JOURNAL OF CHEMICAL EDUCATION

Article

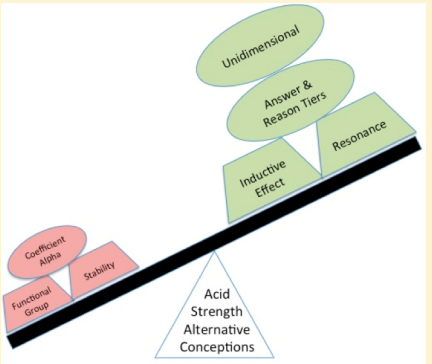
pubs.acs.org/jchemeduc

Students' Understandings of Acid Strength: How Meaningful Is Reliability When Measuring Alternative Conceptions?

Stacey Lowery Bretz* and LaKeisha McClary

Department of Chemistry & Biochemistry, Miami University, Oxford, Ohio 45056, United States

ABSTRACT: Most organic chemistry reactions occur by a mechanism that includes acid−base chemistry, so it is important that students develop and learn to use correct conceptions of acids and acid strength. Recent studies have described undergraduate organic chemistry students' cognitive resources related to the Brønsted−Lowry acid model and the Lewis acid model, providing both qualitative and quantitative analyses of these understandings. To drive changes in pedagogy and curriculum, however, faculty need to be able to quickly assess students' conceptions of acids and acid strength. We recently reported on the development and assessment of a nine-item, multiple-tier, multiple-choice concept inventory about acid strength, named ACID I. Coefficient α for ACID I was calculated to be below 0.70. In this manuscript, we demonstrate that despite this low coefficient α, the data generated by ACID I are indeed reliable. Thus, the purpose of this paper is to (i) report two significant alternative conceptions about acid strength that persist in organic chemistry students' minds after nearly two semesters, and (ii) discuss the meaning of reliability for concept inventories, including a description of additional measures for the reliability of data collected using ACID I. Two types of test conditions were employed within second-semester organic chemistry courses in two different regions of the United States: a course at a medium-sized, midwestern liberal arts university and a large, southeastern research university.



KEYWORDS: *Second-Year Undergraduate, Chemical Education Research, Organic Chemistry, Misconceptions/Discrepant Events, Testing/Assessment, Acids/Bases*

FEATURE: Chemical Education Research

INTRODUCTION

Acid−base chemistry is fundamental to many chemical and biochemical processes. Yet, studies have shown students from secondary school through graduate school have difficulties understanding and applying acid−base principles/theories.[1−5] Learning acid−base chemistry can be particularly challenging for students due to the use of multiple models of acids (i.e., Arrhenius, Brønsted−Lowry, and Lewis), often with little explication of their differences and limitations.[6] Many alternative conceptions that students and preservice teachers hold related to acid−base chemistry have been reported.[7−10] Despite these reports, preservice teachers[11] and college instructors[12] often remain unaware of alternative conceptions their students (are likely to) hold about acid−base chemistry.

Teacher knowledge of chemistry students' alternative conceptions is necessary to develop instructional strategies that promote constructivist learning because such conceptions are integrated into students' knowledge structures and affect students' thinking and reasoning.[13−15] Knowledge structures that students use to make predictions, generate explanations, draw conclusions, and develop hypotheses are called mental models.[16−18] Mental models are dynamic, internal representations that an individual constructs of his or her world.[19] Students often generate mental models that are functional, though incomplete.[20] Mental models can be unique to a situation, and also to the individual.[17] With experience, mental models may become coherent, though not necessarily free of incorrect ideas, and more resistant to change.[21,22] Novice chemistry students hold mental models for many concepts, including bonding,[23−26] chemical equilibrium,[27] and acids and bases.[28,29] Most of these reports sampled high school or general chemistry students. Few studies, however, have reported mental models that advanced chemistry students employ to solve domain-specific problems.[1,30,31]

Advanced chemistry students have been shown to invoke mental models to generate solutions for qualitative tasks. Bhattacharyya[1] described a single mental model that graduate organic chemistry students employed to explain trends in acid strength. McClary and Talanquer[31] showed that undergraduate organic chemistry students used one of four mental models to predict, explain, and justify trends in acid strength; these expressed mental models were hybrids of intuitive beliefs and scientific models. Such synthetic models have also been reported in children[32] and high school students.[29] The McClary and Talanquer study[31,33] involving undergraduate organic chemistry students was the basis for developing ACID I,[34] a nine-item concept inventory to elicit students' conceptions of acid strength.

ACID I is a research tool that can be administered as a diagnostic instrument in classrooms either as a formative

ACS Publications © XXXX American Chemical Society and Division of Chemical Education, Inc.

dx.doi.org/10.1021/ed5005195 | J. Chem. Educ. XXXX, XXX, XXX−XXX

assessment or to evaluate instructional interventions. Other instruments that measure alternative conceptions or intuitive assumptions related to chemistry concepts include particulate nature of matter,[35,36] covalent and ionic bonding representations,[37] kinetic particle theory,[38] solution chemistry,[39] acids and/or bases,[9,29] oxidation−reduction reactions,[40] chemical equilibrium,[41,42] electrolysis,[43] and enzyme−substrate interactions[44]

ACID I was designed to investigate undergraduate organic students' mental models, based upon three prediction tasks from McClary and Talanquer[31] that required students to conceptually understand the factors that affect acid strength. Some first-semester organic chemistry (OC1) students relied almost exclusively upon explicit features of structural representations of organic compounds when asked to rank, explain, and justify trends in acid strength.[33] When ACID I was administered to students in the second week of a second-semester organic chemistry course (OC2), two alternative conceptions were measured and described.[34]

METHODS

Goals and Research Questions

This study extends previous research on organic chemistry students' understandings of acid strength and uses ACID I to measure the alternative conceptions that OC2 students held about acid strength at the end of the semester, the prevalence of these alternative conceptions, and how strongly the alternative conceptions were held in students' minds. In addition, this study investigated the reliability of data collected with ACID I, a concept inventory that used students' own words as response options for the multiple choice items.[31,33]

Three research questions framed this research study:

- Do alternative conceptions as measured after OC1 diminish with additional instruction during OC2?
- What alternative conceptions related to acid strength were identified among OC2 students?
- How appropriate is coefficient α for ACID I, specifically, and for concept inventories in general?

Instrument Design

Creating the prediction tasks, deciding upon the format of each item, and choosing response options for ACID I have been described in detail.[34] In summary, three sets of items were created, each framed around a different trio of compounds that had previously proven challenging for students in organic chemistry to understand with regard to acid strength (Figure 1).[31,33] The first item in each set asked students to choose a reason to best explain why one compound was most acidic. The second and third items asked students to indicate the trend in acid strength for the two remaining compounds and to select a reason for their answer. They were also asked to indicate their confidence for each answer and each reason.[45] The original confidence scale published by Caleon and Subramaniam[45] was converted from a Likert scale to an interval scale, i.e., 0% (just guessing) to 100% (absolutely confident), in order to better quantify the confidence of respondents (Figure 2). While answer tiers inform educators about what students know, reason tiers provide insights into why students have particular conceptions.[46] Confidence tiers offer an indication of how strongly conceptions are held in students' minds.[40,45]

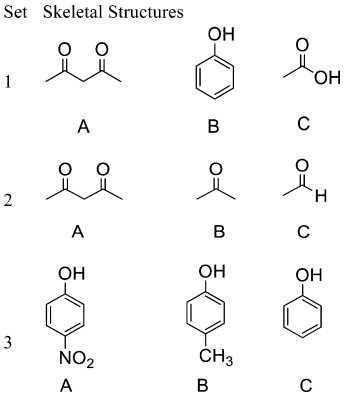


Figure 1. Three sets of structures used in ACID I to elicit students' conceptions of acid strength.

How confident are you about your response?

Place an X anywhere on the scale.

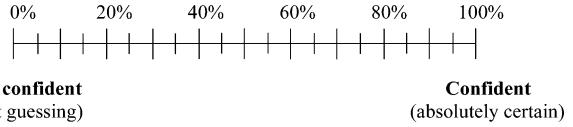


Figure 2. Format of confidence rating for items on ACID I.

Validity

Both the face validity and content validity of ACID I were previously established with content experts.[34] The 2D, skeletal structures of compounds were consistent with students' notions of representations in organic chemistry. Furthermore, the paper-and-pencil, multiple-choice format was a familiar, reasonable approach to quickly gather data about organic chemistry students' understanding of acid strength. Content validity was established by asking organic chemistry instructors who teach undergraduate and graduate organic chemistry courses to determine the extent to which ACID I, framed within the context of predicting and explaining trends in acid strength, reflected the key concepts necessary to understand acid strength within organic chemistry.

Settings, Participants, and Data Collection

ACID I was administered under two different conditions at two institutions; a liberal arts college (LA) and a research university (RU). To answer the research questions, these two institutions were chosen because they taught organic chemistry using different curricula. Two sections (n = 152) of organic chemistry were taught by Professor X at LA using a spiral curriculum: topics were taught broadly first-semester and more in depth second-semester.[47] At RU, Professor Y taught one section and Professor Z taught two sections of OC2 (total n = 226) using a traditional two-semester organic chemistry curriculum where topics were introduced according to the order they appeared in the textbook. Students who were enrolled in OC2 at both institutions completed ACID I during regularly scheduled lecture periods. The three professors signed letters of support as part of the Institutional Review Board process, and student participants consented in writing to be part of this research study.

dx.doi.org/10.1021/ed5005195 | J. Chem. Educ. XXXX, XXX, XXX−XXX

Table 1. Psychometrics for Pre/Post and Post-Test Conditions

Statistic	LA Repeaters (n = 52)		All Students (n = 290)
	Pre-test	Post-test	Post-test
$M_{\text{total score}}$	3.04 ± 1.59	3.33 ± 1.92	3.09 ± 1.70
CF_{student} (%)[a]	54.10 ± 15.67	63.00 ± 12.40	60.76 ± 15.76
Item difficulty[b]	0.13−0.71	0.12−0.65	0.17−0.57
	(M = 0.34 ± 0.20)	(M = 0.37 ± 0.15)	(M = 0.34 ± 0.13)
Item discrimination[b]	0.19−0.80	0.33−1.00	0.40−0.78
	(M = 0.51 ± 0.22)	(M = 0.69 ± 0.22)	(M = 0.55 ± 0.15)
Item reliability[b]	−0.06−0.62	0.25−0.56	0.30−0.54
	(M = 0.37 ± 0.21)	(M = 0.44 ± 0.10)	(M = 0.41 ± 0.08)
Coefficient α	0.39	0.54	0.39

[a]CF_{student} = mean confidence per student. [b]Item difficulties, discriminations, and reliabilities are reported as ranges.

Two test conditions were employed to investigate whether an additional semester of instruction had an effect on students' alternative conceptions. For example, learning about more complex reactions in OC2 where acid strength is an important concept (e.g., enolate chemistry) might have prepared students to consider additional structural features of each trio of acids (e.g., inductive effect) rather than focus upon surface features (e.g., the number of hydrogen atoms).

Under Condition 1 (pre/post condition), ACID I was administered to half the LA students in OC2 during week 2 of the spring semester. (The other students participated in pilot-testing a different concept inventory during this time. The LA students were randomly divided into the two concept inventory groups by their seating in the lecture hall.) Then in week 13, all LA students in OC2 completed ACID I. Thus, some students in OC2 completed ACID I twice (LA Repeaters), while others completed ACID I just once (LA Nonrepeaters). Under Condition 2, OC2 students at RU completed ACID I only once, during week 14. The motivation for this second test condition was to examine the generalizability of the findings from Condition 1.

Data Analysis

Only students who provided an answer and a confidence rating to all questions were included in the analysis. Each item on ACID I was scored as 0 if answered incorrectly, as 1 if answered correctly, and summed to compute a total score for each participant. The minimum total score possible was 0, while the maximum total score possible was 9. Item difficulty, item discrimination, and item reliability (i.e., point biserial coefficients) were computed.[48] Mean confidence was calculated for each item (CF_{item}) and for each student (CF_{student}).[45] Data from a total of 290 students (LA, n = 121; RU, n = 169) were analyzed and described below.

Only distracters that were chosen by at least 35% of students, (i.e., at least 10% above chance for an item with four response options) were considered for further analysis because these conceptions were deemed *significant*.[45] *Significant* specific cases with $CF_{\text{item}} \geq 50\%$ were considered *genuine*, while those cases with $CF_{\text{item}} < 50\%$ were considered *spurious*. *Genuine* cases were considered to be held more strongly in students' minds while *spurious* cases were considered to be less strongly held and transient.

Analyses of data from both test conditions, including the measurement of alternative conceptions, are presented separately, followed by a discussion of the findings and implications for teaching.

RESULTS

Condition 1: Pre/Post

Fifty-two students (i.e., LA Repeaters) completed ACID I in January (pre) and April (post) 2011. Summary statistics are provided in Table 1. The maximum possible score for ACID I is 9; therefore, the mean total scores suggest this was a difficult concept inventory for all students. Values for item difficulty, which is a measure of how many (what percent) of students answered a question correctly, range from 13% to 71%, with means of 34−37%, again indicating the students found the questions on ACID I to be difficult. The α values are low, raising a question about the internal consistency of the students' responses. (Challenges with interpreting α values, specifically for a concept inventory, are discussed later in this manuscript.)

Because each student in the LA Repeater sample had two total scores, normalized gains[49] were computed in order to investigate whether an additional semester of instruction had a positive effect on students' understanding. While 85% of students had a non-zero normalized gain, slightly more students had a positive normalized gain (n = 23) than a negative normalized gain (n = 21). The average normalized gain for the latter group was −0.44, whereas the average normalized gain for the former group was +0.36. Therefore, the average normalized gain for LA Repeaters was effectively zero.

Although data obtained from the pre/post condition were repeated measures, it was not appropriate to calculate a Pearson coefficient. Students had three months additional instruction (including concepts about acid strength) between the pre/post administrations; therefore, the reliability as determined by a test-retest condition would not be as meaningful.[50] Instead, the coefficient α was calculated for each administration of ACID I. Coefficient α was 0.39 on the pre-test and 0.54 on the post-test.

As previously reported, ACID I measures two alternative conceptions: *functional group determines acid strength* and *stability determines acid strength*.[34] Students who reason from the *functional group* misconception focus upon specific structure/composition features to explain trends in acid strength rather than implicit, electronic factors such as polarizability, inductive effect, and resonance. (For example, students who reason from this misconception decide that acetic acid and phenol are both more acidic than pentane-2,4-dione (see Figure 1, set 1) because the latter does not contain OH but the other two compounds do, and OH is associated with the carboxylic acid functional group −COOH.) On each of items 1, 6, and 7, the *functional group* misconception was selected by more than 35% of students whose mean confidences ranged

dx.doi.org/10.1021/ed5005195 | J. Chem. Educ. XXXX, XXX, XXX−XXX

from 45.34−64.65% on the pre-test to 57.76−75.66% on the post-test (Table 2). Students who reason from the stability

Table 2. Frequencies of the Most Common Distracter and Mean Confidence per Item in the Pre/Post Condition

	Pre-test (January, n = 52)		Post-test (April, n = 52)	
Item	Frequency of distracter (%)	CF_{item} (%)	Frequency of distracter (%)	CF_{item} (%)
1[a]	67.31	55.91 ± 23.62	61.54	73.55 ± 17.41
2[a]	84.62	63.28 ± 23.06	71.15	50.96 ± 22.10
3[a]	63.46	81.32 ± 17.81	61.54	68.63 ± 21.98
6	38.46	45.34 ± 20.94	46.15	57.76 ± 21.88
7[a]	48.08	64.65 ± 21.68	57.69	75.66 ± 18.09

[a]Five items measured one of two alternative conceptions: *functional group determines acid strength* or *stability determines acid strength*. With the exception of Item 6 on the pre-test, all items measured a *genuine* significant case. The most frequent distracters are provided in Table 3.

determines acid strength misconception reason that molecules which are more stable are less likely to react, i.e., they are weaker acids. On both items 2 and 3, the *stability* misconception was used by more than 35% of students whose mean confidences ranged from 50.96−68.63% on the pre-test to 63.28−81.32% on the post-test (Table 2).

Condition 2: Post Only

In April, 290 students completed ACID I (Table 1). Total scores were not normally distributed, so a Kruskal−Wallis nonparametric test was used to determine that there were no significant differences among any of the sections (LA Repeaters, LA Nonrepeaters, 3 RU sections). Therefore, data for all students was combined for the purposes of further statistical analyses.

DISCUSSION AND IMPLICATIONS FOR TEACHING

While chemistry experts rely on structure/composition features of substances to predict thermodynamic and kinetic behavior within chemical systems, novices indiscriminately rely on structure/composition to make predictions.[51,52] Therefore, *functional group determines acid strength* is considered an alternative conception. With regard to *stability determines acid strength*, many chemistry students have difficulties understanding thermodynamic stability in the context of chemical bonding[49] and acid strength.[31] ACID I was designed to measure when organic chemistry students consider conjugate base stability as the best reason to explain trends in acid strength, but not to fully investigate students' understandings of thermodynamic stability. It is plausible, of course, a subset of participants may try to reason from both alternative conceptions as measured by ACID I.

Within the data summarized by question in Table 2, 4 particular distracters were chosen by more than 35% of participants, independent of test condition (Table 3). Notably, the confidence regarding these distracters elicited genuine conceptions ($CF_{item} \geq 50\%$) rather than spurious ones ($CF_{item} < 50\%$), suggesting that these items reliably detect both the *functional group* and the *stability* misconceptions in students who have completed either one or two semesters of organic chemistry. The fact that an additional semester of instruction had no noticeable impact upon student thinking about acidity has considerable implications for the teaching and assessment of acids and acid strength in organic chemistry.

Acid strength is an emergent property of chemical substances, i.e., it relies on competing intrinsic and extrinsic factors such as molecular composition, structure, and solvent interactions rather than mere cause-effect relationships. Students' understandings of the influences of molecular composition were measured using ACID I. For example, Set 1 (i.e., Items 1, 2, and 3) explored whether students' conceptions of acid strength mostly relied upon structure/composition or upon electronic factors (inductive effect or resonance) when answering items about the relative acid strength of pentane-2,4-dione, phenol, and acetic acid. The results showed that for *both* the pre/post condition *and* the post condition, most students selected acetic acid as the most acidic substance in the set because "Compound C is a carboxylic acid". Similarly with Item 7, the most frequent response to best explain why *p*-nitrophenol was more acidic than both *p*-methylphenol and phenol was "Compound A has an NO_2 group". As with earlier reports, organic chemistry students, whether during their first-semester[31,33] or at the start of their second semester,[34] continued during the second semester to rely upon prior knowledge related to structure/composition of molecules rather than implicit, electronic factors to make decisions about acid strength.

Structure and composition are emphasized in general chemistry because these factors greatly influence both the physical and chemical properties of substances. While it is problematic that even after two semesters of organic chemistry that students continue to rely predominantly upon these surface features, instruction can potentially shift students' attention toward more conceptual, intrinsic reasoning.[53] Even though students can provide both Brønsted−Lowry and Lewis definitions of acids, few can apply their knowledge to make decisions about acid strength,[2,33] to explain observed trends in acid strength[1,31] or to propose reasonable reaction mechanisms.[54,55] Distinguishing whether a reactant will behave as an

Table 3. Specific Cases of Two Alternative Conceptions Identified Using ACID I under Two Experimental Conditions

Item	Specific Cases of Alternative Conceptions (AC)	AC Strength	AC Frequency, %
1	Acetic acid is more acidic than both phenol and pentane-2,4-dione because it is a carboxylic acid. *Functional group determines acid strength*[a]	genuine	53.8
2	Phenol is more acidic than pentane-2,4-dione. *Functional group determines acid strength*[a]	genuine	67.6
3	Phenol is more acidic than pentane-2,4-dione because the benzene better stabilizes the conjugate base than the carbonyl groups of pentane-2,4-dione. *Stability determines acid strength*[a]	genuine	54.1
7	*p*-Nitrophenol is more acidic than *p*-methylphenol and phenol because *p*-nitrophenol has a nitro ($-NO_2$) group. *Functional group determines acid strength*[a]	genuine	44.5

[a]Alternative conceptions are written in italics. Data are from the post condition (n = 290). Students primarily relied on structure/composition properties of acids to predict and explain trends in acid strength.

dx.doi.org/10.1021/ed5005195 | J. Chem. Educ. XXXX, XXX, XXX−XXX

acid/base or as an electrophile/nucleophile is also particularly challenging for students.[2,55]

It bears noting that the findings in this study were independent of whether the curriculum was the traditional "march through functional groups" or the novel spiral curriculum.[47] Organic chemistry instructors should facilitate students' shifts toward more conceptual reasoning by routinely conducting formative assessment of their students' understanding of acid−base theories as new mechanisms are taught. Requiring students to *explicitly identify* which species are functioning as acids, bases, electrophiles, or nucleophiles *and explain why* each species is best described as belonging to one of those categories is recommended. ACID I is a suitable tool for formative assessment of students' understandings of acid strength, particularly as students may be relying on alternative conceptions to draw inferences when thinking about acids and acid strength. Results from ACID I can then be used to inform subsequent teaching and assessment practices within a course.

Using diagnostic tools to routinely assess students' understandings can also assist instructors in identifying prerequisite concepts and principles from general chemistry that are critical to success in organic chemistry. In a survey of 23 college organic chemistry educators in the U.S. about such prerequisite concepts and principles, Duis[12] found that acid−base chemistry was cited by many organic chemistry faculty as a fundamental organic chemistry concept, yet one from general chemistry that required review because it was difficult for students. Informal surveys similar to the one Duis conducted should occur within departments to maximize continuity of learning throughout the chemistry curriculum. Furthermore, discussions with faculty in other science disciplines are recommended to help students appreciate and understand how a chemistry concept such as acid strength plays a role in biological systems, or, e.g., how physics is involved in measuring dissociation constants. Finally, educators must help students make meaningful connections and place value on those connections by continually, and effectively, assessing them.[56,57] In particular, students' higher-order cognitive skills should be developed and routinely assessed.[58] Smith, Nakhleh, and Bretz[59] have described a framework to assist practitioners in designing exams to assess students' lower-order and high-order cognitive skills.

IMPLICATIONS FOR CHEMISTRY EDUCATION RESEARCH

When designing assessments to measure students' learning, an important psychometric to consider is reliability, which measures how consistent *data* are (rather than the instrument itself).[40,60,61] With respect to diagnostic instruments, both instructors and researchers must be assured to a reasonable extent that the data accurately reflect students' understandings of the targeted concept(s). Coefficient α is a useful measure of internal consistency when instruments cannot be administered multiple times; in this way, reliability can also be thought of as minimizing random error.[62,63] For ACID I, the coefficient α of data collected in January with 89 students was 0.41.[34] In this study, considering only the LA Repeaters, coefficient α was 0.39 in January, which was similar to the larger January sample and the larger April sample, and yet it was 0.54 in April (Table 1). The substantial increase in internal consistency raises the question: *what constitutes meaningful measurement of reliability for diagnostic tools?*

Certainly classical test theory is used widely throughout the chemistry education research community, and it has influenced how researchers think about assessment design and the psychometric analyses of assessments (see refs 34−36, 39, 40, 44, 46, 56, 57, 60, 61, 64−66). Diagnostic instruments in science education literature typically report coefficient α to demonstrate internal consistency among students' responses on the instruments, although other several quantitative measures for reliability exist.[67] Conceptually, reliability is the degree to which an instrument repeatedly produces similar or the same results. An important assumption of coefficient α is normality, i.e., total scores must be normally distributed. However, students with identical scores do not necessarily hold identical misconceptions. Therefore, even if students' total scores on a diagnostic tool designed to measure alternative conceptions were normally distributed, it is unlikely that the alternative conceptions themselves would be normally distributed. Considered through the lens of constructivism as a model for how learning happens, students' understandings depend upon their prior knowledge and experiences, and such knowledge is unlikely to be consistently recalled or applied to answer items on cognitive assessments. As students engage in learning, they try to connect what they already know to what they need to know, and in doing so, many things can happen: concepts can be integrated correctly into prior knowledge; OR concepts can be connected but the nature of the relationship between them is incorrect; OR concepts can be isolated from those they ought to be connected to, resulting in fragmented knowledge; OR concepts that ought to be learned never are, resulting in gaps. This combination of errors, fragments, and gaps will undoubtedly result in less than consistent responses by students on diagnostic assessments. Therefore, the traditional threshold of 0.7 as indicative of acceptable reliability is actually a flawed metric when it comes to diagnostic assessments. Coefficient α cannot capture these complexities in students' understandings, nor was it designed to do so.

Streiner[68] has described four inappropriate uses of coefficient α: (1) tests that measure how many items are completed in a fixed period of time, (2) tests where items are presented in order of difficulty, (3) when the answer to one item depends on the answer to a previous item, and (4) tests with more than one dimension. ACID I is a diagnostic tool that includes items with separate answer tiers and reason tiers; it also does not measure a unidimensional construct. Again, another important caution when using the coefficient α: high internal consistency does not imply unidimensionality.[69] Similarly, measuring students' understandings of a single concept does not imply that the test itself was unidimensional. In fact, we argue that under most circumstances, a concept inventory will fail this important assumption for the coefficient α because by nature, chemistry as a discipline relates interconnected ideas. If the diagnostic tool does not assess understanding of the many related concepts (i.e., it has low construct validity), then inferences drawn from data may be inaccurate.[68,69] While ACID I focuses on the concept of acid strength, the construct (i.e., students' understandings of acid strength) also depends on students' understandings of concepts such as chemical equilibria and the conventions used to represent structures. Thus, coefficient α is not an appropriate measure of reliability for ACID I, or for that matter, most diagnostic instruments in chemistry (science) education.

Recently, chemistry (science) education researchers have become critical of adopted practices that are commonly employed and reported in chemistry (science) education literature.[64,65,70] For example, Adams and Wieman[70] argued

dx.doi.org/10.1021/ed5005195 | J. Chem. Educ. XXXX, XXX, XXX−XXX

that for instruments designed to measure students' understandings (such as those intended to measure alternative conceptions), high internal consistency (e.g., coefficient α) is likely indicative of redundant items, and for that reason, test−retest reliability is a better measure than internal consistency. However, repeated measures data from test-retest conditions may be less reliable if the students are familiar with the instrument or if the topic was explicitly taught between the test and the retest.

ACID I was designed to measure students' understandings of acids and acid strength. Acid−base chemistry is a concept that is typically taught multiple times throughout a second semester organic chemistry course. Therefore, a test−retest condition was not possible in this study. Instead, we used the post condition to examine reliability. Reliable data were expected to show that students from different universities could respond to ACID I at similar points in the semester and have similar responses for each item, for total scores, and that students held a number of the same alternative conceptions related to acid strength.

Recently, Cooper, Underwood, and Hilley[66] used chi-square analysis to determine if responses from two similar groups of students were different on the implicit instruction of Lewis structures instrument (IILSI). We were unable to statistically determine, however, whether students' responses on ACID I differed based on university because the data did not meet the assumptions for a chi-squared test of independence. However, an independent samples Kruskal−Wallis test was performed, and it showed that the distributions of the total scores among the students were not significantly different. As mentioned earlier, Items 1, 2, 3, and 7 elicited the same significant specific case of either *functional group* or *stability* across both samples. On the basis of these findings, the data from ACID I are considered to be reliable.

CONCLUSIONS AND FUTURE RESEARCH

After two semesters of organic chemistry, and two semesters of general chemistry, in most cases, students who completed ACID I held two alternative conceptions, namely, *functional group determines acid strength* and *stability determines acid strength*. Items 1, 2, and 3, in particular, strongly elicited these conceptions from a majority of participants, both in January and again in April. Specific distracters in the January data that were spurious cases of an alternative conception were chosen by fewer than 35% of participants in April; only genuine cases persisted after further instruction in organic chemistry. That Items 1, 2, 3, 7 (and 6 in the pre/post condition only) elicited *functional group* or *stability* with even greater mean confidences suggests that many students finish introductory organic chemistry without gaining a conceptual understanding of acid strength (Tables 1 and 3).

Additional studies are warranted to further explore students' understanding of organic acidity and to address several limitations of the study reported herein. First, the representations themselves could be modified to explicitly include all hydrogen atoms. Note that in the current study, the structure of ethanal as drawn includes only the aldehydic hydrogen atom, but not the three (more acidic) α hydrogens. Under what circumstances do students attend to just explicitly drawn hydrogen atoms vs those not explicitly drawn? Similar studies could be conducted with Kekule structures including all hydrogen atoms or with space-filling models. Second, students were not provided any information regarding the K_a for each compound. Therefore, the opportunity for them to consider relative acidities of the acid vs the conjugate acid, or to consider the implicit electronic features of the conjugate base, remains unexplored. Third, an additional important limitation of the research presented here is the absence of any discussion of solvent effects upon entropy and, ultimately, free energy and acidity.[71,72]

While some specific cases persisted and ostensibly strengthened in students' minds during OC2 (e.g., acetic acid is most acidic because it is a carboxylic acid [Items 1, 2, and 3]), other specific cases were chosen by less than 35% of the students (e.g., pentane-2,4-dione is most acidic because pentane-2,4-dione has two carbonyl groups [Items 4, 5, and 6]). Scenarios like these illustrate that students' alternative conceptions are challenging to assess with high internal consistency because students' knowledge is often fragmented. Many reliability measures are predicated on the assumption of normally distributed data, yet non-normal distributions of alternative conceptions are likely to be the rule, rather than the exception. Nevertheless, reporting the reliability of data collected from a given instrument is necessary for stakeholders (e.g., educators, researchers, administrators) to draw inferences and make decisions about student learning or perhaps teaching effectiveness. Researchers need to more carefully educate audiences regarding the limitations of oft-used reliability measures such as coefficient α when reporting their research in presentations and publications.

Coefficient α was not appropriate to use to establish the reliability of ACID I data. Some item responses depended on the answer from previous items (i.e., scoring separately an answer tier followed by a reason tier), and students' understandings of acid strength could not be considered a unidimensional construct. Thus, coefficient α for students' responses on ACID I varied depending on the test condition. Four items, however, were able to detect *genuine significant* cases of two alternative conceptions. Further evidence to suggest that the data generated by ACID I were reliable is that these cases were previously investigated with first-semester organic chemistry students at a third university in the southwestern United States.[31,33] Not only did the qualitative data from this prior research allowed us to develop ACID I,[34] but they also they provided rich context to make sense of the quantitative data and interpret the reliability.

The confidence scale used in this research was interval (0−100%), rather than the original Likert scale published by Caleon and Subramaniam.[45] However, the distinctions of spurious (CF_{item} < 50%) and genuine (CF_{item} >50%) as labels for misconceptions were retained. Clearly, there are limitations to reporting confidence in terms of a percent, and the authors make no claim that the data are accurate to ±1%. The 0−100% interval scale has also recently been used with cluster analysis to document evidence of "over-confidence," i.e., the Dunning-Kruger effect,[73] in a study to explore students' understandings of redox reactions.[74] While this scale has been used twice now with two different samples of students in two different domains of chemistry (organic acidity and redox reactions), additional studies regarding how to measure and quantify confidence are warranted.

Further research is needed to better explore the appropriateness and the meaningfulness of psychometrics adopted into chemistry (science) education research from other disciplines. Doing so will lead to more robust methods, which in turn can lead to more standardized community practices, especially with

dx.doi.org/10.1021/ed5005195 | J. Chem. Educ. XXXX, XXX, XXX−XXX

regard to reporting. With regard to chemistry students' alternative conceptions, research that explores more deeply the possible origins of such alternative conceptions continues to be warranted.[75] Knowing where the gaps in understanding exist and their origins will further improve instructional pedagogy and assessments.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bretzsl@miamioh.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under NSF DRK-12 Grant No. 0733642. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

(1) Bhattacharyya, G. Practitioner Development in Organic Chemistry: How Graduate Students Conceptualize Organic Acids. *Chem. Educ. Res. Pract.* **2006**, *7* (4), 240−247.

(2) Cartrette, D. P.; Mayo, P. M. Students' Understanding of Acids/Bases in Organic Chemistry Contexts. *Chem. Educ. Res. Pract.* **2011**, *12* (1), 29−39.

(3) Cros, D.; Maurin, M.; Amouroux, R.; Chastrette, M.; Leber, J.; Fayol, M. Conceptions of First-Year University Students of the Constituents of Matter and the Notions of Acids and Bases. *Int. J. Sci. Educ.* **1986**, *8* (3), 305−313.

(4) Demerouti, M.; Kousathana, M.; Tsaparlis, G. Acid-Base Equilibria, Part I. Upper Secondary Students' Misconceptions and Difficulties. *Chem. Educ.* **2004**, *9* (2), 122−131.

(5) Nakhleh, M. B. Students' Models of Matter in the Context of Acid-Base Chemistry. *J. Chem. Educ.* **1994**, *71* (6), 495−499.

(6) Carr, M. Model Confusion in Chemistry. *Res. Sci. Educ.* **1984**, *14* (1), 97−103.

(7) Bradley, J. D.; Mosimege, M. D. Misconceptions in Acids and Bases: A comparative Study of Student Teachers with Different Chemistry Backgrounds. *S. Afr. J. Chem.* **1998**, *51*, 137−145.

(8) Kousathana, M.; Demerouti, M.; Tsaparlis, G. Instructional Misconceptions in Acid-Base Equilibria: An Analysis from a History and Philosophy of Science Perspective. *Sci. Educ.* **2005**, *14* (2), 173−193.

(9) Rahayu, S.; Chandrasegaran, A. L.; Treagust, D. F.; Kita, M.; Ibnu, S. Understanding Acid-Base Concepts: Evaluating the Efficacy of a Senior High School Student-Centered Instructional Program in Indonesia. *Int. J. Sci. Math. Educ.* **2011**, *9* (6), 1439−1458.

(10) Ross, B.; Munby, H. Concept Mapping and Misconceptions: A Study of High-School Students' Understandings of Acids and Bases. *Int. J. Sci. Educ.* **1991**, *13* (1), 11−23.

(11) Dreschler, M.; van Driel, J. Teachers' Perceptions of the Teaching of Acids and Bases in Swedish Upper Secondary Schools. *Chem. Educ. Res. Pract.* **2009**, *10* (2), 86−96.

(12) Duis, J. M. Organic Chemistry Educators' Perspectives on Fundamental Concepts and Misconceptions. *J. Chem. Educ.* **2011**, *88* (3), 346−350.

(13) Driver, R.; Erikson, J. Theories-in-Action: Some Theoretical and Empirical Issues in the Study of Students' Conceptual Frameworks in Science. *Stud. Sci. Educ.* **1983**, *10* (1), 37−60.

(14) Gentner, D. Psychology of mental models. In *International Encyclopedia of the Social and Behavioral Sciences*; Smelser, N.J., Bates, P.B., Eds.; Elsevier Science: Amsterdam, 2002; pp 9683−9687.

(15) Talanquer, V. Commonsense Chemistry: A Model for Understanding Students' Alternative Conceptions. *J. Chem. Educ.* **2006**, *83* (5), 811−816.

(16) van Geenan, E. W.; Witteman, C. L. M. How Experts Reason: The Acquisition of Experts' Knowledge Structures. *Knowl. Eng. Rev.* **2006**, *21* (4), 335−344.

(17) Greca, I. M.; Moreira, M. A. Mental Models, Conceptual Models, and Modelling. *Int. J. Sci. Educ.* **2000**, *22* (1), 1−11.

(18) Talanquer, V. Exploring Dominant Types of Explanations Built by General Chemistry Students. *Int. J. Sci. Educ.* **2010**, *32* (18), 2393−2412.

(19) Johnson-Laird, P. *Mental Models: Towards a Cognitive Science of Language, Inference, And Consciousness*; Cambridge University Press/Harvard University Press: Cambridge, MA, 1983.

(20) Norman, D. A. Some Observations on Mental Models. In Gentner, D., Stevens, A. L., Eds.; *Mental Models*; Psychology Press: London, 1983; pp 7−14.

(21) Vosniadou, S. Capturing and Modeling the Process of Conceptual Change. *Learn. Instr.* **1994**, *4* (1), 45−69.

(22) Vosniadou, S. Mental Models in Conceptual Development. In Magnani, L., Nersessian, N., Eds.; *Model-Based Reasoning: Science, Technology, Values*; Kluwer Academic Press: New York, 2002; pp 353−368.

(23) Coll, R. K.; Treagust, D. F. Learners' Mental Models of Chemical Bonding. *Res. Sci. Educ.* **2001**, *31* (3), 357−382.

(24) Coll, R. K.; Treagust, D. F. Investigation of Secondary School, Undergraduate, and Graduate Learners' Mental Models of Ionic Bonding. *J. Res. Sci. Teach.* **2003**, *40* (5), 464−486.

(25) Taber, K. S. Mediating Mental Models of Metals: Acknowledging the Priority of the Learner's Prior Learning. *Sci. Educ.* **2003**, *87* (5), 732−758.

(26) Luxford, C. J.; Bretz, S. L. Moving Beyond Definitions: What Student Generated Models Reveal about Their Understanding of Covalent Bonding and Ionic Bonding. *Chem. Educ. Res. Pract.* **2013**, *14*, 214−222.

(27) Chiu, M.-H.; Chou, C.-C.; Liu, C.-J. Dynamic Processes of Conceptual Change: Analysis of Constructing Mental Models of Chemical Equilibrium. *J. Res. Sci. Teach.* **2002**, *39* (8), 688−712.

(28) Artdej, R.; Ratanaroutai, T.; Coll, R. K.; Thongpanchang, T. Thai Grade 11 Students' Alternative Conceptions for Acid-Base Chemistry. *Res. Sci. Technol. Educ.* **2010**, *28* (2), 167−183.

(29) Lin, J.-W.; Chiu, M.-H. Exploring the Characteristics and Diverse Sources of Students' Mental Models of Acids and Bases. *Int. J. Sci. Educ.* **2007**, *29* (6), 771−803.

(30) Kraft, A.; Strickland, A. M.; Bhattacharyaa, G. Reasonable Reasoning: Multi-Variate Problem Solving in Organic Chemistry. *Chem. Educ. Res. Pract.* **2010**, *11* (4), 281−292.

(31) McClary, L.; Talanquer, V. College Chemistry Students' Mental Models of Acids and Acid Strength. *J. Res. Sci. Teach.* **2011**, *48* (4), 396−413.

(32) Vosniadou, S.; Brewer, W. F. Mental Models of the Day/Night Cycle. *Cognit. Sci.* **1994**, *18* (1), 123−183.

(33) McClary, L.; Talanquer, V. Heuristic Reasoning in Chemistry: Making Decisions about Acid Strenght. *Int. J. Sci. Educ.* **2011**, *33* (10), 1433−1454.

(34) McClary, L.; Bretz, S. L. Development and Assessment of a Diagnostic Tool To Identify Organic Chemistry Students' Alternative Conceptions Related to Acid Strength. *Int. J. Sci. Educ.* **2012**, *34* (5), 2317−2341.

(35) Nyachwaya, J. M.; Mohamed, A.-R.; Roehrig, G. H.; Wood, N. B.; Kern, A. L.; Schneider, J. L. The Development of an Open-Ended Drawing Tool: An Alternative Diagnostic Tools for Assessing Students' Understanding of the Particulate Nature of Matter. *Chem. Educ. Res. Pract.* **2011**, *12* (2), 121−132.

(36) Stains, M.; Escriu-Sune, M.; Alvarez de Santizo, M. L. M.; 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.

(37) Luxford, C. J.; Bretz, S. L. Development of the Bonding Representations Inventory to Identify Student Misconceptions about Covalent and Ionic Bonding Representations. *Chem. Educ.* **2014**, *91* (3), 312−320.
(38) Treagust, D. F.; Chandrasegaran, A. L.; Crowley, J.; Yung, B. H.; Cheong, I. P.-A.; Othman, J. Evaluating Students' Understanding of Kinetic Particle Theory concepts Relating to the State of Matter, Changes of State, and Diffusion: A Cross-National Study. *Int. J. Sci. Math. Educ.* **2010**, *8* (1), 141−164.
(39) Adadan, E.; Savasci, F. An Analysis of 16−17-year-old Students' Understanding of Solution Chemistry Concepts Using a Two-Tier Diagnostic Instrument. *Int. J. Sci. Educ.* **2012**, *34* (4), 513−544.
(40) Brandriet, A. R.; Bretz, S. L. The Development of the Redox Concept Inventory as a Measure of Students' Symbolic and Particulate Redox Understandings and Confidence. *J. Chem. Educ.* **2014**, *91* (8), 1132−1144.
(41) Özmen, H. Determination of Students' Alternative Conceptions about Chemical Equilibrium: A Review of Research and the Case of Turkey. *Chem. Educ. Res. Pract.* **2008**, *9* (3), 225−233.
(42) Voska, K. W.; Heikkinen, H. W. Identification and Analysis of Student Conceptions Used To Solve Chemical Equilibrium Problems. *J. Res. Sci. Teach.* **2000**, *37* (2), 160−176.
(43) Sia, D. T.; Treagust, D. F.; Chandrasegaran, A. L. High School Students' Proficiency and Confidence Levels in Displaying their Understanding of Basic Electrolysis Concepts. *Int. J. Sci. Math. Educ.* **2012**, *10* (6), 1325−1345.
(44) Bretz, S. L.; Linenberger, K. J. Development of the Enzyme-Substrate Interactions Concept Inventory. *Biochem. Mol. Biol. Educ.* **2012**, *40* (4), 229−233.
(45) Caleon, I.; Subramaniam, R. Do Students Know What They Know and What They Don't Know? Using a Four-Tier Diagnostic Test to Assess the Nature of Students' Alternative Conceptions. *Res. Sci. Educ.* **2010**, *40* (3), 313−337.
(46) Treagust, D. F. Development and Use of Diagnostic Tests to Evaluate Students' Misconceptions in Science. *Int. J. Sci. Educ.* **1988**, *10* (2), 159−169.
(47) Grove, N. P.; Hershberger, J. W.; Bretz, S. L. Impact of a Spiral Curriculum on Student Attrition and Learning. *Chem. Educ. Res. Pract.* **2008**, *9* (2), 157−162.
(48) Ding, L.; Beichner, R. Approaches to Data Analysis of Multiple-Choice Questions. *Phys. Rev. Spec. Top.−Phys. Educ. Res.* **2009**, *5*, 020103−1−020103−17.
(49) Hake, R. R. Interactive-Engagement versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses. *Am. J. Phys.* **1998**, *66* (1), 64−74.
(50) Peter, J. P. Reliability: A Review of Psychometric Basics and Recent Marketing Practices. *J. Mark. Res.* **1979**, *16* (1), 6−17.
(51) Maeyer, J.; Talanquer, V. The Role of Intuitive Heuristics in Students' Thinking: Ranking Chemical Substances. *Sci. Educ.* **2010**, *94*, 963−984.
(52) Taber, K. S. College Students' Conceptions of Chemical Stability: The Widespread Adoption of a Heuristic Rule out of Context and Beyond its Range of Application. *Int. J. Sci. Educ.* **2009**, *31*, 1333−1358.
(53) Chi, M. T. H.; Feltovich, P. J.; Glaser, R. Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Sci.* **1981**, *5* (2), 121−152.
(54) Bhattacharyya, G.; Bodner, G. M. 'It Gets Me to the Product:' How Students Propose Organic Mechanisms. *J. Chem. Educ.* **2005**, *82* (9), 1402−1407.
(55) Ferguson, R.; Bodner, G. M. Making Sense of the Arrow-Pushing Formalism among Chemistry Majors Enrolled in Organic Chemistry. *Chem. Educ. Res. Pract.* **2008**, *9* (2), 102−113.
(56) Bretz, S. L. Navigating the Landscape of Assessment. *J. Chem. Educ.* **2012**, *89* (6), 689−691.
(57) Holme, T. A. Assessment Data and Decision Making in Teaching. *J. Chem. Educ.* **2011**, *88* (8), 1017−1017.
(58) Zoller, U. Science Education for Global Sustainability: What is Necessary for Teaching, Learning, and Assessment Strategies? *J. Chem. Educ.* **2012**, *89* (3), 297−300.
(59) Smith, K. C.; Nakhleh, M. B.; Bretz, S. L. An Expanded Framework for Analyzing General Chemistry Tests. *Chem. Educ. Res. Pract.* **2010**, *11* (3), 147−153.
(60) 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.
(61) Gerlach, K.; Trate, J.; Blecking, A.; Geissinger, P.; Murphy, K. Valid and Reliable Assessments to Measure Scale Literacy of Students in Introductory College Chemistry Courses. *J. Chem. Educ.* **2014**, DOI: 10.1021/ed400471a.
(62) Cronbach, L. J. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* **1951**, *16* (3), 297−334.
(63) Cronbach, L. J. Test Reliability: Its Meaning and Determination. *Psychometrika* **1947**, *12* (1), 1−16.
(64) Bretz, S. L. Designing Assessment Tools to Measure Students' Conceptual Knowledge of Chemistry. In *Tools of Chemistry Education Research*; Bunce, D., ; Cole, R., Eds.; American Chemical Society: Washington, D.C., 2014; pp 155−168.
(65) Lewis, S. E.; Lewis, J. E. The Same or Not the Same: Equivalence as an Issue in Educational Research. *J. Chem. Educ.* **2005**, *82* (9), 1408−1412.
(66) Cooper, M. M.; Underwood, S. M.; Hilley, C. Z. Development and Validation of the Implicit Information from Lewis Structures Instrument (IILSI): Do Students Connect Structures with Properties? *Chem. Educ. Res. Pract.* **2012**, *13*, 195−200.
(67) Salkind, N. J. Getting It Right Every Time: Reliability and Its Importance. In *Tests & Measurement for People Who (Think They) Hate Tests & Measurement*; SAGE Publications: Thousand Oaks, CA, 2006; pp 37−62.
(68) Streiner, D. L. Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *J. Pers. Assess.* **2003**, *80* (1), 99−103.
(69) Segars, A. Assessing the Unidimensionality of Measurement: A Paradigm and Illustration within the Context of Information Systems Research. *Omega* **1997**, *25* (1), 107−121.
(70) Adams, W. K.; Wieman, C. E. Development and Validation of Instruments to Measure Learning of Expert-like Thinking. *Int. J. Sci. Educ.* **2011**, *33* (9), 1289−1312.
(71) Rochester, C. H.; Rossall, B. Steric Hindrance and Acidity. Part 3. Enthalpies and Entropies of Ionization of Phenols in Methanol. *Trans. Faraday Soc.* **1969**, *65*, 1004−1013.
(72) Parsons, G. H.; Rochester, C. H. Enthalpies and Entropies of Ionization of 4-Substituted Phenols in Methanol + Water Mixtures. *J. Chem. Soc., Faraday Trans. 1* **1975**, *71*, 1069−1082.
(73) Dunning, D. The Dunning-Kruger Effect: On Being Ignorant of One's Own Ignorance. In *Advances in Experimental Social Psychology*; Olson, J.M., Zanna, M.P., Eds.; Elsevier Academic Press: San Diego, CA, 2011; Vol. *44*, pp 247−296.
(74) Brandriet, A. R.; Bretz, S. L. Measuring 'Meta-Ignorance' through the Lens of Confidence: Examining Students' Redox Misconceptions about Oxidation Numbers, Charge, and Electron Transfer. *Chem. Educ. Res. Pract.* **2014**, *15* (4), 729−746.
(75) Hammer, D. Misconceptions or P-Prims: How May Alternative Perspectives of Cognitive Structure Influence Instructional Perceptions and Intentions. *J. Learn. Sci.* **1996**, *5* (2), 97−127.