ACS CHEMICAL HEALTH & SAFETY

Editors' Choice
Open Access

pubs.acs.org/acschas

Review

Validity and Reliability of Survey Data: Key to Empowering Chemical Health and Safety Research

Qi Cui,* Jordan T. Harshman, and Regis Komperda

Cite This: *ACS Chem. Health Saf.* 2024, 31, 121−126

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: This work provides a guide for researchers and practitioners to develop and administer surveys within the context of chemical health and safety research. It discusses the challenges and key factors in developing health and safety surveys, focusing on evidencing validity and reliability in the field of psychometrics. The discussion encompasses survey design, question construction, ethical data collection, and the use of pilot studies for testing. The paper highlights the importance of adhering to the *Standards* for survey evaluation, advocating for the validity and reliability of survey data akin to accuracy and precision in benchtop applications. This work seeks to enhance the robustness of survey data, thereby reinforcing the foundation upon which chemical health and safety research can advance.

KEYWORDS: *validity, reliability, survey*

SURVEY METHODOLOGY IN CHEMICAL HEALTH AND SAFETY RESEARCH

In chemical health and safety research, the use of survey methodology enables researchers to measure and quantify the intricate interplay among workplace practices, safety, and health-related outcomes. The role of surveys in assessing attitudes, practices, and knowledge is critical for the development and refinement of safety protocols, chemical hygiene practices, and hazard assessment tools. Enhancing our understanding of chemical health and safety, informing regulatory updates, and shaping effective risk management strategies are possible only by analyzing survey data that demonstrate evidence for validity and reliability. Just as precision and accuracy are fundamental in benchtop applications, evidencing the validity and reliability of survey data is equally critical and should be approached with comparable rigor. Therefore, it is paramount to emphasize the role of robust survey methods in contributing to the efficacy and progression of health and safety standards.

Survey instruments are systematic tools designed to gather information from respondents in a structured manner. These tools vary in complexity, from simple questionnaires to complex scales, and are crafted to align with specific research objectives. Simple questionnaires typically consist of straightforward, closed-ended questions, whereas complex scales involve multidimensional questions that measure various aspects of a subject.[1] Two primary types of surveys are prevalent in chemical health and safety research: cross-sectional and longitudinal.[2] Cross-sectional surveys provide a snapshot of a population at a single point, capturing a wide range of data points concurrently, while longitudinal surveys track and analyze changes over time, offering valuable insights into trends and potential cause-and-effect relationships.

The development of health and safety surveys comes with its challenges, particularly concerning the evidence that demonstrates the validity and reliability of survey data, which is the field of psychometrics.[3] Validity refers to the extent to which the survey measurements accurately capture what they purport to measure, while reliability pertains to the consistency of these measurements over time. The multifaceted nature of chemical health and safety, with its complex hazards and diverse settings, demands that data generated by surveys be both valid and reliable before inferences are made. If understanding and policy are based on information generated from survey data that have not demonstrated these psychometric facets, then the validity of assessment and management of chemical health and safety risks is severely jeopardized.

The purpose of this article is to underscore the critical importance of validity and reliability in survey instrument data within the context of chemical health and safety research. By dissecting the methodological underpinnings that contribute to the robustness of survey data, this work aims to empower researchers and practitioners with the knowledge to develop

Received: December 4, 2023
Revised: January 23, 2024
Accepted: January 29, 2024
Published: February 13, 2024

https://doi.org/10.1021/acs.chas.3c00111
ACS Chem. Health Saf. 2024, 31, 121−126

and deploy surveys that yield valid and reliable insights, ultimately enhancing the effectiveness of health and safety interventions in the chemical industry.

SURVEY DESIGN AND IMPLEMENTATION IN HEALTH AND SAFETY

Survey Purpose. The foundational step in constructing a survey begins with delineating its purpose clearly. Examples may include assessing risk perceptions, evaluating the efficacy of safety measures, or determining compliance with health regulations. Consider the three example statements of purpose in Table 1 below for examples of how to clearly delineate the purpose of a survey.

Survey Design. The development of questions is a critical aspect where both language and structure play pivotal roles in influencing the quality of data collected.[5] This process must adhere to best practices that include the use of simple, familiar words, avoiding ambiguous meanings, and ensuring questions are specific, concrete, and structured to avoid bias.[1] Some examples of unintentionally ambiguous questions are those that ask two questions in a single statement, referred to as "double-barreled" in the case of an item such as "My organization has a safe and well-organized chemical storage environment"; if the survey respondent agrees with half the statement but not the other, they may not be able to provide a meaningful response to the question. To minimize bias in responses, it can be helpful to start with questions that are simple and straightforward to answer before proceeding to more difficult or complex questions or those that may ask the respondent to reveal sensitive information, such as accident rates or safety violations (Table 2). Additionally, researchers should note that it is important to be detail-oriented in relation to the wording of an item's stem and its response options, if present. As an example of stem wording, consider how "I am aware of risks" may invoke entirely different thoughts in participants than "I am aware of potential risks." Also, asking participants to respond on a Likert scale (agree−disagree) measures something very different than a frequency scale (never−always).

When deciding on the structure of a survey or the format of items, consider both the burden on the respondent and those analyzing the data. Best practices include grouping similar questions together or including an introductory statement that describes the information the respondent should gather before beginning, such as training records, incident reports, or other documents. As another example, collection of numeric information is typically easier if respondents can enter numeric values that can be analyzed directly, rather than prematurely collapsing numeric data into categories such as 0−1, 2−5, >5, etc. These categories can always be created from the raw data, but if recorded as categorical responses, the raw numbers can never be known, which may limit the statistical analyses that can be conducted.

Ethical Considerations and Institutional Review. An overlooked but essential aspect of survey design involves ethical considerations and the possible need for an Institutional Review Board (IRB) review. In academic settings, researchers can consult their university's Office of Research Integrity or a similar entity, where an IRB administrator or committee is often housed. Notably, the specific name and presence of such offices may vary, especially in smaller or specialized institutions. In industrial or federal research settings, similar review bodies are typically established to oversee ethical

Table 1. Examples of Survey Purpose Statements with Varying Levels of Detail and Clarity

Categorization	Purpose	Comments
Vague	This survey measures safety assessment among industrial chemists.	The term "safety assessment" is incredibly broad, making the validation efforts incredibly easy as any questions pertaining to safety are likely related to the broad construct of "safety assessment".
Somewhat Delineated	This survey measures industrial chemists' ability to recognize risks and assess chemical hazards.	While this purpose somewhat delineates specific facets of safety assessment, it is still unclear how those are conceptualized.
Detailed	This survey measures the ability of synthetic industrial chemists within research teams at large industries to (1) recognize risks throughout an entire synthetic pathway (across a team of workers) and (2) assess the specific chemical hazards and potential dangers involved in a specific synthetic path in accordance with the RAMP[4,a] framework.	This purpose adequately describes what is meant by the "R and A" of RAMP within the specific context and defines the intended population, which may be operationalized differently for different populations.

[a]Note: the term "RAMP" refers to the four principles of safety: Recognize hazards, Assess risks of hazards, Minimize risks of hazards, and Prepare for emergencies.

https://doi.org/10.1021/acs.chas.3c00111
ACS Chem. Health Saf. 2024, 31, 121−126

Table 2. An Example of Survey Design

Question Type	Question Content	Purpose
Simple, rating scale	On a scale from 1 to 5, how would you rate the overall safety of your work environment?	To begin engaging with the topic of safety in a nonspecific manner.
Sensitive, closed-ended	Have you ever witnessed a safety violation in your workplace? (Yes/No)	To begin broaching more sensitive topics, yet still in a binary, less intrusive manner.
Sensitive, open-ended	If yes, please describe the nature of the safety violation and how it was addressed.	To gather specific information on sensitive issues like safety violations, only from those who affirmatively respond to the previous question.

compliance. Researchers unfamiliar with these procedures are encouraged to refer to resources like the CITI Program's Human Subjects Research sections.

Obtaining IRB approval is a crucial and legal requirement for ensuring that the research adheres to ethical standards and protects the rights of participants, particularly for surveys intended for public dissemination or academic research. This process involves a detailed evaluation of the survey's content, data collection methods, and confidentiality measures. It is important to note that not all surveys require IRB review. For example, internal surveys conducted within a company for organizational purposes, such as surveying employees, may not fall under the IRB purview. However, research intended for publication or external distribution typically necessitates IRB oversight. Failure to secure IRB approval in these cases can result in significant setbacks, such as the inability to use the collected data for publication or academic purposes.[6] Hence, it is imperative for researchers to submit their survey protocol to an IRB for review and approval prior to administering the survey, ensuring that the research is ethically sound and valid for scientific dissemination.

Pilot Testing. Testing the survey through a pilot study is pertinent to identifying any issues or biases in the design and analysis plan. This process includes evaluating questions to minimize biases like social desirability[7] and ensuring an optimal flow of questions from general to specific, especially for sensitive topics. As illustrated in the study by Ménard et al., which focused on accident experiences and reporting practices in chemistry and biochemistry laboratories, conducting a pilot study is crucial for refining the survey to align with the study's objectives before its final administration.[8] It is also helpful for researchers to see what data are generated to identify any pitfalls in the analysis plan, as researchers have access to the data and can trial-run analyses.

Validity and Reliability of Survey Data. Evidencing the validity of survey responses ensures that the survey accurately reflects the constructs it is intended to measure.[3,9−11] It involves a meticulous process where every aspect of the survey, from question design to response options, aligns with the intended measures. Reliability, on the other hand, ensures that the survey produces consistent results for repeated applications under similar conditions. This stability is crucial for tracking changes and trends over time.[3,12] Having consistent scores over multiple occasions ensures that the scores can be reliably obtained and are not due to chance.

Both validity and reliability are grounded in the *Standards for Educational and Psychological Testing*[3] developed by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The *Standards* provide criteria for evaluating tests, testing practices, and the effects of test use, with "test" encompassing a wide range from affective scales and surveys to traditional knowledge assessments. It supports the development of instruments that are not only insightful but also consistent and replicable in their application.

Sources of Validity Evidence. Validity in survey research is not a matter of simply writing survey items aligned with the topic of interest; it encompasses a range of evidence sources. Each of these sources sheds light on distinct facets of validity while collectively upholding its unified concept.[3,11,13,14] It is crucial to evaluate the validity evidence every time a survey instrument is used, meaning that the common phrase "this survey has previously been validated" opens researchers to serious threats to validity. As highlighted by Nunnally, challenges in ensuring measurement quality can stem not just from the instrument but also from sampling errors related to the diversity of respondents.[15] This means that an identical survey instrument can yield varying degrees of validity when applied to different demographic or population groups, underscoring the need for careful validation in diverse settings.

It is important to note that the following types of evidence for validity are things that all researchers should consider, but not every survey necessitates *evidencing all types*. It is ultimately up to the researchers to determine which types of evidence will provide the most compelling case that their survey measures what they designed it to measure. For example, not all surveys are designed with a clear structure of constructs (see "Evidence Based on Internal Structure" for more details), so this type of evidence would not be appropriate for such a survey.

Evidence Based on Test Content. The concept of "test content" encompasses the themes, wording, and format of items, tasks, or questions in a survey as well as the guidelines for procedures regarding administration and scoring. This definition aligns with the concept of content validity as recognized in various studies.[16,17] Evidence based on test content focuses on the relevance and representativeness of the survey content in relation to the intended construct.[12] This is achieved through a rigorous examination of the survey items to confirm that they are representative of the domain of interest.

The documentation of data from existing literature, external experts, the author team, and other relevant resources can provide evidence of how the collection of items, tasks, or questions aligns with the detailed purposes of the survey. For example, to develop a semiquantitative survey designed to measure chemicals experts' ability to control the exposure of chemicals hazardous to health at the workplace, the validation of the tool's content was enriched via feedback from four expert panels.[18] This process ensured the relevance, representativeness, and comprehensiveness of the survey content, aligning it closely with the intended constructs of chemical safety and health.

Evidence Based on Response Processes. Evidence based on response process refers to the fit between the constructed survey questions and the actual performance or responses of the participants.[3] This links the individuals providing the data (the participants) to the nature of the data (the responses). The underlying cognitive activities that participants engage in

https://doi.org/10.1021/acs.chas.3c00111
ACS Chem. Health Saf. 2024, 31, 121−126

while responding to questions form the basis of response processes.[19] Analyzing these processes is critical to confirm that the methods used by respondents are in alignment with the survey designer's intentions.[3]

One effective method for examining response processes is through cognitive interviews. These in-depth interviews are designed to elicit detailed feedback from participants about specific survey items,[20] shedding light on their thought processes when responding to survey items. They typically consist of asking participants to respond to the survey and think out loud as they do so in the presence (virtual or otherwise) of the researchers. This allows the survey developer to explore in detail the participants' thinking and understanding of the items when they respond in certain ways to the individual item. For example, in the study of the investigation into laboratory technicians' use and interpretation of hazard communication elements on chemical labels, the authors employed cognitive interviews to gather feedback from participants about the survey items and made small adjustments to avoid misinterpretations.[21]

Evidence Based on Internal Structure. Evidence based on internal structure is defined as the extent to which the relationships among survey items align with the underlying construct(s) that the survey scores are intended to represent and interpret.[3,22] This aspect of evidence critically examines the efficacy of the survey instrument in producing valid measures of the targeted construct(s), which are referred to as latent. A latent (meaning "hidden") construct is a complex mental structure that usually cannot be measured accurately by a single item. For example, there are many facets of understanding personal protective equipment (PPE)—what it is, why it is needed, context-specific awareness of PPE, *etc*. A construct may be "understanding of PPE" and if that is a goal of the survey measure, the researchers have a responsibility to provide evidence that responses to survey items demonstrate a coherent mental structure.

In order to generate robust measures of a latent construct, a specific structure of the underlying instrument derived data must be defined. Evidence related to internal structure involves examining how individual items within the instrument are interrelated.[3] Latent constructs can be unidimensional (all items interrelate with only one latent construct), multidimensional (each item relates to one or more latent constructs), or structured in complex hierarchies. Regardless, it is imperative that the relationships among the items consistently reflect the underlying structure.

The scoring methods play a crucial role in the interpretation of these constructs. The process of scoring, whether by summing item scores to create one score or using different scales, must reflect the survey's dimensionality. A survey intended to yield a single composite score, for example, should be unidimensional. In the case of multidimensional structures, it is important to report multiple scores that align with the various scales. Additionally, more complicated scoring systems might not assign equal value to each item, and such variations should be justified.[23] Factor analysis, both exploratory (EFA) and confirmatory (CFA), is vital in this context.[24] EFA helps in identifying potential latent constructs and ensuring their alignment with the survey's intended outcomes, while CFA tests the fit of these constructs within the predefined model. These analytical techniques guarantee that the scoring system accurately represents the targeted constructs.

In a study conducted by Liu et al.,[25] the authors employed exploratory factor analysis and confirmatory factor analysis to elucidate the relationship between observed and latent variables within questionnaire categories. This approach demonstrates how factor analysis can be utilized to affirm that the items within a survey are coherently structured and effectively measure the underlying construct, in this case, the impact of chemistry learning motivation on freshmen's laboratory safety perception.

Evidence Based on Relations to Other Variables. Evidence based on relations to other variables involves confirming that the survey scores or responses relate to other measures in a predictable manner.[3] For example, one would expect that higher scores on a positive safety climate survey would correlate with lower accident rates.[26] If a research team is developing a survey measuring the safety climate, they may analyze correlations from their survey responses to accident frequency. Establishing such correlations provides evidence that the survey is measuring the intended construct accurately.

Evidence Based on Test Consequences. Finally, survey writers may need to establish evidence based on consequences of the intended use of responses.[3] This is less common, as it only applies in the specific circumstance where responses will be used to make decisions that carry significant consequences. For example, a survey intends to measure an organization's understanding and implementation of RAMP[4] concepts in its operations. If the survey designers intend to implement a cutoff score that indicates unacceptable safety and petition agencies to use the instrument to distribute fines and other penalties, then the researchers have an onus to provide evidence that organizations that score below the threshold are deserving of the penalties.

Sources of Reliability Evidence. The *Standards* put forward the term "measurement error" as the central key in establishing reliability evidence.[3] Errors in measuring human cognition are unavoidable and represent variances in how consistently participants (mis)interpret items and response options (internal consistency) or how the same participant varies in responses to the same question over time (temporal stability). These are measured by various statistics in derived from theoretical assumptions of what constitutes error and how to quantify it.[3]

Evidence Based on Temporal Stability. To ensure that a survey consistently measures respondents' performance on the construct of interest, researchers may choose to administer the same survey to the same sample of participants at a set time after initial administration. The correlation between the scores and/or responses from these two administrations is then computed.[3,27,28] Such a method for assessing reliability is predicated on the assumption of the score stability across these administrations. This assumption is warranted only when no significant changes or developments (such as new learning experiences or growth) have occurred between the survey administrations that could potentially alter the participants' relationship to the construct being measured.

Temporal stability may not be expected in some applications. A Likert item such as "I believe I am responsible for the safety of myself and others in the workplace" is a belief that is expected to remain constant over long periods of time. Therefore, researchers may want to demonstrate that participants respond consistently across two administrations of the same question, a few days apart. However, a Likert item such as "I wear PPE every day" would likely solicit different

https://doi.org/10.1021/acs.chas.3c00111
ACS Chem. Health Saf. 2024, 31, 121−126

responses depending on what the participant was doing most recently (i.e., running a synthesis versus completing general office work).

Evidence Based on Internal Consistency. Internal consistency is complicated by assumptions made by the researchers about the theoretical definition of measurement error, a discussion beyond the scope of this commentary. Generally speaking, participants who respond to an item in a particular way may be more inclined to respond to similar items in a predictable fashion. Thus, an observed consistency in responses to similar items indicates a strong internal consistency, underscoring the reliability of the survey data in measuring a particular concept.[3] In a case study from a bleach processing plant, the reliability was evaluated using Cronbach's α coefficient for each risk perception dimension.[29] Alternatively, we might expect participants of similar "ability" (e.g., understanding of SDS) to respond consistently to certain items, which may also demonstrate internal consistency. Demonstration of measurement error under this framework would like to involve a technique called item response theory, which is a statistical approach that models the relationship between survey takers' overall level of ability, trait, or proficiency being measured and their performance on individual test items.[30]

■ SOME FINAL IMPLEMENTATION SUGGESTIONS

Evidencing the validity and reliability of survey data is analogous to evidencing accuracy and precision in benchtop applications and should be considered as being just as rigorous. Many of the procedures to adequately evidence validity and reliability are straightforward and accessible to researchers without training in psychometrics, while others are deeply complicated and contextualized by thousands of articles, perspectives, and field-specific standards. A vital direction for future research is the collaboration with chemistry education researchers (CER map: https://chem.uncg.edu/popova/cer-website/), who commonly possess relevant psychometric expertise garnered through years of training and experience along with knowledge of chemical safety standards and practices. As with many undertakings in the field of chemistry, the project quality and impact can often be enhanced by forming a team of experts with relevant qualifications. We therefore strongly encourage researchers in the health and safety fields to consider collaborating with experts in survey design and data analysis.

Finally, we look forward to the researchers involved in chemical health and safety establishing their own sets of standards and norms in the area of survey development. Every field that employs surveys has unique considerations and approaches to assessing the evidence of validity and reliability to align with its specific contexts. For example, recent reviews and editorials have highlighted examples of how these standards have developed within the field of chemistry education.[31−36] We strongly encourage members of this field to approach this crucial task with the same dedication to detail and science employed in their chemistry.

■ AUTHOR INFORMATION

Corresponding Author

Qi Cui − *Division of Biological Sciences, University of California, San Diego, California 92093, United States;* orcid.org/0000-0002-3034-1143; Email: q1cui@ucsd.edu

Authors

Jordan T. Harshman − *Department of Chemistry and Biochemistry, Auburn University, Auburn, Alabama 36849, United States;* orcid.org/0000-0003-0783-465X

Regis Komperda − *Department of Chemistry and Biochemistry, San Diego State University, San Diego, California 92182, United States;* orcid.org/0000-0003-4837-7141

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chas.3c00111

Author Contributions

The manuscript was written through the contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

IRB, Institutional Review Board; AERA, American Educational Research Association; APA, American Psychological Association; NCME, National Council on Measurement in Education; PPE, personal protective equipment

■ REFERENCES

(1) Krosnick, J. A.; Presser, S. Question and Questionnaire Design. *Handbook of Survey Research*, 2nd ed.; Wright, J. D., Marsden, P. V., Eds.; Elsevier: San Diego, CA, 2009.

(2) Hamilton, L.; Ravenscroft, J. *Building Research Design in Education*; Bloomsbury Publishing: 2018.

(3) American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*; American Educational Research Association: Washington, DC, 1999.

(4) Hill, R. H., Jr Recognizing and understanding hazards—The key first step to safety. *J. Chem. Health Saf.* **2019**, *26* (3), 5−10.

(5) Converse, J. M.; Presser, S. *Survey questions: Handcrafting the standardized questionnaire*; Sage Publications, Inc.: 1986.

(6) Zinn, S. R.; Slaw, B. R.; Lettow, J. H.; Menssen, R. J.; Wright, J. H., II; Mormann, K.; Ting, J. M. Lessons Learned from the Creation and Development of a Researcher-Led Safety Organization at The University of Chicago. *ACS Chem. Health Saf.* **2020**, *27*, 114−124.

(7) Gaia, A. *Social Desirability Bias and Sensitive Questions in Surveys*. In Atkinson, P., Delamont, S., Cernat, A.; Sakshaug, J. W., Williams, R. A., Eds.; SAGE Research Methods Foundations: 2020.

(8) Ménard, A. D.; Flynn, E.; Soucie, K.; Trant, J. F. Accident Experiences and Reporting Practices in Canadian Chemistry and Biochemistry Laboratories: A Pilot Investigation. *ACS Chemical Health & Safety* **2022**, *29* (1), 102−109.

(9) Cronbach, L. *In Educational Measurement*, 2nd ed.; Thorndike, R., Ed.; American Council on Education: Washington, DC, 1971; p 443.

(10) Kane, M. T. An argument-based approach to validity. *Psychol. Bull.* **1992**, *112*, 527.

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

(12) Carmines, E. G.; Zeller, R. A. *Reliability and Validity Assessment*; Sage: Beverly Hills, CA, 1979.

(13) Messick, S. Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educ. Res.* **1989**, *18*, 5.

(14) Cronbach, L. J. *In Test Validity*; Wainer, H.; Braun, H., Eds.; Erlbaum: Hillsdale, NJ, 1988.

(15) Nunnally, J. C.; Bernstein, I. H. *Psychometric Theory, 3rd Ed.*; McGraw-Hill: New York, 1994.

(16) Messick, S., Ed. *Test Validation*, 3rd ed.; American Council on Education, National Council on Measurement in Education: Washington, DC, 1993.
(17) Anastasi, A. *Psychological Testing*; MacMillan: New York, 1988.
(18) Mohamed, M. M.; Abdul Aziz, H.; Abdul Manaf, N.; Lian See, T. Semi-Quantitative Chemical Expert Tool for Occupational Safety and Health (Use and Standards of Exposure of Chemicals Hazardous to Health) Regulations 2000. *ACS Chem. Health Saf.* **2023**, *30* (1), 9−20.
(19) Embretson, S. E. A general latent trait model for response processes. *Psychometrika* **1984**, *49*, 175.
(20) Willis, G. B. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*; Sage Publications: Thousand Oaks, CA, 2005.
(21) Rossete, C. A.; Ribeiro, M. G. Laboratory Technicians' Use and Interpretation of Hazard Communication Elements on Chemical Labels. *ACS Chem. Health Saf.* **2021**, *28*, 211−223.
(22) Murphy, K. R.; Davidshofer, C. O. *Psychological Testing: Principles and Applications*, 6th ed.; Prentice Hall: Upper Saddle River, NJ, 2005.
(23) DiStefano, C.; Zhu, M.; Mindrila, D. Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research & Evaluation.* **2019**, *14*, 20.
(24) Bandalos, D. L.; Finney, S. J. Factor analysis: Exploratory and confirmatory. In Hancock, G. R., Stapleton, L. M., Mueller, R. O., Eds.; *The reviewer's guide to quantitative methods in the social sciences*; Routledge/Taylor & Francis Group: 2019; pp 98−122.
(25) Liu, X.; Jin, X.; Wang, X. Influence of Chemistry Learning Motivation on Freshmen' Laboratory Safety Perception. *ACS Chem. Health Saf.* **2022**, *29* (6), 485−493.
(26) Read, B. R.; Zartl-Klik, A.; Veit, C.; Samhaber, R.; Zepic, H. Safety Leadership That Engages the Workforce to Create Sustainable HSE Performance. In *All Days*; SPE: 2010; p SPE-126901-MS.
(27) McIntire, S. A.; Miller, L. A. *Foundations of Psychological Testing: A Practical Approach*, 2nd ed.; Sage Publications: Thousand Oaks, CA, 2006.
(28) Bless, C.; Higson-Smith, C. *Fundamentals of Social Research Methods, an African Perspective*, 3rd ed.; Juta: Lansdowne, South Africa, 2000.
(29) Reed, P.; Marin, L. S.; Zreiqat, M. Impact of Toolbox Training on Risk Perceptions in Hazardous Chemical Settings: A Case Study from a Bleach Processing Plant. *ACS Chem. Health Saf.* **2023**, *30* (3), 129−138.
(30) Bock, R. D.; Gibbons, R. D. *Item Response Theory*; John Wiley & Sons: 2021.
(31) Barbera, J.; Harshman, J.; Komperda, R. The Chemistry Instrument Review and Assessment Library (CHIRAL): A New Resource for the Chemistry Education Community. *J. Chem. Educ.* **2023**, *100* (4), 1455−1459.
(32) Komperda, R.; Pentecost, T. C.; Barbera, J. Moving beyond Alpha: A Primer on Alternative Sources of Single-Administration Reliability Evidence for Quantitative Chemistry Education Research. *J. Chem. Educ.* **2018**, *95* (9), 1477−1491.
(33) Maric, D.; Fore, G. A.; Nyarko, S. C.; Varma-Nelson, P. Measurement in STEM education research: A systematic literature review of trends in the psychometric evidence of scales. *Int. J. STEM Educ.* **2023**, *10* (1), 39.
(34) Deng, J. M.; Streja, N.; Flynn, A. B. Response Process Validity Evidence in Chemistry Education Research. *J. Chem. Educ.* **2021**, *98* (12), 3656−3666.
(35) Lewis, S. E. Considerations on validity for studies using quantitative data in chemistry education research and practice. *Chem. Educ. Res. Pract.* **2022**, *23*, 764−767.
(36) Barbera, J.; VandenPlas, J. R. 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*; Bunce, D. M., Ed.; ACS Symposium Series; American Chemical Society: 2011; Vol. *1074*, pp 177−193.

https://doi.org/10.1021/acs.chas.3c00111
ACS Chem. Health Saf. 2024, 31, 121−126