What am I actually measuring? The role of validity evidence in STEM education research

Overview

• What is validity and why does it matter?

• How can DBER instruments examine validity?
  • Examination of the Science Motivation Questionnaire

• A DBER community resource of instruments
  • CHEmistry Instrument Review and Assessment Library (CHIRAL)
Why does validity matter?

“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” (p. 536)

Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence

Janelle A. Arjoon, Xiaoying Xu, and Jennifer E. Lewis

Department of Chemistry, University of South Florida, Tampa, Florida 33620, United States

Keeping Up-to-Date with Chemical Education Research Standards

Claims can only be taken seriously and acted upon if evidence is provided about the trustworthiness of the data upon which these claims are based. Reviewers systematically requested evidence for the validity and reliability of the data collected, whether the data were collected with a new or already existing and published instrument” (p. 2214)

Chemistry Education Research and Practice

This editorial is meant to provide considerations on how validity has been presented and reviewed among papers submitted to Chemistry Education Research and Practice (CERP) that analyze quantitative data. Authors submitting to CERP are encouraged to make an explicit case for validity and this editorial describes the varying sources of evidence that can be used to organize the evidence presented for validity.

Contemporary Test Validity in Theory and Practice: A Primer for Discipline-Based Education Researchers

Second, potential test users should be critical in their evaluation of existing instruments, and should not merely assume a strong validity argument exists for an instrument’s score interpretations and uses with a particular population. Potential users should look to the instrumentation (or methods) sections of published articles for key information.... Altogether, such practices should advance the quality of measurement within the realm of discipline-based education research (p. 15)
Measurement Error – Physical Science

**SYSTEMATIC ERROR**
- **Accuracy:**
  - Normal boiling point of water is 99.974 °C
  - Can calibrate to known standard

**RANDOM ERROR**
- **Precision:**
  - Repeated measurements are not identical
  - Can quantify the consistency among values

**SYSTEMATIC ERROR**
- Observed value = True value + Error
- 99.517 °C = 99.974 °C + Error
- Error = –0.457 °C

**RANDOM ERROR**
- Standard deviation = 0.247 °C
- Standard error of the mean = 0.143 °C
- 95% CI = [99.096 °C; 100.323 °C]

---

Measurement Error – Psychometrics

**SYSTEMATIC ERROR**
- **Validity:**
  - Often do not know the true value
  - Understanding
  - Motivation

**RANDOM ERROR**
- **Reliability:**
  - Consistency among repeated measurements
  - Test-retest reliability (correlation)
  - Consistency among values in one administration
  - Reliability coefficients (KR-20, alpha, omega)
"Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (p. 11)

“The term reliability has been used in two ways in the measurement literature...to refer to the reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test...in a more general sense, to refer to the consistency of scores” (p. 33)

“There is no single, preferred approach to quantification of reliability/precision...No one method of investigation is optimal in all situations” (p. 41)


Now Free! https://www.testingstandards.net/open-access-files.html
Evaluating Validity Evidence in STEM Education Research

- The amount and type of evidence is related to how the data will be used
  - The stakes of the results provide a guideline for the level of quality
- Evidence must be collected and evaluated each time an instrument is used
  - Even if the instrument has extensive use in the literature
  - Especially if you want to modify anything about the instrument

Call to Action from SMQ II Authors

“The items of the revised questionnaire, like the items of the original questionnaire, were designed so that the word science in each item can be replaced with the word biology, chemistry, or physics...
By this means it is possible to create discipline-specific questionnaire versions for these and other science disciplines, but researchers should examine each item to ensure it is representative of their target discipline and, when using any version of the questionnaire, establish its validity.” (p. 1173)
Science Motivation Questionnaire (SMQ)

- Development history well documented
  - SMQ (2006, 2007)
    - 30 items on 6 scales
    - 369 non-science majors in biology
  - SMQ (2009)
    - 30 items on 5 scales
    - 770 non-science majors in biology
  - SMQ II (2011)
    - 25 items on 5 scales
    - 313 non-science majors in biology
    - 367 science majors in biology

- Used frequently by other researchers
  - Languages: German, Greek, Japanese, Korean, Spanish, Thai, Turkish
  - Topics: Astronomy, Biology, Calculus, Chemistry, Computer Science, Histology, Math, Nanotechnology, Organic Chemistry, Pharmacy, Physics, Technology


In a chemistry course, how would you or a typical student answer?

“The science I learn is relevant to my life”

“The chemistry I learn is relevant to my life”

<table>
<thead>
<tr>
<th>Response</th>
<th>Never</th>
<th>Rarely</th>
<th>Sometimes</th>
<th>Usually</th>
<th>Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Chemistry (n = 287)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science Chemistry</td>
<td>6%</td>
<td>28%</td>
<td>33%</td>
<td>66%</td>
<td>49%</td>
</tr>
<tr>
<td>Introductory Chemistry (n = 373)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science Chemistry</td>
<td>20%</td>
<td>35%</td>
<td>39%</td>
<td>45%</td>
<td>17%</td>
</tr>
</tbody>
</table>
Validity – Confirmatory Factor Analysis

Latent variable

Intrinsic Motivation

Indicator variables (Items)


SMQ II Phase 1: Validity – Internal Structure

SMQ II Phase 1: Validity – Internal Structure

Correlation Matrix for 10 SMQ II Items by Latent Variable

- The science I learn is relevant to my life
- Learning science is interesting
- Learning science makes my life more meaningful
- I am curious about discoveries in science
- I enjoy learning science
- I put enough effort into learning science
- I use strategies to learn science well
- I spend a lot of time learning science
- I prepare well for science tests and labs
- I study hard to learn science


SMQ II Phase 1: Validity
Confirmatory Factor Analysis

Data-model fit for SMQ II by wording and course type

<table>
<thead>
<tr>
<th>Course</th>
<th>Wording</th>
<th>n</th>
<th>χ²</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Chemistry</td>
<td>Science</td>
<td>146</td>
<td>483</td>
<td>0.94</td>
<td>0.93</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Chemistry</td>
<td>141</td>
<td>468</td>
<td>0.96</td>
<td>0.96</td>
<td>0.07</td>
</tr>
<tr>
<td>Introductory Chemistry</td>
<td>Science</td>
<td>189</td>
<td>487</td>
<td>0.97</td>
<td>0.97</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Chemistry</td>
<td>184</td>
<td>657</td>
<td>0.94</td>
<td>0.94</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Target Fit Index Values ≥0.95

For all models WLSMV estimator used, df = 265, and p < 0.01

SMQ II Phase 1: Validity
Confirmatory Factor Analysis

Data-model fit for SMQ II by wording and course type

<table>
<thead>
<tr>
<th>Course</th>
<th>Wording</th>
<th>n</th>
<th>$\chi^2$</th>
<th>Data-Model Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Chemistry</td>
<td>Science</td>
<td>146</td>
<td>483</td>
<td>Unacceptable</td>
</tr>
<tr>
<td></td>
<td>Chemistry</td>
<td>141</td>
<td>468</td>
<td>Somewhat acceptable</td>
</tr>
<tr>
<td>Introductory Chemistry</td>
<td>Science</td>
<td>189</td>
<td>487</td>
<td>Somewhat acceptable</td>
</tr>
<tr>
<td></td>
<td>Chemistry</td>
<td>184</td>
<td>657</td>
<td>Unacceptable</td>
</tr>
</tbody>
</table>

For all models WLSMV estimator used, $df = 265$, and $p < 0.01$

Phase 1 Conclusions

- In general, there are issues with SMQ II functioning in all four contexts
- Current form of SMQ II is not appropriate for within-class or cross-class comparisons of science and chemistry motivation


---

SMQ II – Phase 2

<table>
<thead>
<tr>
<th>Survey of Experts</th>
<th>Interview Students</th>
<th>Revise Items</th>
<th>Administer to New Sample</th>
<th>Test Internal Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experts in DBER (n = 8) or educational measurement (n = 4)</td>
<td>40 general chemistry students</td>
<td>9 items revised 4 item added Response scale changed</td>
<td>9 colleges and universities Biology and Chemistry courses 2487 student responses Randomized wording</td>
<td>Exploratory Factor Analysis Confirmatory Factor Analysis</td>
</tr>
</tbody>
</table>


Learning science will help me get a good job
I think about the grade I will get in science
I am confident I will do well on science labs and projects

Expert Review of Response Scale

<table>
<thead>
<tr>
<th>Frequency Scale</th>
<th>Likert Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never</td>
<td>Disagree</td>
</tr>
<tr>
<td>Rarely</td>
<td>Somewhat disagree</td>
</tr>
<tr>
<td>Sometimes</td>
<td>Neither disagree nor agree</td>
</tr>
<tr>
<td>Usually</td>
<td>Somewhat agree</td>
</tr>
<tr>
<td>Always</td>
<td>Agree</td>
</tr>
</tbody>
</table>

Cognitive Interviews with Students

“All the time, that’s just another thing, I gotta get good grades to get into the program. To get a good job.”

“I did pretty well on my labs. We don’t have any kind of projects in gen chem”

SMQ II Phase 2: Validity
Confirmatory Factor Analysis

Data-model fit for single-factor mSMQ II models

<table>
<thead>
<tr>
<th>Scale</th>
<th>Course</th>
<th>Wording</th>
<th>n</th>
<th>$\chi^2$</th>
<th>Data-Model Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic Motivation (df = 5)</td>
<td>General Chemistry</td>
<td>Science</td>
<td>835</td>
<td>69</td>
<td>Acceptable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chemistry</td>
<td>855</td>
<td>166</td>
<td>Somewhat acceptable</td>
</tr>
<tr>
<td></td>
<td>Preparatory Chemistry</td>
<td>Science</td>
<td>139</td>
<td>18</td>
<td>Somewhat acceptable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chemistry</td>
<td>137</td>
<td>42</td>
<td>Somewhat acceptable</td>
</tr>
<tr>
<td></td>
<td>General Biology</td>
<td>Science</td>
<td>258</td>
<td>37</td>
<td>Somewhat acceptable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Biology</td>
<td>263</td>
<td>48</td>
<td>Somewhat acceptable</td>
</tr>
<tr>
<td>Grade Motivation (df = 2)</td>
<td>General Chemistry</td>
<td>Science</td>
<td>835</td>
<td>4</td>
<td>Acceptable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chemistry</td>
<td>855</td>
<td>7</td>
<td>Somewhat acceptable</td>
</tr>
<tr>
<td></td>
<td>Preparatory Chemistry</td>
<td>Science</td>
<td>139</td>
<td>7</td>
<td>Somewhat acceptable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chemistry</td>
<td>137</td>
<td>3</td>
<td>Somewhat acceptable</td>
</tr>
<tr>
<td></td>
<td>General Biology</td>
<td>Science</td>
<td>258</td>
<td>2</td>
<td>Acceptable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Biology</td>
<td>263</td>
<td>2</td>
<td>Acceptable</td>
</tr>
</tbody>
</table>

WLSMV estimator used; cut-off values CFI and TLI ≥ 0.95 and RMSEA ≤ 0.05

SMQ II Phase 2: Validity
Confirmatory Factor Analysis

Data-model fit for five-factor mSMQ II model

<table>
<thead>
<tr>
<th>Course</th>
<th>Wording</th>
<th>n</th>
<th>$\chi^2$</th>
<th>Data-Model Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Chemistry</td>
<td>Science</td>
<td>417</td>
<td>251</td>
<td>Acceptable</td>
</tr>
<tr>
<td></td>
<td>Chemistry</td>
<td>426</td>
<td>334</td>
<td>Somewhat acceptable</td>
</tr>
<tr>
<td>Preparatory Chemistry</td>
<td>Science</td>
<td>139</td>
<td>189</td>
<td>Acceptable</td>
</tr>
<tr>
<td></td>
<td>Chemistry</td>
<td>137</td>
<td>194</td>
<td>Acceptable</td>
</tr>
<tr>
<td>General Biology</td>
<td>Science</td>
<td>128</td>
<td>179</td>
<td>Acceptable</td>
</tr>
<tr>
<td></td>
<td>Biology</td>
<td>130</td>
<td>191</td>
<td>Acceptable</td>
</tr>
</tbody>
</table>

WLSMV estimator used, $df = 142$, and $p < 0.01$; cut-off values CFI and TLI ≥ 0.95 and RMSEA ≤ 0.05

Conclusions about SMQ

- **Significant** revision required
  - Total motivation scores are **inappropriate**
  - Individual motivation scales did not show consistent functioning
  - Individual scale constructs have **weak** theoretical support
    - Grade and career motivation were data-driven
    - Strong correlations between intrinsic and career motivation are problematic
      - Also observed by other researchers or collapsed into single factor during factor analysis


Implications for Measuring Motivation

- Only looking at original instrument papers may not provide the full picture
- Extensive use does not imply an instrument functions in all settings
- Addressing multiple sources of validity evidence is beneficial
- Other motivation scales have more grounding in theoretical frameworks

<table>
<thead>
<tr>
<th>Reference</th>
<th>Sample</th>
<th>Translations &amp; Modifications</th>
<th>Factors (correlated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glynn et al., 2011</td>
<td>340 university biology students</td>
<td>None</td>
<td>5</td>
</tr>
<tr>
<td>Ardura &amp; Pérez-Bitrián, 2018</td>
<td>530 high school students</td>
<td>Spanish; Physics and chemistry wording</td>
<td>5</td>
</tr>
<tr>
<td>González et al., 2017</td>
<td>520 high school students</td>
<td>Spanish; Physics wording; Only self-efficacy</td>
<td>1</td>
</tr>
<tr>
<td>Kwon, 2016</td>
<td>334 middle school students</td>
<td>Technology wording; Changed response scale</td>
<td>5</td>
</tr>
<tr>
<td>Salta &amp; Koulougliotis, 2015</td>
<td>330 high school students</td>
<td>Greek; Chemistry wording; Removed lab references</td>
<td>5</td>
</tr>
<tr>
<td>Tosun, 2013</td>
<td>306 high school students</td>
<td>Turkish; Chemistry wording; Removed six items</td>
<td>5</td>
</tr>
<tr>
<td>Tosun, 2013</td>
<td>266 university students</td>
<td>Turkish; Chemistry wording; Removed five items</td>
<td>5</td>
</tr>
<tr>
<td>Vasques et al., 2018</td>
<td>203 university students (Pre)</td>
<td>Japanese; Removed a self-efficacy item</td>
<td>5</td>
</tr>
<tr>
<td>Vasques et al., 2018</td>
<td>230 university students (Post)</td>
<td>Japanese; Removed a self-efficacy item</td>
<td>5</td>
</tr>
<tr>
<td>Yamamura &amp; Takehira, 2017</td>
<td>165 pharmacy students</td>
<td>Japanese; Pharmacy wording; Removed 12 items including all self-efficacy items</td>
<td>4 with cross loadings</td>
</tr>
</tbody>
</table>

*Fit index values recommended by Hu and Bentler (1999) are CFI > 0.95; RMSEA ≤ 0.06 and SRMR ≤ 0.08*
Lessons Learned from the SMQ II

• Validity and reliability information needs to be easier to find and keep track of
  • What is the best way to organize and share psychometric data for instruments?
  • How can we keep track of this information for every use of an instrument?
  • How can we synthesize this information for every instrument?

A Psychometric Resource for DBER

• CHeistry Instrument Review and Assessment Library (CHIRAL)
  [chiral.chemedx.org]

• Centralized resource for instrument information and psychometric summary

**Undergraduate Researchers:**
- Dylan Alcazar
- Tyree Baker
- Victoria Delgado
- Briana Enriquez
- Jan Lipovaca
- Allison Phillips
- Bryanna Mendenhall
- Eleanor Quirk
- Erica Sisouphanthong
- Grace Trinh
- Duong Vu

[Web resource]

**SDSU | San Diego State University**

**Portland State University**

**Expert Panels**

[D. Kat Lazenby]

[Kristin Tenney]

[Tina Marcroft]

[Dr. Jordan Harshman]

[Dr. Jack Barbera]
CHIRAL Features

- Integrated into Chemical Education Xchange chiral.chemedx.org

- Features
  - Searchable by topic or population
  - List uses and alternative versions or translations
  - Identification of publications providing psychometric data
  - Peer-reviewed panel summary synthesizing psychometric data
  - Glossary of psychometric and instrument terms

- Current status
  - Identified over 800 unique publications containing ~850 instruments
  - Over 1300 instances of instrument use
  - CHIRAL website displays information on 526 instruments
  - Hosted peer review panels for 40 instruments from 2021 – 2023
### Chemistry And Biology Assessment For Biochemistry

**Published:** 2011, 2011  
**Population:** Students, Undergraduate  
**Format:** Multiple Choice  
**Questions:** 24

<table>
<thead>
<tr>
<th>Overview</th>
<th>Evidence</th>
<th>Review</th>
<th>Versions</th>
<th>Citations</th>
</tr>
</thead>
</table>

**Summary**

- **Original author(s):** Villafane, S.M., Loercher, J., Minderhaus, V., & Lewis, J.E.
- **Year original instrument was published:** 2011, 2011

**Inventory**

- **Number of items:** 24
- **Number of versions/translations:** 1
- **Total implementations:** 7

**Language:** English  
**Country:** United States  
**Format:** Multiple Choice

**Intended population (S):** Students  
**Undergraduate:**

**Domain:** Cognitive  
**Topic:** Biochemistry

---

### Chemistry And Biology Assessment For Biochemistry

**Published:** 2011, 2011  
**Population:** Students, Undergraduate  
**Format:** Multiple Choice  
**Questions:** 24

<table>
<thead>
<tr>
<th>Publications</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original development paper</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uses in the instrument in data collection</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Modified version of existing instrument</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Evaluation of existing instrument</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Reliability**

- Test-retest reliability
- Internal consistency
- Coefficient (Cronbach's alpha)
- McDonald's Omega
- Inter-rater reliability
- Person separation
- Generalizability coefficients
- Other reliability evidence

**Validity**

- Expert judgment
- Response process
- Factor analysis, IRT, Rasch analysis
- Differential item function
- Evidence based on relationships to other variables
To What Extent Does CER Report Psychometrics?

- **Pre-registration, data and analysis code:** [https://osf.io/cq43f](https://osf.io/cq43f)
- **Why Chemistry Education Research and Practice (CERP) was selected:**
  - Freely available
  - Represents an international audience
  - All articles are considered chemistry education research
- **CERP has been publishing since 2000**
  - We started with all publications since 2010 (N = 767)
  - Extracted information about the instruments and validity/reliability evidence

**Panel Review: Chemistry & Biology Assessment for Biochemistry**

(Post last updated June 14, 2022)

**Review panel summary**

The Chemistry & Biology Assessment for Biochemistry (C&B&A) is intended to assess foundational understanding of chemistry and biology topics that are viewed as prerequisite to biochemistry learning. The instrument consists of 24 items, divided into groups of three questions associated with each of eight topics: bond energy, free energy, London dispersion forces, pH/pKa, hydrogen bonding, alpha helix, amino acids, and protein function. It has been tested with undergraduate biochemistry students at a variety of US institutions, some of which are classified as Hispanic-serving institutions by the US Department of Education.

The developers of the C&B&A described the process of instrument development in great detail [1]. Of note is the attention given to **content validity**. The items were written by a team of PhD faculty members in chemistry, biology, and biochemistry. The multiple choice items are highly structured so that across all three items designed for each topic, the four multiple choice options contain a single consistent correct idea and three common misconceptions. After an initial writing of many items, the final items selected were again revised for consistency and clarity by the authors. The data produced by the 24-item instrument were assessed for internal structure validity using confirmatory factor analysis. The fit statistics used to determine whether the eight-factor solution matched the data were satisfactory. However, the panel was unable to fully assess the internal structure validity evidence because the complete model was not described.

While cognitive interviews were described in a subsequent paper [2, supplementary information], it is unclear how the information from these may have been used to improve the items or the degree to which they provided supportive evidence of **response process validity**.

Two publications [1, 2] reported **single administration reliability** using **coefficient alpha** values for each of the eight subscales. These values ranged widely and it is unclear whether they provide sufficient evidence for the **reliability** of the data.

While **difficulty** and **discrimination** values were not reported by item, the low overall scores in all the studies examined [1-5] suggest that the items are quite difficult and therefore may be unlikely to discriminate well between high and low achieving students.

**Recommendations for use**

The C&B&A has been used and tested with undergraduate biochemistry students in the US and is only available by contacting the authors [3] directly.

Traditionally, the C&B&A is scored separately for each topic, with a student either getting all three items correct (scored as correct), or getting two or fewer items correct (scored as incorrect). When scored in this way, student scores tend to be quite low. **Misper** scores in study [1] on each topic were between 0.05 and 0.33, meaning that about 5-33% of students got all three questions right for that topic. When the instrument was scored as a single scale, as in [5], the average score on the pre-test was 7/21 (33%), and 10/21 (48%) on the post-test. Thus, we recommend that instructors consider use of the C&B&A in ways that do not
How are measurement instruments used and/or evaluated in CERP?

There are no clear trends in targets of measurements over time.

**RESULTS**

**Cognitive constructs are the most common target**

<table>
<thead>
<tr>
<th>Topic</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive</td>
<td>54.9%</td>
</tr>
<tr>
<td>Affective</td>
<td>34.2%</td>
</tr>
<tr>
<td>Behavioral</td>
<td>9.8%</td>
</tr>
<tr>
<td>Metacognition</td>
<td>3.3%</td>
</tr>
<tr>
<td>Evaluation</td>
<td>3.0%</td>
</tr>
</tbody>
</table>

**Yet, instruments measuring affective constructs are most commonly administered in multiple publications**

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attitude Toward the Subject of Chemistry Inventory version 2 (ASCI V2)</td>
<td>8</td>
</tr>
<tr>
<td>Attitude Toward the Subject of Chemistry Inventory (ASCI V1)</td>
<td>4</td>
</tr>
<tr>
<td>College Chemistry Self-Efficacy Scale – Cognitive Skills Scale (CCSS – CSS)</td>
<td>4</td>
</tr>
<tr>
<td>Lawson's Classroom Test of Formal Reasoning (Greek)</td>
<td>4</td>
</tr>
</tbody>
</table>

*Out of 430 appearances of instruments in publications; instruments used fewer than 10 times are omitted*
How are measurement instruments used and/or evaluated in CERP?

- Use (n = 400; 93%) instrument is used to generate data to address research question not related to validity & reliability
- Original (n = 304; 71%) presentation of a newly developed (or modified) instrument
- Modification (n = 126; 29%) modification (e.g., subset of items, translation) of an existing instrument
- Evaluation (n = 277; 64%) investigation of instrument data quality (validity & reliability)

N = 430 instrument appearances in publications

RESULTS
How are measurement instruments used and/or evaluated in CERP?

- Use (n = 400; 93%) instrument is used to generate data to address research question not related to validity & reliability
- Original (n = 304; 71%) presentation of a newly developed (or modified) instrument
- Modification (n = 126; 29%) modification (e.g., subset of items, translation) of an existing instrument
- Evaluation (n = 277; 64%) investigation of instrument data quality (validity & reliability)

To what extent do CER researchers provide psychometric evidence for instrument data, as reported in CERP?


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

![Diagram showing psychometric evidence sources and validity](image1)

**RESULTS**


---

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

![Diagram showing psychometric evidence sources and validity](image2)

**RESULTS**


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

**RESULTS**

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

---


N = 430 instrument appearances in publications
Summary

- Validity must be collected and evaluated each time an instrument is used
  - DBER fields are aware, but is still not always reported
- Synthesizing the psychometric data for an instrument is complex, and we hope CHIRAL helps!

Lewis, S. E. Considerations on Validity for Studies Using Quantitative Data in Chemistry Education Research and Practice. Chemistry Education Research and Practice 2022.
Stains, M. Keeping Up-to-Date with Chemical Education Research Standards. Journal of Chemical Education 2022, 99 (6), 2213–2216.

Acknowledgments

• Collaborations & Funding

Dr. Jungwon Kim
Sharai Mendez
Tina Marcroft
Abel Sekone
Haley Palm

Undergraduate Researchers:
Briana Enriquez
Bryanna Mendenhall
Erica Sisouphanthong (Team Lead)
Duong Vu