

Response process validity evidence in chemistry education research

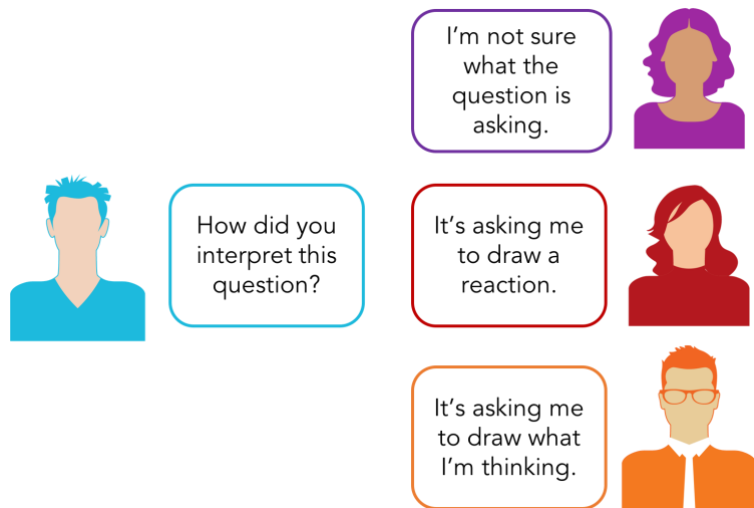
Jacky M. Deng, Nicholas Streja, and Alison B. Flynn*

Department of Chemistry and Biomolecular Sciences, University of Ottawa, Ottawa, ON, Canada, K1N 6N5

5 ABSTRACT

Response process validity evidence can provide researchers with insight into how and why participants interpret items on instruments (*e.g.*, tests, questionnaires). In the chemistry education research literature and in the social sciences more broadly, there has been variable use and reporting of response process aspects of studies. This manuscript's objective is to support researchers in
10 developing purposeful, theory-driven protocols to investigate response processes. We highlight key considerations for researchers who are interested in using cognitive interviews in their research, including: the theoretical basis for response process, collecting response process validity evidence through cognitive interviews, and using that evidence to inform instrument modifications.

GRAPHICAL ABSTRACT



15

INTRODUCTION

Instruments are invaluable tools of measurement for researchers that allow us to directly and indirectly observe otherwise inaccessible phenomena.^{1,2} For example, chemical instruments like spectrometers are regularly used to quantify physical and chemical entities and phenomena, with measurements providing the user with evidence related to the entities and phenomena being measured. Similarly, social science instruments—which include assessments, tests, surveys, and observation protocols—seek to measure variables within humans (respondents) directly and indirectly.^{1,3} Among the goals of discipline-based education research (DBER), one is to “help identify and measure appropriate learning objectives and instructional approaches that advance students toward those objectives”.⁴ As DBER has grown, so too have the number of instruments designed to investigate different variables of interest.^{5–10} Those instruments need to generate data of high validity and reliability. However, resources describing how to collect validity evidence in DBER contexts are limited.¹¹ This manuscript’s objective is to support researchers in developing and reporting protocols to investigate a specific type of validity evidence—response process.

In some cases, instruments may not yet exist for a particular goal, requiring instruments to be created and supported with appropriate validity evidence to determine whether they measure what they intend to measure. The effectiveness and suitability of social science instruments can vary depending on the goals and context of a particular evaluation. An instrument designed for a particular purpose in one context may need to be re-evaluated, re-configured, or re-designed to provide meaningful measurements in pursuit of new research goals and/or within a new context. For example, an instrument designed to measure a particular variable within a university context may produce less valid data when used in a high school context.

DEFINING VALIDITY

Validity and types of validity evidence

Validity can be considered analogous to the accuracy of an instrument. The validity of an instrument is a measure of the validity of *the data collected* by the instrument for a *specified respondent population*.¹² Collecting validity evidence is imperative in test development—without evidence for validity of the data obtained from a particular instrument in specific research context, there is uncertainty about what the instrument is actually measuring.¹³

Evidence for validity can be collected in several ways. The four main types of validity evidence are: test content, response process, internal structure, and relationships with other variables (Figure 1 **Error! Reference source not found.**).¹³ Each type of evidence can be collected using specific and established research methods. For example, evidence of test structure is often collected through quantitative means, such as factor analysis. In contrast, response-process validity evidence is often collected through qualitative inquiry (e.g., cognitive interviews, open-response questions).

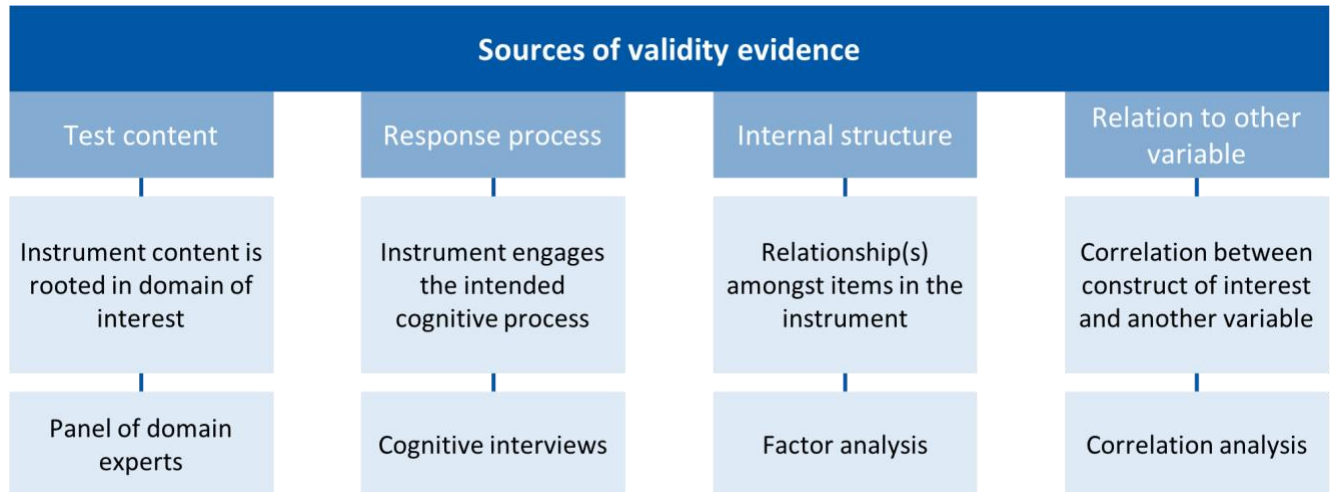


Figure 1: Sources of validity evidence, including examples of methods to collect each type of evidence.^{13,14}

This work provides insight into response process validity. We describe (1) this type of validity and an associated theoretical framework, (2) how response process validity has been reported in the literature to date, and (3) how to collect response process data, analyze the data, and use the findings to improve the validity of data collected from an instrument.

Response-process validity evidence

Response process represents one of the primary sources of validity evidence and is described as “evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees”.¹³ By investigating respondents’ response processes, researchers gain an understanding of how responses were generated, including what cognitive resources are activated, what barriers are faced by respondents, and whether interpretations of items differ from what developers intended.^{15–17} In the context of instrument development, response process evidence can provide researchers with insight into whether the questions they are asking are:

(1) activating intended constructs and prompting respondents to respond with those constructs in mind and (2) structured in a fashion that allows respondents to readily interpret and respond.

The theoretical basis for response process can be based in Tourangeau's four-stage cognitive model.¹⁸ According to this model, in order to accurately respond to an item, a respondent must be able to: (1) comprehend the item, (2) retrieve information relevant to the item's content, (3) make a judgement based upon the recall of knowledge relevant to the item, and (4) provide their answer within the given item structure. Breakdowns at each of these stages can be a result of various facts, including item wording, clarity, difficulty, and knowledge of relevant ideas held by respondents and/or their motivation.

The following is an example of how Tourangeau's model can be used to interpret response process data. A researcher may be interested in learning about students' self-efficacy in organic chemistry may design the item: "I am knowledgeable in acid–base chemistry", in which respondents are asked to respond on a Likert scale (*e.g.*, strongly disagree to strongly agree). Tourangeau's model states that responding to this item requires respondents to engage in the following cognitive processes: (1) comprehending and interpreting critical terms and phrases within the item, such as what the terms "knowledgeable" and/or "acid–base chemistry" mean, (2) identifying and retrieving the necessary cognitive resources and/or information relevant to the item, (3) choosing to report an accurate estimation of their belief within the given item structure, and (4) selecting or producing an answer that is feasible within the given item structure. Evidence **for** response process validity would mean that the respondent's understanding of the item matches the item's intended meaning and that the respondent is able to report an accurate account of their response within the given item structure. Evidence **against** response process validity can manifest as difficulties that arise during at least one of Tourangeau's stages. For example, a respondent believing that "acid–base chemistry" is too ambiguous and broad a term to provide an accurate response would provide evidence *against* response process validity for this item (challenge experienced in stage 1 of Tourangeau's model). Other examples of evidence against response process validity could include difficulties with understanding the item as a whole (stage 1); the item activating irrelevant ideas within respondents (stage 2); respondents overestimating or underestimating their ability in their final response (*e.g.*, feeling pressure to provide

95 a positive response; stage 3); and/or respondents feeling as if they are not able to respond accurately within the test structure (*e.g.*, wanting to elaborate on their response more openly; stage 4).

To understand the current state of response process interview reporting in chemistry education research (CER), we conducted a literature review of research articles that involved collecting response process evidence for creating or modifying instruments in chemistry education. We searched both the
100 *Journal of Chemical Education* and *Chemistry Education Research and Practice* using the queries “response process”, “concept inventory”, “instrument **and** response process”, “questionnaire **and** response process”, “survey **AND** response process”, “cognitive interview”, and limited the search results to articles published within the last ten years (2011–2021). Reviews and any studies that did not state the use of response process interviews were excluded, as we were specifically interested in how
105 these reports described their response process protocols, how data gathered from investigations were analyzed, and how these data were used to guide instrument modifications (if needed). Co-author NS conducted the preliminary review and coded the protocol, analysis, and modifications described in each article in terms of three categories: *none reported*, *general descriptions provided* (*i.e.*, the authors mention an analysis, but do not provide explicit details), and *specific descriptions provided* (*i.e.*, the
110 authors explain their process and resulting decisions with details). For modifications, we captured whether the authors stated that *no modifications were needed* following response process investigations. After NS’ coding, co-author JMD then reviewed the same articles and conducted the same analysis; the co-authors discussed the codes and analysis until they reached agreement. The codebook and exemplar quotes from each article reviewed are described in the Supporting
115 Information.

The findings from the 39 articles reviewed suggest that how response process validity evidence is reported in research investigations varies widely within CER (Figure 2), reinforcing previous work evaluating the state of measurement in CER. A 2013 study found that of 20 instruments described in 37
120 publications, only four instruments reported response-process validity evidence, with only two of these four providing specific details about protocols and the nature of changes to instruments due to response process evidence.¹⁹ This trend appears to be present in other research fields, as well: a 2011 review found that few research studies explicitly used or described response process evidence as a source of substantive validity evidence, and that this trend has remained relatively static in the last 20

years compared to other sources of validity evidence.²⁰ Similarly, a 2008 review of validity
 125 investigations in psychology and education found response process was only described for 1.8% of
 measures.^{21,22} Other reviews have suggested that the majority of work reporting response process
 evidence were validation studies in health research.²³ Explicitly describing how evidence for response
 process is gathered and used to inform changes to instruments will be essential for developing a
 broader research community understanding of the validity process and of constructs measured by a
 130 given instrument, and also deepens understanding of what is being measured by an instrument.^{19,24}

The findings from our and others' reviews suggest a need for resources to support researchers in
 collecting and reporting response process validity evidence. From our review, cognitive interviews
 were the most common method for collecting response process validity evidence (Supporting
 Information). Therefore, the following sections will describe how researchers can design and enact
 135 cognitive interviews to systematically collect and report validity evidence based on response process.

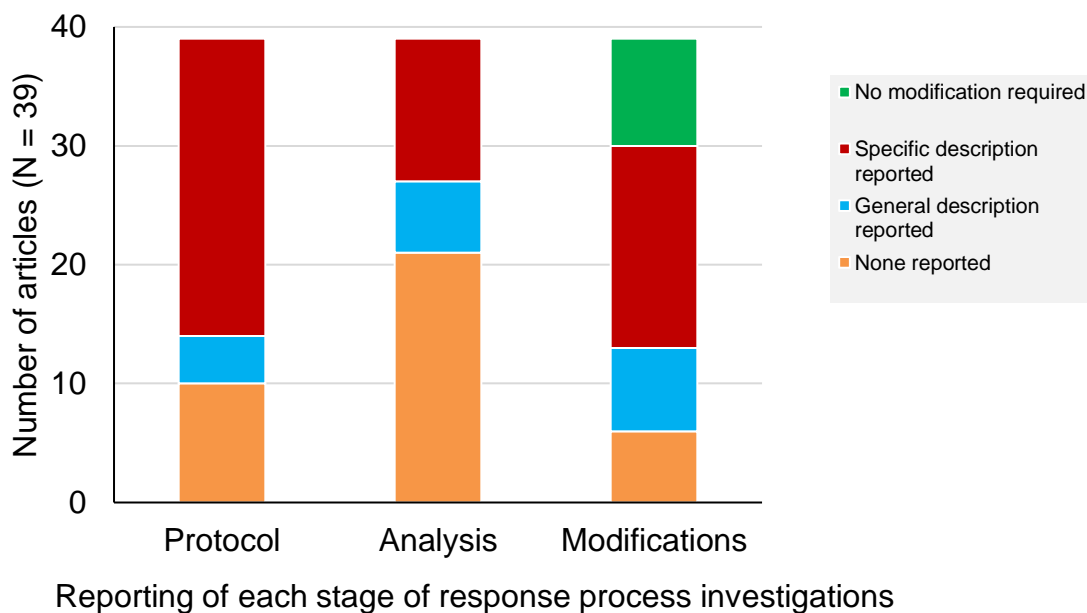


Figure 2: Overview of how response process evidence investigations protocols, analyses, and modifications have been described in CER literature between 2011 to 2021, $N = 39$.

COGNITIVE INTERVIEWS TO COLLECT RESPONSE PROCESS VALIDITY EVIDENCE

Cognitive interviews can be used in instrument development to inform item revision decisions by providing evidence of validity based on both test content (clarity and relevance of items)²⁵ and response process (thought processes involved in responding to an item).^{13,26,27} When the goal is to collect evidence for response process validity, the focus of cognitive interviews (often called response process interviews) is to study how respondents cognitively process (think about) and respond to materials presented to them, with a specific emphasis on what barriers and pitfalls arise during response process.^{18,28} From these inquiries, qualitative data can be obtained that illuminates respondents' responses and behaviors, allowing researchers to determine whether item interpretations are (mis)aligned with developers' intentions and to then identify ways to modify items to support validity.^{29,30} When completed early in instrument development, cognitive interviews enable researchers to identify and correct issues in understanding and content before resources may be devoted to more extensive validity testing (*e.g.*, factor analysis). Evidence of response process can range from collecting and indexing respondents' statements about their response, to using theory to interpret respondents' statements in an effort to formulate generative mechanisms of how and why respondents decided to respond in the way(s) they did. In the following sections, we will discuss practical considerations for researchers planning to use cognitive interviews to collect evidence for response process, including designing and organizing protocols, and interpreting and using data acquired from interviews to inform instrument modifications.

160 Designing cognitive interview protocols

Cognitive interviews are multi-stage process involving the following: identifying item intent, data collection, analysis, and comparison of respondent interpretation to intended meaning (Table 1). In the first stage, researchers develop items for a particular instrument and document the intent of each item prior to cognitive interviewing, as well as any associated constructs the item is intended to investigate. These descriptions serve as the basis for comparison against respondents' interpretations.

Table 1: Example of the cognitive interview process. Adapted from Peterson et al., 2017.

Stage 1: Prepare item intent and any associated construct dimensions	Stage 2: Develop interview protocol	Stage 3: Conduct cognitive interviews	Stage 4: Analysis of interview data	Stage 5: Compare to intent and dimensions; decide on modification strategy
<p>Item: I am knowledgeable in acid–base chemistry (1-Strongly Disagree to 5-Strongly Agree)</p> <p>Construct: self-efficacy</p> <p>Intent: Assess respondent’s perception of self-efficacy beliefs related to acid–base chemistry</p>	<p>Anticipated relevant cognitive operations: Comprehension</p> <p>Probe: What does the word “knowledgeable” mean to you?</p>	<p>Respondent 1: In this case, knowledgeable was a bit vague. I know that acids donate protons and bases accept protons, but I’m not sure I could apply these ideas.”</p> <p>Respondent 2: Knowledgeable means that I can do well on a test, so I would say I am knowledgeable because I did well</p>	<p>Key phrases: “a bit vague”; “means that I can do well on a test”; “tough to say without more information”</p> <p>Summary: Respondents have various interpretations of the item and some feel the item is not specific enough to report an accurate judgement</p>	<p>Conclusion: The item’s intention was to assess respondents’ self-efficacy beliefs related to acid–base chemistry. Data show that the item is too vague for respondents to share a consistent interpretation.</p> <p>Recommendation: Revise the item to focus on a specific learning outcome related to acid–base chemistry.</p>

		<p>on my last chemistry exam.</p> <p>Respondent 3:</p> <p>I feel like I am [knowledgeable] because I think I understand the main ideas, but it's tough to say without more information.</p>		<p>For example, "I can identify the most acid proton and the most basic atom in an acid–base reaction".</p> <p>Conduct another round of cognitive interviews to investigate the impact of modification.</p>
--	--	--	--	---

Once items have been prepared, an interview protocol can be developed to gather information on respondents' response processes. A general framework based on Tourangeau's four-stage cognitive model of response process can be used to guide this process.¹⁸ Specifically, prompts may seek to address the following questions:

- 175
1. How do you (the respondent) interpret the item?
 2. What information do you believe to be relevant for responding to the item?
 3. How and why did you decide on their given response?
 4. What challenges did you face when providing their response within the given item structure, if any?

180 A list of potential prompts to address each stage of response process can be found in the Supporting Information.

Most protocols in current CER literature focus on addressing Question 1, with interviews being used evaluate the readability and clarity of test items by asking students to read items aloud and then

explain the meaning in their own words (Supporting Information). Other response process
185 investigations have sought to develop rich, causal understandings of the how and why behind response
processes through purposeful integration of theoretical and methodological frameworks. For example,
Hosbein and Barbera (2020) asked respondents to not only read the item aloud and describe their
interpretation, but also to choose an answer and describe why they chose the answer they did. The
extent to which researchers choose to investigate and analyze response process will depend on the
190 goals of the research, as well as the feasibility within the context of the investigation.^{23,24,26,31,32}

When designing cognitive interview protocols, researchers may also choose to focus not only
cognitive processes, but also other, non-cognitive processes involved in answering test items, such as
self-referential, motivational, affective, situated, and dialogical response processes.³² For example,
methods can focus not only on what and why respondents are engaging in a particular response
195 process cognitively, but also how motivated respondents feel during the test-taking experience (*e.g.*,
fatigue) or have particular affective feelings towards particular items.

Herein, we provide four general approaches to cognitive interviews (Table 2).^{33,34} The four
approaches can be broadly categorized within the continuum of think-aloud (TAP) and verbal probing
protocols—they differ in terms of their relative flexibility/rigidity, resulting in distinct strengths and
200 weaknesses depending on the goals of the research.³⁴

205

210

Table 2: Summary of qualitative interview approaches, including examples in a response process interview context. Adapted from Patton (2015) and Willis (2011).

Approach	Key features	Advantages	Disadvantages
<i>Informal conversational</i>	No pre-planned structure; researcher generates questions spontaneously within flow of conversation	Interview can be tailored to each participant and build off of previous interviews	Identifying trends in data can be difficult; increased risk of interviewer effects, so requires skilled interviewer
<i>Interview guide</i>	Specific questions/topics prepared before interview, with freedom to investigate spontaneous lines of inquiry	Researcher can focus conversation on several key topics, while still being flexible enough to maintain rapport and engage in details of inquiry	Increased rigidity can limit the scope of the interview, such that unanticipated key perspectives could be missed
<i>Standardized, open-ended</i>	List of open-ended questions prepared before interview; researcher follows this list with limited deviation	Focused questioning ensures that all participants are asked the same questions, making data easier to analyze	Increasing rigidity further, so increased possibility that unanticipated key perspectives are missed

Fixed-response

List of fixed-response questions prepared before interview; researcher presents questions and options to interviewee	Fixing questioning and response options minimizes variability between responses; minimizes onus on interviewer to manage the conversation and often easier to collect data from larger samples to identify trends in response process	Very rigid; no opportunities to investigate response process perspectives beyond those provided to respondents
--	---	--

220 The *informal conversational approach* is the most open-ended approach to interviewing. These types of interviews are characterized by their lack of pre-planned structure and reliance on the researcher to generate questions within the flow of the interaction. The main strength of the conversational interview is its flexibility; questions can be specific to the interaction, allowing for deepened communication with the participant. Conversational interviews also allow the interviewer to take information from previous interviews to guide questioning in later interviews. The main weakness of 225 informal conversational interviews is the difficulty involved in comparing and analyzing the data from interviews. The variability in questioning between interviews places the onus largely on the interviewer to generate useful data. The conversational interviewer must be skilled at generating rapid insights and questions. Because of their lack of structure, conversational approaches may be more susceptible to interviewer effects, leading questions, and biases, when unaccounted for, each threatening the 230 integrity of the data collected from interviews.³⁵ Patterns and trends within the dataset may also take longer to identify from conversational interviews, as responses from participants will vary as different participants are asked different sets of questions during their respective interviews.

Open-ended approaches may complicate the data collected from cognitive interviews, as researchers may be unsure where and how to make changes to an instrument if there is a large amount of variability in the topics discussed by interviewees. However, the flexibility associated with the informal conversational approach may also be useful in investigating response process, specifically in research contexts where multiple rounds of interviews are feasible. For example, interviewers interested in where participants experience confusion when completing an instrument might begin interviews in an unstructured fashion with an initial round of interviews to allow researchers to identify items that are variably interpreted between respondents. Researchers can then use the information from these initial interviews to develop more structured, targeted questioning in future rounds of interviews.

The *interview guide approach* is characterized by researchers preparing a list of specific questions and/or issues to be explored prior to the interview. This approach ensures that each interviewee is confronted with the same general subjects of inquiry; however, the researcher maintains a degree of flexibility in this approach; one is free to formulate spontaneous responses to interviewee responses and establish a conversational approach within the confines of the predetermined subject of inquiry. The main strength of the interview guide approach is that it provides the interviewer with a basic map or framework of how to facilitate the interview. This makes data collection more systematic across the sample. However, though there is some flexibility in this approach, the increased rigidity of the interview guide approach increases the risk of omitting important topics outside the purview of the protocol.

For response process investigations, the interview guide approach can be useful if the interviewer is interested in interviewees' thoughts and processes on specific instrument items. Researchers might have anticipated concerns about specific items, so they can focus their questioning during interviews on these items, while leaving room for interviews to discuss other issues and/or items. The interview guide approach can also be useful for cognitive interviews involving multiple rounds of interviews.³⁶ Researchers may choose to employ a conversational approach in initial rounds, allowing interviewees the freedom to discuss whatever concerns and interpretations come to mind during their response processes. Researchers could then identify any consistent concerns from these initial rounds, using this information to formulate an interview guide approach for subsequent rounds.

The *standardized open-ended approach* is characterized by careful and complete wording of each interview question prior to the interview. This approach imposes greater structure on the interview by not only fixing the general organization of the interview, but also fixing the wording of each question and the sequence of questions within the interview. However, questions are still open-ended, meaning that interviewees construct their own responses instead of choosing a response from a selection of pre-determined possibilities. Like the interview guide approach, one of the strengths of the standardized open-ended approach is that interviewees answer the same questions, limiting the variability between interviews and allowing researchers to more easily organize/analyze the data and compare between interviewees. However, the rigidity of this approach limits flexibility; the interviewer is granted few opportunities to pursue unanticipated topics or issues introduced during the interview, which can result in the loss of important data.

In cognitive interviews, a standardized open-ended approach may be useful if the interviewer wanted to learn about interviewees' perspectives on specific aspects of an instrument and/or a pre-determined selection of instrument items. The researcher might identify a selection of items to discuss during the interview and ask each interviewee the same questions about each item. This would allow the researcher to gather targeted information about specific items and ensure that all interviewees have an opportunity to respond to the same set of questions during the interview.

In the *fixed-response approach*, questions and response options are determined in advance of the interview. Each interview follows exactly the same sequence of questions, and interviewees are instructed to select fixed responses instead of constructing their own. This approach allows for simpler data collection and analysis, and also affords the interviewer with the opportunity to ask many questions in a short amount of time. However, the fixed-response approach constrains the opportunities afforded by qualitative inquiry. Because respondents must fit their thoughts and experiences within pre-determined categories, the approach may misrepresent respondents' actual thoughts and experiences. Therefore, we suggest that researchers avoid relying on the fixed-response approach as the initial approach in gathering evidence for response process.

For cognitive interviews, the fixed-response approach may be useful in cases where researchers have already determined specific respondent concerns from initial rounds of interviews and want feedback on whether updates to items have remedied these specific concerns. For example,

researchers can use an open approach in an initial round of interviews to determine common issues expressed by respondents and then make initial changes to test items to address these initial concerns. The researchers can then ask additional rounds of respondents whether these points of confusion persist or not using version of the instrument that multiple-choice options based on interviewees' concerns from the previous round (*e.g.* "Did you experience any of the following as you answered the item?").

The four approaches described above exist on a continuum and researchers are certainly not limited to selecting only one the approaches in their research. The approaches described can also be combined to develop a more comprehensive understanding of response process. For example, one study conducted open interviews with an initial sample of respondents to identify initial points for revision, and then evaluated the impact of these revisions through targeted pilot administrations of their instrument.³⁷ Protocols can also be combined within a single interview; researchers might adopt a guided or standardized approach in the first part of the interview protocol to focus on specific items, and then allow the interviewee to freely address any additional items or concerns during the latter part of the interview.

Cognitive interviews can be conducted either concurrently or retrospectively. Concurrent cognitive interviews require interviewees to verbalize their experiences and response processes as they work through test items. For example, one study provided participants with a copy of their test and, as they worked through the test, asked each participant to read each item aloud, explain what was being asked, choose an answer, and explain their choice.⁹ A benefit of concurrent protocol is that insights from the interviewee should be more accurate than if the interviewee were recalling their mental processes at a later point in time. However, by asking interviewees to verbalize their cognitive processes, researchers risk disrupting participants' response processes, meaning that the response process expressed by the interviewee may be less similar to the response process of an actual test respondent.³⁴ For example, asking interviewees to verbalize their response process may increase cognitive demand and introduce response process disruptions that would not be encountered if respondents were to complete the test on their own.

Cognitive interviews can also be conducted retrospectively. In this case, interviews are conducted after the interviewee has had a chance to complete the instrument without disruption. For example,

320 one study asked respondents to complete the survey in full before engaging in a questioning period.³⁸
However, though retrospective interviews ease the cognitive demand on interviewees and allow them
to complete the test without disruption, interviewees may be more likely to forget and introduce bias
as they try to recall their response process at a later point in time.

One way to mitigate the challenges associated with each of these approaches is to combine the
325 two approaches. One study conducted interviews that incorporated both concurrent and retrospective
components;³⁷ in this case, respondents were asked to actively think-aloud as they answered test
items (concurrent). After they had completed the items, the researchers then asked respondents about
their interpretations of test items (retrospective).

Sampling

330 Validity is a measure of the validity of the data collected by the instrument for a specific target
population, so participants invited to cognitive interviews should be as closely aligned as possible with
members of the intended test-taking population.¹³ Cognitive interviews should seek a diversity of
perspectives; therefore we suggest that sampling for cognitive interviews is most effective when
restricted to the test-taking population, but recruitment from this population is random and/or
335 purposeful with diversity in mind. As with any research with human subjects, Research Ethics Board
approval or exemption must be obtained before recruiting for and conducting cognitive interviews.

Cognitive interviews are a qualitative method of inquiry, so sample sizes will vary depending on the
specific goals of investigations. From CER literature describing the use of cognitive interviews, sample
sizes have ranged from 5 to 25 (Supporting Information).^{9,37-43} Recommendations from other sources
340 range from 5 to 15,⁴⁴⁻⁴⁶ with some research suggesting that increasing sample sizes beyond this range
does not seem to uncover new response process problems.⁴⁷

Using sequential rounds of interviewing can also be advantageous when collecting evidence of
response process,^{28,36} as some have done in CER.^{9,37,48} This strategy is particularly useful in instrument
development because it provides researchers an opportunity to investigate the impact of instrument
345 modifications following initial interviews. For example, researchers may interview five respondents and
make changes to an instrument based on the evidence gathered from this first round. To better
understand the impact of these changes, the researcher can then conduct a second round of
interviews to identify (1) whether instrument modifications have resolved or exacerbated issues from

the first round and (2) any new challenges faced by respondents in the second round of interviews.

350 Other options can also include conducting a round of interviews to learn about respondents' interpretations of items, revising items based on these interviews, and then conducting a pilot study by administering the revised instrument to samples from a target population to.³⁷ However, conducting interviews in rounds can mean that true sample size for a response process investigation may not be known at the beginning of the study. Researchers may begin their research with a specific sample size
355 in mind but that number can be adjusted based on what is learned from interviews, as inquiry deepens, or when data saturation is observed.⁴⁹

ANALYZING COGNITIVE INTERVIEW DATA TO INFORM INSTRUMENT MODIFICATIONS

Of the three stages of response process investigations, analysis of response process data was found to have the highest frequency of *none reported*. Variable reporting of response process data analysis
360 has also been reported in other reviews of response process investigations in social science literature.²⁰⁻²² There is variation in analysis methods for cognitive interviewing.⁴⁵ Analysis of cognitive interview data can be less intensive, characterized by interviewers taking notes of problems during the interview process and comparing notes across interviews.⁵⁰ For example, one analysis is described as follows: "If the rationale students provided was not aligned with the researcher's intended meaning,
365 the item was flagged for modification or removal. If students gave a rationale that was aligned with the intended meaning, there was validity evidence that the item was being interpreted as intended and could be used in future implementations of the scale in the given context".⁵¹ In contrast, analyses involving coding schemes and classifying response process issues taxonomically have also been employed.^{36,52} For example, one analysis stated that "audio files of the interviews were transcribed by
370 a commercial transcription service. All analysis was conducted on the interview transcripts, led by author Hosbein. Cognitive interview analysis was used by author Hosbein to establish a coding rubric. This entailed carefully reading each transcript and documenting the language students used in explaining their responses. These reviews were used to establish initial codes regarding students' scale-based language... Author Hosbein and a secondary (non-author) researcher independently coded all 40
375 transcripts using the initial rubric. After discussing discrepancies in coding, the rubric was revised and additional transcripts were independently re-coded. The final rubric had an acceptable Cohen's kappa value of 0.8 and was used by author Hosbein to re-code the remaining transcripts."⁵³ Research on the

range of analytical methods employed has identified benefits and challenges with both—less intensive analyses unearth fewer response problems and may be more prone to bias,^{50,54} while more intensive analyses may not be feasible—one may simply lack the resources (*e.g.*, time, funding) to engage in a full qualitative analysis of response process data, especially if instrument administration is time-dependent.³⁶ In response to the ambiguity in the literature about what is “best” and to serve a range of research contexts, researchers have devised qualitative methods that seek to balance feasibility and reliability—we adapted one such approach for chemistry education context in Table 1.^{26,45}

There is limited literature guidance on how to modify items based on cognitive interview data. In CER, most response process investigations have focused on item readability and clarity, resulting in most modifications being relatively minor changes to item wording. For example, one study described how: "Several small changes were made to the wording of some of the survey items. One such change involved the words 'excited' and 'intrigued' in some of the survey items; students indicated that such words were too enthusiastic, and that the word 'interested' would be more fitting, which reinforced the importance of such item wording".⁵⁵ Though there is no single or “best” approach for analyzing and modifying items based on cognitive interview data, readers will benefit from the explicit descriptions of such investigations whenever possible. To best highlight how and why students interpret items during instrument development/modification for the research community, reports can explicitly describe:

1. Who is sampled to participate in any response process investigations?
2. What protocols are enacted and what is the rationale for these protocols?
3. Why are data analyzed and used to inform instrument modifications (if at all)?

Explicitly describing how evidence for response process is gathered and used to inform changes to instruments will be essential for developing a broader research community understanding of the validity process and of constructs measured by a given instrument, and also deepens understanding of what is being measured by an instrument.^{19,24}

CONDUCTING COGNITIVE INTERVIEWS REMOTELY

There are many opportunities afforded by conducting interviews remotely, including: hosting instruments online, minimizing video and audio recording costs, and allowing researchers to record participants’ drawings as they speak.⁵⁶ In this section, we highlight some considerations for conducting cognitive interviews remotely.

First, many online survey methods allow tests to be completed virtually (*e.g.*, SurveyMonkey, Qualtrics). Hosting tests on these online programs can be beneficial for facilitating response process interview dialogue when screen-sharing is possible. With screen-sharing, both the interviewer and the interviewee are able to view the relevant items simultaneously, similar to in-person.

Videoconferencing software with built-in recording capabilities also make it easier to manage audio and video data from interviews. For example, Zoom can be used to simultaneously record audio and video during interviews, which may allow researchers to dedicate more focus to interviewees' statements, notetaking, and formulating questions in more dynamic, think-aloud approaches.

Recording audio and video simultaneously may be more cost-effective than using separate devices (*i.e.*, video camera, microphone).⁵⁶ Other useful options include the ability to transcribe audio for cloud recordings on Zoom, which can streamline interview analyses.⁵⁷ However, care must be taken with regards to research ethics in these cases, as cloud storage of interview data may conflict with existing institutional review board (IRB) principles and approvals. Hosting the discussion virtually may also be beneficial in reducing interviewees' feelings of pressure, which may afford more direct articulations of participants' response processes.⁵⁶

There are also many options available for drawing molecules in an online setting, critical to many areas of chemistry. Participants can draw using pen and paper with either a webcam pointing to the page or by taking a photo/scan and sending to the research (*e.g.*, email, upload to a secure folder, drop into the videoconference chat). They could draw using a notepad app (*e.g.*, Drawboard, Notability).^{58,59} Whiteboard apps offer a combination of options, including drawing, photo sharing, and collaboration (*e.g.*, Miro, Mural; Figure 3).^{60,61} With all the electronic approaches, participants can readily share screen with the interviewer. If participants are working collaboratively, the interviewer can mute and turn their own video off to fade into the background. Offering a few different options provides opportunities for students with variable access to technologies to participate equitably.

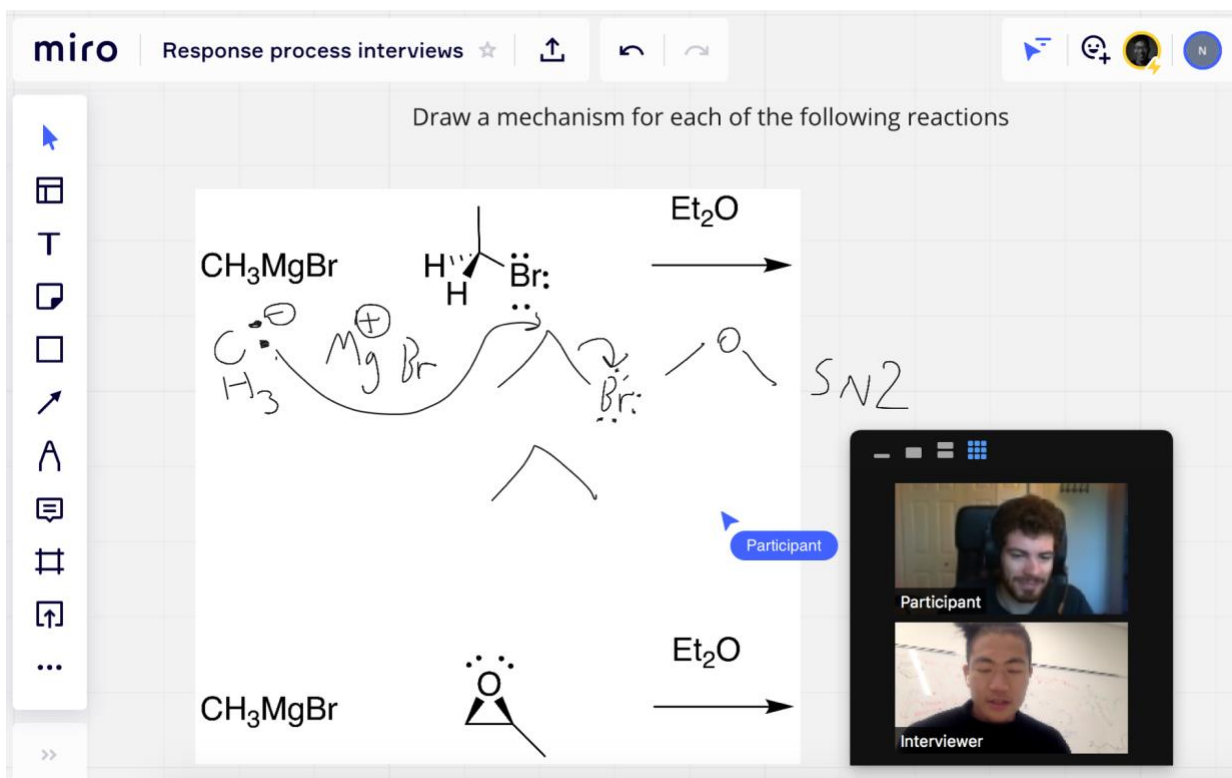


Figure 3: Example of using whiteboard apps to facilitate virtual interviews.

ALTERNATIVES TO COGNITIVE INTERVIEWS

Interviews in particular can be resource-intensive and subject to threats and limitations such as interviewer effects, disproportionately motivated participants, and small sample sizes. Although
 435 cognitive interviews have been the predominant method to investigate response process throughout the social sciences and discipline-based education research, other methods exist that may be more appropriate and feasible for a particular context. Alternatives that can be used instead of or in parallel with response process interviewing to develop more comprehensive understandings of response
 440 processes include:

- *Open-response questions*, which prompt respondents to provide their thoughts, opinions, beliefs, *etc.* about the test-taking experience. This method also allows for greater number of participants if administered in the form of an online survey, for example.^{62–64}
- *Vignettes*, which are short stories or descriptions of a hypothetical respondent used to
 445 investigate a subject's cognitive processing during instrument-related decisions.^{65–67}

Vignettes can be valuable when a small number of key comprehension issues are identified

and targeted at the outset of the validity investigation. For example, if researchers were interested in how respondents interpret and respond to the item “I identify as a visible minority in chemistry”, the researchers might provide a vignette of a hypothetical individual to respondents and ask their opinion on how this hypothetical individual should respond.

- *Card sort tasks*, in which respondents are provided a set of concepts and prompted to organize them in a closed sort (within pre-determined themes) or open sort (without pre-determined themes).⁶⁸ Card sort tasks are valuable in learning how a respondent organizes concepts or items and what a respondent believes a concept or item includes or excludes. For example, if researchers were building an instrument that measured students’ knowledge of different spectroscopic methods, they might ask respondents to organize instrument items in terms of the different methods (*e.g.*, NMR, IR) to determine how respondents are perceiving the content of the items.
- *Focus groups* can be used to facilitate dialogue between respondents, which can be helpful to investigate respondents’ perceptions about an instrument at a general level and their thinking about the concepts/themes of the instrument (unlike interviews which can be used to focus on specific items). Focus groups have the advantage that participants can add to each other’s ideas (*e.g.*, build upon, agree, disagree) but have the simultaneous disadvantage that participants can be influenced by the others’^{69–71}.
- *Eye-tracking* can be used to study where participants look, in what order, and the amount of time they spend looking various elements of an item. Eye-tracking has also been used to collect data on participants’ pupil dilation, item fixation and duration alongside think-aloud interviews to evaluate challenges during response process.^{72–76}

COLLECTING RESPONSE PROCESS VALIDITY EVIDENCE IN TEACHING

Collecting evidence for validity for course assessments can help ensure that inferences about student achievement of learning outcomes are as accurate as possible.^{13,77} Specific to response process, instructors may speak with several students to gain insight into students’ response processes, similar to the cognitive interviews described above. Instructors may also provide opportunities for students to describe how they arrived at their answer or how they approached specific questions on an

assessment (e.g., providing an option to write out their thinking after a multiple-choice question).^{77–79} By analyzing/evaluating students' responses, instructors gain insight into students' response processes, which can inform how to evaluate responses and/or tailor instruction. Instructors can also ask similar types of questions throughout the sequence of a course and refine the questions over time. For example, instructors can ask a question across different formative and summative assessments and review student responses to refine the question as the course progresses, similar to conducting cognitive interviews in rounds (Figure 4).

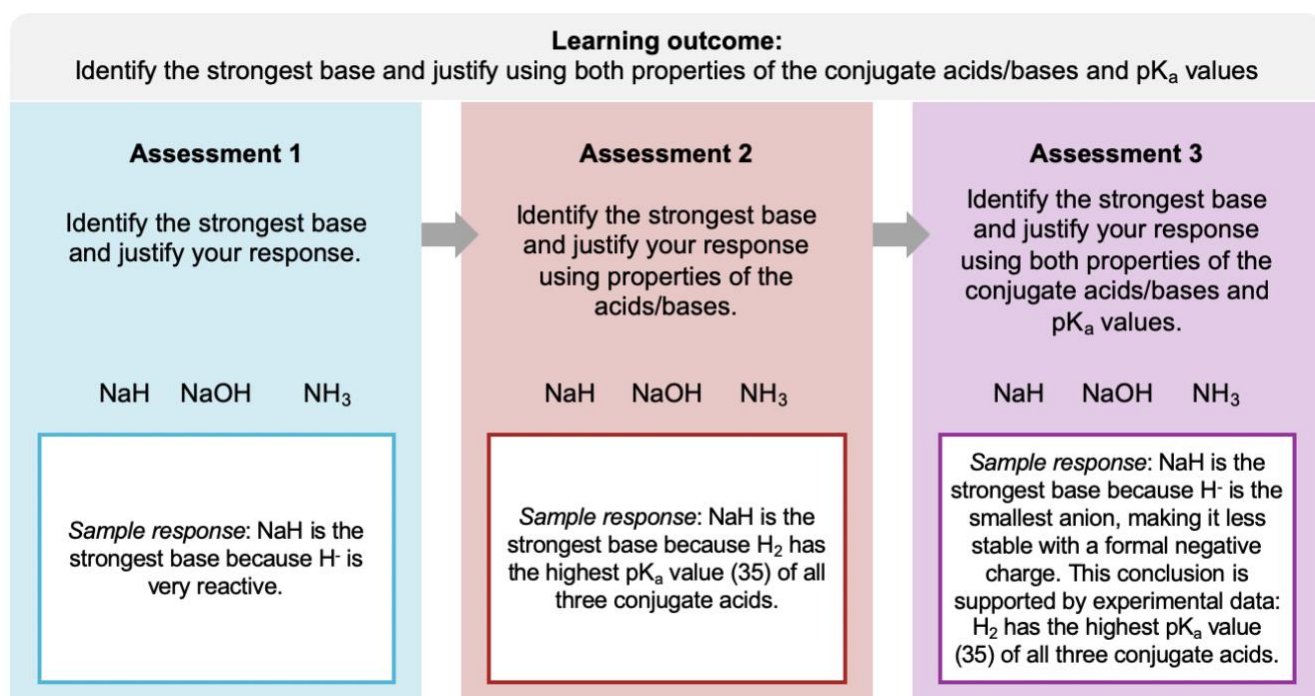


Figure 4: Example of revising items based on response process evidence gathered from students' responses across multiple assessments.

CONCLUSIONS

In this manuscript, we highlight key considerations using cognitive interviews to enhance validity, including: (1) how Tourangeau's four-stage cognitive model can be used to guide response process investigations,¹⁸ (2) different approaches to cognitive interviewing and the advantages/challenges associated with each, and (3) suggested practices when analyzing and using cognitive interview data to inform instrument modifications. Cognitive interviews can contribute to the piece of an instrument's overall validity argument by providing evidence of respondents' response processes. Transparent

reporting of interview methods, analysis, and results are necessary to further solidify the validity evidence of an instrument. We examined 39 CER articles in which the authors used cognitive interviews to collect response process validity evidence. There was variation in how cognitive interviews were reported, specifically with respect to interview protocols, analysis of interview data, and/or resulting modifications to instruments. This manuscript's goal is to support researchers in improving the validity of data obtained from research and educational instruments, developing purposeful protocols to collect and refine instruments based on evidence on students' response processes, and ultimately improving the quality of data collection toward improving research outcomes.

ASSOCIATED CONTENT

Supporting Information

The SI contains a summary of response-process validity reporting in CER articles and a list of possible prompts to probe response process during cognitive interviews.

AUTHOR INFORMATION

Corresponding Author

*E-mail: alison.flynn@uottawa.ca

ACKNOWLEDGMENTS

Co-author JD acknowledges the Vanier Canada Graduate Scholarship (CGS) Program for funding.

REFERENCES

- (1) Opfer, J. E.; Nehm, R. H.; Ha, M. Cognitive Foundations for Science Assessment Design: Knowing What Students Know about Evolution. *J. Res. Sci. Teach.* **2012**, *49* (6), 744–777.
- (2) Campbell, C. E.; Nehm, R. H. A Critical Analysis of Assessment Quality in Genomics and Bioinformatics Education Research. *CBE Life Sci. Educ.* **2013**, *12* (3), 530–541.
- (3) Bollen, K. A. Latent Variables in Psychology and the Social Sciences. *Annu. Rev. Psychol.* **2002**, *53* (February 2002), 605–634.
- (4) National Research Council. *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering*; Washington, D.C., 2012.

- 520 (5) Sirum, K.; Humburg, J. The Experimental Design Ability Test (EDAT). *Bioscene* **2011**, *37* (1), 9–16.
- (6) Smith, M. K.; Wood, W. B.; Knight, J. K. The Genetics Concept Assessment: A New Concept Inventory for Gauging Student Understanding of Genetics. *CBE Life Sci. Educ.* **2008**, *7* (4), 422–430.
- (7) Denofrio, L. A.; Russell, B.; Lopatto, D.; Lu, Y. Mentoring: Linking Student Interests to Science
525 Curricula. *Science (80-.)*. **2007**, *318* (5858), 1872–1873.
- (8) Smith, M. K.; Jones, F. H. M.; Gilbert, S. L.; Wieman, C. E. The Classroom Observation Protocol for Undergraduate Stem (COPUS): A New Instrument to Characterize University STEM Classroom Practices. *CBE Life Sci. Educ.* **2013**, *12* (4), 618–627.
- (9) Atkinson, M. B.; Popova, M.; Croisant, M.; Reed, D. J.; Bretz, S. L. Development of the Reaction
530 Coordinate Diagram Inventory: Measuring Student Thinking and Confidence. *J. Chem. Educ.* **2020**.
- (10) Wren, D.; Barbera, J. Gathering Evidence for Validity during the Design, Development, and Qualitative Evaluation of Thermochemistry Concept Inventory Items. **2013**.
- (11) Reeves, T. D.; Marbach-Ad, G. Contemporary Test Validity in Theory and Practice: A Primer for
535 Discipline-Based Education Researchers. *CBE Life Sci. Educ.* **2016**, *15* (1), 1–9.
- (12) National Research Council. *How People Learn: Brain, Mind, Experience, and School*; National Academies Press: Washington, D.C., 2000.
- (13) American Psychological Association; American Educational Research Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*;
540 American Educational Research Association: Washington, DC, 2014.
- (14) Lapierre, K. Exploring the Influence of a Redesigned Organic Chemistry Curriculum on Students' Interpretations of Reactions and Self-Efficacy Beliefs in Organic Chemistry, University of Ottawa, 2019.
- (15) Hughes, K. Comparing Pretesting Methods: Cognitive Interviews, Respondent Debriefing, and
545 Behavior Coding. *Surv. Methodol.* **2004**, *2*, 1–20.
- (16) Presser, S.; Couper, M. P.; Lessler, J. T.; Martin, E.; Martin, J.; Rothgeb, J. M.; Singer, E. Methods for Testing and Evaluating Survey Questions. *Public Opin. Q.* **2004**, *68* (1), 109–130.
- (17) Ryan, K.; Gannon-Slater, N.; Culbertson, M. J. Improving Survey Methods With Cognitive

Interviews in Small- and Medium-Scale Evaluations. *Am. J. Eval.* **2012**, *33* (3), 414–430.

- 550 (18) Tourangeau, R.; Rips, L. J.; Rasinski, K. *The Psychology of Survey Response*; Cambridge University Press: New York, New York, USA, 2000.
- (19) Arjoon, J. A.; Xu, X.; Lewis, J. E. Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence. *J. Chem. Educ.* **2013**, *90* (5), 536–545.
- 555 (20) Zumbo, B. D.; Shear, B. R. The Concept of Validity and Some Novel Validation Methods. In *Northeastern Educational Research Association*; Rocky Hill, CT, 2011; p 56.
- (21) Cizek, G. J.; Bowen, D.; Church, K. Sources of Validity Evidence for Educational and Psychological Tests: A Follow-up Study. *Educ. Psychol. Meas.* **2010**, *70* (5), 732–743.
- (22) Cizek, G. J.; Bowen, D.; Church, K. Sources of Validity Evidence for Educational and Psychological Tests. *Educ. Psychol. Meas.* **2008**, *68* (3), 397–412.
- 560 (23) Padilla, J.; Benítez, I. Validity Evidence Based on Response Processes. *Psichothema* **2014**, *26* (1), 136–144.
- (24) Sorana Laneanu, M. Response Processes as a Source of Validity Evidence in Self-Report Measures: Theoretical and Methodological Implications, The University of British Columbia (Vancouver), 2016.
- 565 (25) Dumas, H. M.; Watson, K.; Fragala-Pinkham, M. A.; Haley, S. M.; Bilodeau, N.; Montpetit, K.; Gorton, G. E.; Mulcahey, M. J.; Tucker, C. A. Using Cognitive Interviewing for Test Items to Assess Physical Function in Children with Cerebral Palsy. *Pediatr. Phys. Ther.* **2008**, *20* (4), 356–362.
- (26) Castillo-Díaz, M.; Padilla, J. L. How Cognitive Interviewing Can Provide Validity Evidence of the Response Processes to Scale Items. *Soc. Indic. Res.* **2013**, *114* (3), 963–975.
- 570 (27) Haladyna, T. M. Developing and Validating Test Items. *Dev. Validating Test Items* **2013**.
- (28) Willis, G. Introduction to Cognitive Interviewing. *Cogn. Interviewing* **2011**, 2–11.
- (29) Fowler, F. J. *Survey Research Methods*; Sage Publications Inc.: Thousand Oaks, CA, 2013.
- (30) Biemer, P. P.; Lyberg, L. E. *Introduction to Survey Quality*; John Wiley & Sons, Inc.: Hoboken, NJ, 2003.
- 575 (31) Zumbo, B. D. Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice. In *The Concept of Validity*; Information Age Publishing, Inc: Charlotte, NC,

2009; pp 65–83.

- 580 (32) Sorana Laneanu, M.; Hubley, A. M. A Model Building Approach to Examining Response Processes as a Source of Validity Evidence for Self-Report Items and Measures. In *Understanding and Investigating Response Processes in Validation Research*; Zumbo, B. D., Hubley, A. M., Eds.; Springer International Publishing: Vancouver, B.C., 2017; pp 115–136.
- (33) Patton, M. Q. Distinguishing Interview Approaches and Types of Interviews. In *Qualitative Research & Evaluation Methods*; 2015; pp 641–648.
- 585 (34) Willis, G. Cognitive Interviewing in Practice: Think-Aloud, Verbal Probing, and Other Techniques. *Cogn. Interviewing* **2011**, 42–65.
- (35) Kreuter, F. Interviewer Effects. In *Encyclopedia of Survey Research Methods*; Lavrakas, P. J., Ed.; Sage Publications Inc.: Thousand Oaks, CA, 2011; pp 370–371.
- (36) Conrad, F. G.; Blair, J. Data Quality in Cognitive Interviews: The Case of Verbal Reports. In *Methods for Testing and Evaluating Survey Questionnaires*; Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Singer, E., Eds.; John Wiley & Sons, Inc., 2004; pp 67–87.
- 590 (37) Stephenson, N. S.; Duffy, E. M.; Day, E. L.; Padilla, K.; Herrington, D. G.; Cooper, M. M.; Carmel, J. H. Development and Validation of Scientific Practices Assessment Tasks for the General Chemistry Laboratory. *J. Chem. Educ.* **2020**, 97 (4), 884–893.
- 595 (38) Barbera, J.; Adams, W. K.; Wieman, C. E.; Perkins, K. K. Modifying and Validating the Colorado Learning Attitudes about Science Survey for Use in Chemistry. *J. Chem. Educ.* **2008**, 85 (10), 1435–1439.
- (39) Hosbein, K. N.; Barbera, J. Development and Evaluation of Novel Science and Chemistry Identity Measures. *Chem. Educ. Res. Pract.* **2020**, 21 (3), 852–877.
- 600 (40) Ferrell, B.; Barbera, J. Analysis of Students' Self-Efficacy, Interest, and Effort Beliefs in General Chemistry. *Chem. Educ. Res. Pract.* **2015**, 16 (2), 318–337.
- (41) Hosbein, K. N.; Alvarez-bell, R.; Callis-duehl, K. L.; Sampson, V.; Wolf, S. F.; Walker, J. P. Development of the Investigation Design, Explanation, and Argument Assessment for General Chemistry I Laboratory. **2020**.
- 605 (42) Mulford, D. R.; Robinson, W. R. An Inventory for Alternate Conceptions among First-Semester

General Chemistry Students. *J. Chem. Educ.* **2002**, *79* (6), 739.

- (43) Stains, M.; Escriu-Sune, M.; De Santizo, M. L. M. A.; Sevian, H. Assessing Secondary and College Students' Implicit Assumptions about the Particulate Nature of Matter: Development and Validation of the Structure and Motion of Matter Survey. *J. Chem. Educ.* **2011**, *88* (10), 1359–1365.
- (44) Beatty, P. C.; Willis, G. B. Research Synthesis: The Practice of Cognitive Interviewing. *Public Opin. Q.* **2007**, *71* (2), 287–311.
- (45) Peterson, C. H.; Peterson, N. A.; Powell, K. G. Cognitive Interviewing for Item Development: Validity Evidence Based on Content and Response Processes. *Meas. Eval. Couns. Dev.* **2017**, *50* (4), 217–223.
- (46) Cresswell, J. W. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*; Pearson: Boston, MA, 2012.
- (47) Blair, J.; Conrad, F. G. Sample Size for Cognitive Interview Pretesting. *Public Opin. Q.* **2011**, *75* (4), 636–658.
- (48) Abell, T. N.; Bretz, S. L. Development of the Enthalpy and Entropy in Dissolution and Precipitation Inventory. *J. Chem. Educ.* **2019**.
- (49) Patton, M. Q. Sample Size for Qualitative Designs. In *Qualitative Research & Evaluation Methods*; Sage Publications Inc., 2015; pp 470–477.
- (50) Willis, G. Analyzing and Documenting Cognitive Interview Results. *Cogn. Interviewing* **2011**, 151–175.
- (51) Hensen, C.; Barbera, J. Assessing Affective Differences between a Virtual General Chemistry Experiment and a Similar Hands-On Experiment. *J. Chem. Educ.* **2019**, *96* (10), 2097–2108.
- (52) Woolley, M. E.; Bowen, G. L.; Bowen, N. K. The Development and Evaluation of Procedures to Assess Child Self-Report Item Validity. **2006**, 687–700.
- (53) Komperda, R.; Hosbein, K. N.; Phillips, M. M.; Barbera, J. Investigation of Evidence for the Internal Structure of a Modified Science Motivation Questionnaire II (mSMQ II): A Failed Attempt to Improve Instrument Functioning across Course, Subject, and Wording Variants. *Chem. Educ. Res. Pract.* **2020**, *21* (3), 893–907.
- (54) DeMaio, T. J.; Landreth, A. Do Different Cognitive Interview Techniques Produce Different

Results? *Methods Test. Eval. Surv. Quest.* **2004**, 89–108.

- (55) Smith, K. C.; Alonso, V. Measuring Student Engagement in the Undergraduate General Chemistry Laboratory. *Chem. Educ. Res. Pract.* **2020**, 21 (1), 399–411.
- (56) Trate, J. M.; Teichert, M. A.; Murphy, K. L.; Srinivasan, S.; Luxford, C. J.; Schneider, J. L. Remote Interview Methods in Chemical Education Research. *J. Chem. Educ.* **2020**, 97 (9), 2421–2429.
- 640 (57) Zoom Video Communications Inc. Using audio transcription for cloud recordings <https://support.zoom.us/hc/en-us/articles/115004794983-Using-audio-transcription-for-cloud-recordings-> (accessed Jan 3, 2021).
- (58) Drawboard Pty Ltd. Drawboard <https://www.drawboard.com/> (accessed Jun 30, 2021).
- 645 (59) Ginger Labs. Notability <https://www.gingerlabs.com/> (accessed Jun 30, 2021).
- (60) RealtimeBoard Inc. Miro <https://miro.com/about/> (accessed Jun 30, 2021).
- (61) Tactivos Inc. Mural <https://www.mural.co/> (accessed Jun 30, 2021).
- (62) Cooper, M. M.; Sandi-Urena, S.; Stevens, R. Reliable Multi Method Assessment of Metacognition Use in Chemistry Problem Solving. *Chem. Educ. Res. Pract.* **2008**, 9 (1), 18–24.
- 650 (63) Lacosta-Gabari, I.; Fernández-Manzanal, R.; Sánchez-González, D. Designing, Testing, and Validating an Attitudinal Survey on an Environmental Topic: A Groundwater Pollution Survey Instrument for Secondary School Students. *J. Chem. Educ.* **2009**, 86 (9), 1099–1103.
- (64) Russell, L. B.; Hubley, A. M. Some Thoughts on Gathering Response Processes Validity Evidence in the Context of Online Measurement and the Digital Revolution. In *Understanding and Investigating Response Processes in Validation Research*; Zumbo, B. D., Hubley, A. M., Eds.; Springer International Publishing: Vancouver, B.C., 2017; pp 229–249.
- 655 (65) Jonas, K. G.; Markon, K. E. Modeling Response Style Using Vignettes and Person-Specific Item Response Theory. *Applied Psychological Measurement*. 2019, pp 3–17.
- (66) Stecher, B.; Le, V. N.; Hamilton, L.; Ryan, G.; Robyn, A.; Lockwood, J. R. Using Structured Classroom Vignettes to Measure Instructional Practices in Mathematics. *Educ. Eval. Policy Anal.* **2006**, 28 (2), 101–130.
- 660 (67) Brovelli, D.; Bölsterli, K.; Rehm, M.; Wilhelm, M. Using Vignette Testing to Measure Student Science Teachers' Professional Competencies. *Am. J. Educ. Res.* **2014**, 2 (7), 555–558.
- (68) Galloway, K. R.; Leung, M. W.; Flynn, A. B. Patterns of Reactions: A Card Sort Task to Investigate

- 665 Students' Organization of Organic Chemistry Reactions. *Chem. Educ. Res. Pract.* **2019**, *20* (1), 30–52.
- (69) Snijkers, G.; Haraldsen, G.; Jones, J.; Willimack, D. *Designing and Conducting Business Surveys*; Wiley: Hoboken, NJ, 2013.
- (70) Snijkers, G. Cognitive Laboratory Experiences: On Pre-Testing Computerised Questionnaires and
670 Data Quality, Utrecht University, 2002, Vol. PH.D. Thes.
- (71) Jobe, J. B. Cognitive Psychology and Self-Reports: Models and Methods. *Qual. Life Res.* **2003**, *12* (3), 219–227.
- (72) Neuert, C. E.; Lenzner, T. A Comparison of Two Cognitive Pretesting Techniques Supported by Eye Tracking. *Soc. Sci. Comput. Rev.* **2016**, *34* (5), 582–596.
- 675 (73) Neuert, C. E. How Effective Are Eye-Tracking Data in Identifying Problematic Questions? *Soc. Sci. Comput. Rev.* **2020**, *38* (6), 793–802.
- (74) Lenzner, T.; Kaczmirek, L.; Galesic, M. Seeing through the Eyes of the Respondent: An Eye-Tracking Study on Survey Question Comprehension. *Int. J. Public Opin. Res.* **2011**, *23* (3), 361–373.
- 680 (75) Menold, N.; Kaczmirek, L.; Lenzner, T.; Neusar, A. How Do Respondents Attend to Verbal Labels in Rating Scales? *Field methods* **2014**, *26* (1), 21–39.
- (76) Elling, S.; Lentz, L.; de Jong, M. Users' Abilities to Review Web Site Pages. *J. Bus. Tech. Commun.* **2012**, *26* (2), 171–201.
- (77) Bonner, S. M. Validity in Classroom Assessments: Purpose, Properties, and Principles. In *SAGE Handbook of Research on Classroom Assessment*; McMillan, J. H., Ed.; Sage Publications Inc.: Thousand Oaks, CA, 2013; pp 87–106.
- 685 (78) Yale Poorvu Centre for Teaching and Learning. Considering Validity in Assessment Design <https://poorvucenter.yale.edu/AssessmentDesignValidity> (accessed Jul 5, 2021).
- (79) Talbot, R. M. On the Importance of the Validity of Classroom Assessments
690 https://medium.com/@Bud_T/on-the-importance-of-the-validity-of-classroom-assessments-fe019f3bdd8c (accessed Jul 5, 2021).