., (2016), *Educational Measurement for Applied Researchers: Theory into Practice*, Springer.

Xu X. Lewis J. E., (2011), Refinement of a chemistry attitude measure for college students. *Journal of Chemical Education*, **88**(5), 561–568.

