

Chemistry Education Research and Practice

Accepted Manuscript



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

1
2
3 **On the prevalence of alternative conceptions on acid-base chemistry among secondary**
4 **students: Insights from cognitive and confidence measures**
5
6
7
8
9

10
11
12
13 Kai Yee Hoe & Subramaniam Ramanathan*

14
15 Nanyang Technological Univeristy - National Institute of Education, Singapore
16

17 * email: subramaniam.r@nie.edu.sg
18
19
20
21
22

23 **Abstract**
24

25
26 This study presents an analysis of alternative conceptions (ACs) on acid-base chemistry
27 harbored by grade 9 students in Singapore. The ACs were obtained by the development and
28 validation of a 4-tier diagnostic instrument. It is among the very few studies in the science
29 education literature that have focused on examining results based also on measures used in
30 the educational psychology literature. Results indicate that the students harbor a range of ACs
31 of varying strengths in relation to the properties of acids and bases, strengths of acids and
32 bases, pH, neutralization, indicators, and sub-microscopic view of acids and bases. The 25-
33 item instrument uncovered 30 ACs. A novel insight from this study is that when students are
34 presented with a test item with all incorrect answer and reason responses but with a blank
35 space for them to fill up their own answer and reason if they disagree with any of these
36 responses, hardly any chose to do so. Overall, the results reiterate the utility of the 4-tier
37 format for identifying ACs and getting some indication of their strengths.
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Introduction

Learning is a complex process in the process of acquiring knowledge. Not all content taught is learnt. Gaps are bound to exist in students' understanding. The presence of alternative conceptions (ACs) can also interfere with students' understanding. These ACs are at odds with the relevant scientific concepts, and have to be identified and addressed if proper conceptual understanding is to be promoted. Not surprisingly, studies on students' understanding of topics in the sciences have been a fruitful field of research, and these have allowed for the documenting of ACs on various topics for use by instructors and researchers.

The focus of this study is on Chemistry. Some may view chemistry as a subject that is difficult to be effectively taught and learnt. Kavanaugh and Moomaw (1981) found that many students have difficulties in understanding fundamental chemistry concepts. Students' difficulties in learning chemistry have been well documented (for example, Gabel & Bunce, 1994). Contributing factors include abstract nature of the subject (Herron, 1975), remoteness of the language used (Glassman, 1967) and different levels of representations used (Gabel, Samuel & Hunn, 1987; Nakhleh & Kracjik, 1994).

In order to understand chemistry, one must be familiar with the various forms of representations such as symbols, pictorial representations and equations. Very often, visualization at the molecular level is crucial for understanding and explaining chemical phenomena. Kozma and Russell (1997) showed that the schemas for chemistry knowledge differ between experts (teachers) and novices (students) in terms of complexity. Novices are lower in representational competence – fluency in using different representations and switching between representations. They concluded that novices normally use only one form of representation and are not quite able to transform it to other forms. Novices also tend to

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

rely on the obvious surface features while the experts have better grasps of the underlying principles. These difficulties carry on into all areas of chemistry. A number of studies have focused on student difficulties with concepts in acid-base chemistry, the focus of our research. A summary of the literature on studies involving ACs on acid and bases can be found in Table 1. As can be seen, students and teacher harbour a range of misconceptions on this topic.

Gaps in studies reviewed

The literature review explored the range of ACs exhibited by students on the topic of acids and bases, and how these were diagnosed. It can be seen that these studies have been done from elementary, secondary, high school, college to university levels. Most of them are focused on a single level except in the case of the Taiwanese National Survey (Chiu, 2007), which reported very comprehensive and concerted efforts to cover elementary to senior high schools over a six year period of study. It has, however, to be noted that this is a comprehensive survey on various topics in Chemistry, with acid-base chemistry being only a sub-set of the survey.

Except for the study by Demircioğlu et al. (2005), the number of ACs uncovered for upper secondary students is relatively few (five to six). It is not likely that these studies covered sufficient breadth and depth. In addition, the Concept Achievement Test developed by Demircioğlu et al. (2005) is, strictly speaking, an achievement test rather than a diagnostic instrument although it doubled up in its role to identify ACs together with interviews.

Also noticeable is that for most of the studies, other than standard assessments and interviews, 2TMC test is one of the diagnostic instruments used. The inclusion of the reason tier in 2-tier multiple choice (2TMC) questions greatly reduces the probability of getting a question correct

1
2
3
4 by guessing (from 0.25 to 0.0625 for 4-response options in each tier), leading to more robust
5
6 conclusions in the studies. However, it does not help to shed more light on why students get
7
8 an answer wrong. A correct answer accompanied by wrong reasoning could mean either the
9
10 existence of an AC or a lack of knowledge – and it is not possible to differentiate between
11
12 these (Caleon & Subramaniam, 2010). In the recent literature, two diagnostic instruments on
13
14 acids/bases have been reported. The 18 item Acid–Base Diagnostic Test (ABDT) of Artdej et
15
16 al. (2010), based on 2-tier MCQ format, revealed 19 ACs among 55 eleventh grade 11
17
18 students. An obvious limitation of the instrument is that the language used is Thai, and it is
19
20 therefore not easy for others to use it. In addition, since two-tier MCQs were used, chance
21
22 factors and lack of metacognitive information can also contribute to its shortcomings. The
23
24 nine item ACID I, in four-tier format, developed by McClary and Bretz (2012) uncovered
25
26 two ACs on acid strength among 104 undergraduates. With the incorporation of confidence
27
28 ratings, wrong answers and/or reasons with high confidence ratings can provide more reliable
29
30 indications of ACs. The magnitude of the confidence ratings also provides some indication of
31
32 the strengths of the ACs (Sreenivasulu & Subramaniam, 2013). The use of 4 tier multiple
33
34 choice (4TMC) tests as a diagnostic instrument for uncovering ACs is relatively new, with
35
36 Caleon and Subramaniam (2010) having used it earlier to diagnose upper secondary students’
37
38 ACs on the topic of waves.
39
40
41
42
43
44
45
46
47

48
49 The use of 4TMC test as a diagnostic instrument, however, is still largely unexplored – we
50
51 managed to locate only a handful of references on this. In terms of number of questions used
52
53 in the studies, the literature indicates the number of one-tier and two-tier MCQs ranging from
54
55 5 to 25 items. Nine items were used by McClary & Bretz (2012) in their 4TMC test. Further
56
57 studies on 4TMC question in different domains are needed before we can unequivocally
58
59 reiterate its utility. The number of ACs identified in the studies (approximately) that are
60

1
2
3 equivalent to the coverage of the 'O' Level chemistry syllabus, the focus of our study, ranges
4
5 from 1 to about 16.
6
7

8
9
10 The principal objectives of this study are to develop a 4-tier diagnostic instrument on acid-
11
12 base Chemistry and to document the prevalence and strengths of the ACs among grade 9
13
14 students.
15
16

17 18 19 20 **Methodology**

21 22 **Development of acid-base Chemistry diagnostic instrument (ABCDI)**

23 24 *Reserch deisgn*

25
26
27 We used the approach of Treagust (1988) for 2-tier instruments to craft the ABCDI but with
28
29 same modifications, which are elaborated below.
30
31

32 33 34 *Ethicx clearance*

35
36 Approval from the univeristy's Institutionnl Review Board was obtained for the conduct of
37
38 this study. All students who participated in tbis study gave informed consent through their
39
40 parents.
41
42

43 44 45 46 *General information*

47
48 The test instrument was crafted in English, which is also the medium of instruction in schools
49
50 in Singapore. Students were taught the topic of acids and bases by their teachers at least two
51
52 months before this study. They were also given at least a week to revise this topic before
53
54 sitting for the diagnostic test.
55
56

57
58
59
60 Phase 1 (Preliminary Phase)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Development of MCQ instrument with open-ended section for explanation

Based on the literature, conversations with secondary school chemistry teachers over the years, and the first author's teaching experience, a 25 item MCQ instrument was developed. Each MCQ comprises a standard stem plus four or five options. The reason tier was a blank space for students to provide their reason for their choice of answer.

Student interviews were not used as an additional approach in this phase to uncover ACs for the following reasons:

(a) While interviews can provide better insights into students' knowledge framework, it is a very time consuming process.

(b) Interviews have to be recorded and transcribed, and the process is tedious. This study aims to also ascertain the effectiveness of crafting 4TMC items without resorting to interviews.

(c) Interviews can only capture certain aspects of conceptions of a small sample. In place of the additional ACs that could be potentially generated via interviews, the author's teaching experience was instead used to generate a number of ACs to be used as distracters for the MCQ items.

The test was administered to Secondary 3 (equivalent to Grade 9 ages about 15-16) students (N=113) from a mainstream government co-educational school. The students take Pure Chemistry as one of the subjects for the Singapore-Cambridge General Certificate of Education. The time allotted for the test was one hour.

1
2
3 Facility index (FI) and discrimination index (DI) were calculated for each item. Distractor
4 analysis for the MCQ items plus analysis of the open-ended responses were also done. These
5 data were used as a guide to modify or delete questions, and to identify potential ACs for use
6 as distracters in framing the second tier for the pilot version of the instrument. In this way a
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Facility index (FI) and discrimination index (DI) were calculated for each item. Distractor analysis for the MCQ items plus analysis of the open-ended responses were also done. These data were used as a guide to modify or delete questions, and to identify potential ACs for use as distracters in framing the second tier for the pilot version of the instrument. In this way a 24 item two-tier MCQ instrument was developed. For each of the answer and reason tiers, a corresponding confidence scale was added, with the scale ranging from 1 (just guessing) to 6 (absolutely confident).

The instrument was sent for validation to three Chemistry academics and a subject head who has taught secondary level Chemistry for more than ten years. They were requested to assess the items in the instrument holistically as well as individually. A checklist that they need to tick (Yes, No and Remarks columns) was also provided, and it focused on the following:

- The questions can be understood by the level of students being evaluated – this is to ensure that the students would not provide a wrong answer due to weak command of the language.
- The questions test students' understanding of acids and bases – this is to ensure that the questions fall within the scope of the syllabus.
- The test is appropriate for the level being evaluated – this is to avoid dependence on knowledge beyond the intended level in order to answer the items correctly.
- The questions are not ambiguous – this is to ensure that the questions would not be misinterpreted.
- Each question has only one correct response in each tier.
- The time to complete the test is reasonable (one hour)
- Each question asks for an answer on only one dimension (not double barrelled)
- The responses accommodate most of the possible answers. – this is to reduce the

1
2
3 chances of guessing or elimination of the least plausible options to answer the
4 question.
5
6

- 7
- 8 • The questions do not lead the respondent to a desired answer. – this is to avoid a
9 situation where the stem provides hints (e.g. grammatical hints) for the answer.
10
11
 - 12 • The questions do not use emotionally loaded or vaguely defined words. – this serves
13 to keep the questions “emotionally neutral”, neither appealing nor discriminating to
14 the students’ state of mind, thus decreasing any ambiguities.
15
16
17
18
19

20
21
22 Overall, the validators felt that the instrument can be used with some minor modifications.
23 Changes to some of the options and rephrasing of a few questions were undertaken to address
24 their concerns. In this way, a pilot version of the instrument was developed.
25
26
27
28
29

30 31 **Pilot study**

32
33 The pilot test instrument was administered to another sample of students (N=116) from
34 secondary 3 in three mainstream government co-educational schools who were studying pure
35 Chemistry. It was found that the one hour duration was more than adequate to complete the
36 test.
37
38
39
40
41
42

43
44
45 Data from this phase were collated and analyzed. Again, FI and FI were computed as well as
46 distractor analyses done. Based on the scores for the answer tier only, FI ranges from 0.06 to
47 0.61, with an average value of 0.28 (SD = 0.17). For getting both-tiers correct, FI ranged
48 from 0.00 to 0.44. As none of the questions was deemed too easy, all the 24 questions were
49 retained. Further validation indicated that no revisions were necessary.
50
51
52
53
54
55
56
57

58
59
60 A sample question is shown below:

21. When a given volume of 1 mol/dm^3 hydrochloric acid is mixed thoroughly with an **equal volume** of 1 mol/dm^3 potassium hydroxide, the resulting solution has

- A more H^+ than OH^- ions.
 B more OH^- than H^+ ions.
 C no H^+ and no OH^- ions.
 D same number of H^+ and OH^- ions.

Confidence Rating	1	2	3	4	5	6
	Just Guessing	Very Unconfident	Unconfident	Confident	Very Confident	Absolutely Confident

Scientific reason for my answer:

- A All the H^+ from hydrochloric acid are completely neutralised by the same number of OH^- from aqueous potassium hydroxide, leaving no H^+ and no OH^- .
 B This is due to the self-ionisation of water.
 C Hydrochloric acid is a stronger acid while potassium hydroxide is a weaker base. More H^+ than OH^- are formed.
 D Potassium hydroxide is a stronger base while hydrochloric acid is a weaker acid. More OH^- than H^+ are formed.

Confidence Rating	1	2	3	4	5	6
	Just Guessing	Very Unconfident	Unconfident	Confident	Very Confident	Absolutely Confident

Main Study

On top of the 24 questions, it was decided to include an additional question to form a 25 item 4TMC instrument for the main study. This item was unique in that no combination of options from the given answer and reason tiers was correct. Students would have to exercise the option of providing their own answer and reason to score a mark if they feel that none of the options provided were satisfactory. All students were briefed about this possibility before the test but no specific reference to any question was indicated. This was an attempt to try out

1
2
3 whether a 4TMC question can ‘mimic’ a traditional opened-ended question where students are
4 to provide their own responses. As part of a diagnostic test, an item like this could help to
5 surface ACs that may not be covered by the options provided. If no or only few students opt to
6 give their own answer and reason, it may indicate that the sample, on the whole, is relying on
7 common test-taking strategies to answer the question.
8
9
10
11
12
13
14
15
16

17 A different sample of 141 students (71 females, 70 males) participated in the main study. As
18 with the preliminary and pilot studies, the samples are of diverse ethnicity – predominantly
19 Chinese followed by Malays, Indians and others.
20
21
22
23
24
25
26

27 **Treatment of Data**

28
29
30
31 For the 4TMC questions in the pilot phase and main study, FI and DI for the answer, reason
32 and both-tiers were computed for overall scores and individual test items. T-tests and
33 Conbach alphas were computed using Microsoft Excel.
34
35
36
37
38
39
40

41 For each 4TMC item, the answer and reason responses were scored separately. When
42 analyzing the answer tier, a score of 1 was given for a correct response and zero was given
43 for a wrong response. When analyzing both-tiers, a score of 1 was awarded only if correct
44 responses for both answer and reason tiers were given by the student; otherwise, zero score
45 was assigned. In addition, the confidence ratings for each option chosen were analyzed, and
46 used as a rough gauge to assess the strengths of the ACs. The relevant confidence measures
47 are as follows:
48
49
50
51
52
53
54
55
56

- 57 (a) CF : mean confidence rating.
58
59 (b) CFC : mean confidence ratings for an item with a correct response.
60

- (c) CFW : mean confidence ratings for an item with an incorrect response.
- (d) CAQ (Confidence Accuracy Quotient)
- (e) CB (Confidence Bias)

Furthermore, the ACs obtained were categorised based on the classification developed by Caleon & Subramaniam (2010):

- (a) Significant alternative conception (SiAC)
- (b) Spurious alternative conception (SpAC)
- (c) Genuine alternative conception (GAC)
- (d) Moderate alternative conception (MAC)
- (e) Strong alternative conception (SAC)

(Please refer to Appendix A for details.)

Treatment of missing data

For the diagnostic test, missing data was treated by the list-wise deletion approach so that complete case analysis can be done. This approach would generally lead to diminution in the power of the statistical tests conducted owing to the reduction in sample size. This can be addressed significantly if the data can be shown to be missing completely at random (MCAR) (Heitjan & Basu, 1996). If data were MCAR, then the reduced sample size can be considered to be a random sample of the larger sample, in which case the statistical power of the tests would be essentially preserved or reduced only minimally. To check for MCAR, the sample was bifurcated into data sets that were complete and those with one or more missing elements. The distribution properties of the data sets were examined. If t-tests for the nonrally distributed data show no significant differences, then the data can be considered to be MCAR, and estimates obtained can be considered to be reasonably unbiased.

Results and Discussion

Missing data

While collating the data, it was found that, a significant number of scripts contained missing information. Analyses of the missing information were carried out, and it was found that the data are predominantly missing completely at random (MCAR) (Appendix B).

Using a smaller sample size comprising the 92 ‘perfect’ scripts would have only a minimal impact on the statistical power of the tests, owing to the reduced sample size, but valid statistical inferences can still be obtained for this sample size. That is, the threats to internal validity are likely to be less.

The reduced sample size in the present study is still especially good when compared to the reduced sample of 89 students used in the study of McClary and Bretz (2012). These researchers also faced the same issue of missing data. Out of the initial 104 participants, only 89 were included in the final data analysis, disregarding those with missing answers or confidence ratings. The authors, however, did not provide statistical justification for the use of only 89 samples, even though the number of scripts with missing data is small – 15 out of 104, or about 14.4%. As the percentage of missing data is small, it is likely that threats to internal validity are minimal.

In view of the foregoing findings, it was decided that the present study would concentrate on the reduced sample size of 92 for three reasons. Firstly, this sample size is comparable to other studies reported in this area; secondly, there is statistical justification advanced to show that the 47 scripts with missing data can be safely omitted from the analysis; and, thirdly,

1
2
3 there is precedence in the literature to disregard missing information in a study using a multi-
4 tier instrument.
5
6
7
8
9

10 **Test Statistics**

11
12
13
14 Cronbach alpha (α) was used to determine the internal consistency of the ABCDI, based on
15 the test scores for answer, reason and both tiers. It was found to be low for the answer tier
16 (0.31) and reason tier (0.38). It rose when both answer and reason tiers are considered
17 together (0.44). These statistics imply an advantage of complementing the answer tier with a
18 reason tier in assessing learning and understanding. The criterion-referenced nature of the test
19 items in ABCDI could have reduced the variability of the test scores. This could result in
20 deflation of the values of alpha as reliability measures depend on the variability in the scores
21 (Popham & Husek 1969) and “variability is not a necessary condition for a good criterion-
22 referenced test” (p3). The present study also involved various sub-topics that were based on
23 different concepts within acid-base Chemistry. The test was therefore not likely to be uni-
24 dimensional (Schmitt, 1996) as the various sub-topics tested did not form a distinctive
25 construct and so the test measured more than one attribute, which could have lowered the
26 value of alpha. The results are similar to those found in the study on waves using 4TMC
27 items (Celeon & Subramaniam, 2010), where the Cronbach alphas for the three
28 corresponding tiers were 0.40, 0.19 and 0.50 respectively. Low Cronbach alphas were also
29 found in a recent study on undergraduates’ ACs on chemical thermodynamics (Sreenivasulu
30 & Subramaniam, 2013) using 4TMC test and by McClary & Bretz, 2012) in their 4TMC test
31 instrument on acid strength.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58

59 The proportion of students who gave correct answers as well as the values of relevant
60 confidence measures per item of ABCDI (N=92) is summarized in Table 8 (Appendix C).

1
2
3 The test is considered difficult for the students as the mean proportion of students providing
4 correct responses is well below 0.5 (Answer tier = 0.30, reason tier = 0.26 and both tiers =
5 0.16). The low FIs are not a major concern as the test was not meant to determine the degree
6 of mastery of a predefined set of learning objectives for assessment purposes. Tests for
7 assessment or placement need a 'spread of scores' to account for the range of concepts or
8 skills examined. The questions also need to be reasonably 'doable'. In contrast, for a
9 diagnostic test, the aim is to surface ACs, and therefore the items used tend to delve rather
10 deep into the topic matter. With two tier items, the difficulty levels are amplified. A low FI
11 does not necessarily imply that an item is unreasonably difficult. Instead, it is likely to be
12 pointing to the existence of an AC or lack of knowledge on a number of items among the
13 students. The DIs, based on scores for both tiers correct, range from 0.00 to 0.70, with an
14 average of 0.24 (SD = 0.20). This is again close to that of the pilot test (-0.03 to 0.69, average
15 0.24). Overall discrimination for the entire instrument, based on scores for both tiers, is
16 considered fair.
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Some patterns can be observed when examining Table 8 for students' confidence ratings. For
4 Q13, 15, 16, 18, 19 and 21, the average confidence rating (both tiers) of students who
5 provided the correct answers (CFC) is below 3.5. This is interpreted to mean that they are not
6 too confident about their answers, even though they scored correctly for these questions.
7
8 Overall, the CFC for the items has a mean value of 3.64 (SD = 0.49).
9
10
11
12
13
14
15
16

17 For about half of the questions, the average confidence ratings of students getting wrong
18 answers (CFW) are higher than 3.5. For questions 1, 2, 4, 6, 11, 12, 14, 15, 17, 18, 19, 21, 22
19 and 23 (14 questions in total), in respect of answer, reason and both tiers, CFWs are higher
20 than the corresponding CFCs. This corresponds to negative CAQs (in one or more of the
21 tiers), indicating the presence of overconfidence. For three of the questions (Q 11, 15 and 21),
22 the students are overconfident across all three tiers.
23
24
25
26
27
28
29
30
31
32

33 The mean CAQs are low (-0.03 for answer tier, 0.07 for reason tier and -0.01 for both tiers),
34 suggesting that, in general, the students could not discriminate well between what they know
35 and what they do not know. In cases of negative CAQs, it is quite likely that ACs may have
36 existed in the schema of the students' knowledge.
37
38
39
40
41
42
43
44
45

46 The CB value (both-tiers) for every item is positive (varies from 0.12 to 0.56), which implies
47 a possibility of over confidence among the students. This is a little surprising as Asian
48 students with predominately Confucian heritage have a greater tendency to be more modest
49 and would not rate themselves highly when asked to self-assess their own performance (Ho,
50 2009). A possible explanation could be that the ACs are deeply entrenched such that the
51 students believed that they have answered correctly according to their 'knowledge and
52 understanding'. The 10 items with the highest CB (for both-tiers) are Q23, 6, 17, 25, 11, 15,
53
54
55
56
57
58
59
60

1
2
3
4 22, 12, 18 and 21, in descending order of CB values.
5
6
7

8 The foregoing observations did not come as a surprise as they were also flagged off in the
9 findings of the preliminary and pilot phases.
10
11

12
13
14
15 Item 25 is a 'special' item, where none of the options provided for the answer and reason tiers
16 are correct. The distractor analysis is summarized in Table 9. The most popular answer-
17 reason combination was D and B (27.17% with a CF of 3.46 for around a quarter of the
18 respondents), followed by the second most popular combination of B and A. Only 2.17%
19 (with CF of 3.75) of the students chose E–E combination (i.e. none of the given answer-
20 reason combinations is correct), which ought to be the 'correct' set of responses provided the
21 written comments are valid. However, after going through all the self-provided responses by
22 the students, none of the responses was found to be acceptable. Answer option D accounted
23 for 48.91% – nearly half of the sample size. The lower percentage of students choosing this
24 free response combinations suggest a likelihood of students' fixation on the idea that the
25 correct answers must be among the given choices (even though it was mentioned on the cover
26 page of the instrument as well as at the briefing before commencement of the test that
27 students can provide their own answer(s) and reason(s) should they find that none of the
28 given options for an item was satisfactory). Apparently, the idea that the pH of an aqueous
29 solution of an acid depends only on its concentration (answer option D) is a very prevalent
30 AC. The next most popular answer option was B (that the pH of an aqueous solution of an
31 acid depends on the strength of the acid only), which accounted for 20.65%. These two
32 options made up a total of 69.56%, which are more than two-thirds of the sample size.
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
Apparently, the students have only partial understanding of the factors affecting the value of
pH. In addition, despite the fact that the term 'only' should be providing some hints that each

of the options alone is not encompassing enough; the very low percentage of students that decided that none of the answer-reason combinations is correct is far too low at 2.17%. These observations can be interpreted in terms of the prevalence of incomplete understanding of the relevant concepts. Perhaps, there might be a certain extent of reluctance on the part of the students to venture beyond the provided options.

		Answer				
		A	B	C	D	E*
Reason	A	0.00% (-)	17.39% (3.56)	4.35% (3.50)	1.09% (4.00)	0.00% (-)
	B	1.09% (3.50)	0.00% (-)	5.43% (3.30)	27.17% (3.46)	5.43% (3.60)
	C	1.09% (3.00)	3.26% (4.50)	0.00% (-)	14.13% (3.77)	0.00% (-)
	D	10.87% (3.45)	0.00% (-)	0.00% (-)	2.17% (3.25)	0.00% (-)
	E*	0.00% (-)	0.00% (-)	0.00% (-)	4.35% (4.00)	2.17% (3.75)*

Table 9 Distractor Analysis for Item 25

* Student self-provided answer or scientific reason
Numbers in brackets are the corresponding confidence ratings

The 30 ACs found are summarized in Table 10 (Appendix D). A discussion of these ACs follow.

Alternative conceptions identified

The ACs identified are discussed in terms of the groups they occur under. Where available, comparisons are also made with respect to findings from other studies.

Properties of Acids and Bases

Students think that acids are associated with corrosiveness (AC1), which is similar to the findings of others (Treagust, 1988, Demircioğlu, Ayas & Demircioğlu, 2005). A significant number of students (36%) also considered acids as being more dangerous and reactive than bases (AC2), which is parallel to “*Acids melt metals, acids are strong and bases are not strong*” (Nakhleh & Krajcik, 1994) or “*Acids are more powerful than bases*” (Sheppard, 2006). It also partially coincides with “*Acids burn and melt everything*” (Özmen et al., 2009, 2012). Both AC1 and AC2 are spurious ACs (SpAC, meaning $CF < 3.5$), which means that they are more likely to be due to lack of knowledge or understanding. Students may have encountered more acids than bases in their daily lives (even tertiary students know more about acids than bases, Cros et al., 1986) and media reports on accidents and attacks involving acids may have led to the preconceptions that acids are rather reactive and dangerous.

The rather high percentage (52%) of students thinking that rain water in an unpolluted area is neutral (AC3) came as a surprise. This is comparable to Banerjee’s (1991) findings that 71% of chemistry teachers and 79% of undergraduate students exhibited the same AC. Among the 52%, about 19% supported their case arguing that rain water is pure water. AC3 is a strong AC (mean $CF \geq 4.0$), which warrants attention as it is likely to be quite deep-seated. We suspect that the students might have been misled by the term ‘unpolluted’ to some extent, so much so that they assumed that the precipitation from the sky would stay chemically pure in

1
2
3 an unpolluted area even though they should be aware that naturally occurring constituents
4 (carbon dioxide in this case) in the atmosphere will dissolve in the rain to make it acidic.
5
6
7

8
9
10 A naïve and superficial correlation of chemical structures with acidity or basicity may explain
11 why more than a quarter (29%) of the students believed that compounds containing H will
12 produce H^+ whilst compounds containing OH will produce OH^- (AC4). These were also
13 found in other studies (Demircioğlu, Ayas & Demircioğlu, 2005; Huang 2004; Özmen et al.,
14 2012). AC4 is a moderate AC (MAC). Simplistically relating compounds containing H as
15 acid is not that surprising. However, the only weak base that students are familiar with at this
16 level is aqueous ammonia. Perhaps the frequent encounters and familiarity with strong bases
17 such as sodium and potassium hydroxides could have dominated their thinking.
18
19
20
21
22
23
24
25
26
27
28
29
30

31 AC5, a spurious AC, is probably due to a failure to recall the definition or may also be a
32 language issue – misinterpretation of ‘basic’ in the term ‘basicity’, that leads students to think
33 that the higher basicity of an acid makes it more basic! The strength of this AC is not high
34 and it could be addressed relatively easily once the distinction between basicity and basic is
35 made clearer.
36
37
38
39
40
41
42
43
44
45

46 There was also a group of students (12%) that have yet to understand that the mechanism of
47 electrical conduction in aqueous solutions is different from that in metals, and believed
48 electrons are also responsible for the conduction of electricity in acid solutions (AC6, MAC).
49 This is similar to the finding in Othman et al. (2008)’s study, which revealed students’ belief
50 that an ionic compound does not conduct in the solid state but in the aqueous state due to the
51 production of free electrons formed in an aqueous solution. These are spurious to moderate
52 ACs, which would require better understanding of the ions present and how they affect the
53
54
55
56
57
58
59
60

1
2
3 conductivity of the solutions in order to be addressed.
4
5
6

7
8 *Strengths of acids and bases*
9

10 For those students asserting that pure ethanoic acid is a stronger acid than aqueous ethanoic
11 acid (AC7), their reasoning was that pure ethanoic acid produces more hydrogen ions. This
12 moderate AC could also be another case of language issue in that the term ‘pure’ may have
13 given a connotation of ‘without impurity’ and so naturally ‘more’ acid is present to produce
14 more hydrogen ions. The fact that an aqueous medium is needed for the acid to dissociate to
15 form hydrogen ions was somehow ignored or side-lined. This would not be too difficult to
16 address once students understand that water is needed before the typical behaviors of acids
17 can be manifested.
18
19
20
21
22
23
24
25
26
27
28

29
30
31 The belief that a stronger acid is one that produces a higher hydrogen ion concentration (AC8)
32 or one with a higher initial concentration (AC9) indicates incomplete understanding of the
33 definitions of strong and weak acids (full versus incomplete dissociation) as well as a partial
34 picture of the factors responsible for the quantity of hydrogen ions produced. These are
35 moderate to strong ACs, indicating quite deeply entrenched beliefs. The same flawed
36 argument was also extended to weak acids in AC13 (MAC). Student thought that a weak acid
37 that produced more hydrogen ions was the stronger of the two weak acids. Having a higher
38 concentration of hydrogen ions could be due to more than one factor, of which acid strength
39 is just one of them. This is consistent with the findings for item 25, where less than 10% of
40 the students rejected all the options where only a single factor is responsible for the value of
41 pH (a measure of hydrogen ion concentration). These ACs are not unique among the students
42 as they are also identified in other studies: “Concentrated acid is strong acid” (Demircioğlu et
43 al., 2005; Özmen et al., 2009; Ross, 1989). It did not really come as a surprise that the same
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 flawed conceptions were also extended to bases – a stronger base is one with a higher initial
4 concentration (AC11, SpAC) or a higher initial number of moles (AC12, SAC). Apparently,
5
6 the students transferred the same flawed thinking to bases.
7
8
9

10
11
12 AC10 again demonstrated another instance of “the more the merrier” way of thinking and
13 therefore, a dibasic acid is considered to be stronger than a monobasic acid. It is not
14
15 surprising that this has also been reflected as a strong AC as the students think that the
16
17 potential ability to produce twice the amount of hydrogen ions with the same starting
18
19 concentration would ‘naturally’ imply a stronger acid. This AC is similar to the sample of
20
21 students in the study in Turkey, who also thought that if an acid has more H in the formula,
22
23 its acidity will increase (Demircioğlu et al., 2005; Özmen et al., 2009).
24
25
26
27
28
29
30

31
32 For AC14 (MAC), students are content with the understanding that a dibasic weak acid
33
34 produces twice the amount of hydrogen ions than a weak monobasic acid. This is closely
35
36 related to AC8. The medium strength of this AC indicates that it is a genuine AC. This is
37
38 most probably a (wrong) extrapolation of AC8 to weak acids. In addition, a dichotomous
39
40 mode of thinking may be involved, that is, either 100% dissociation or partial dissociation,
41
42 and all ‘partial’ dissociations are equally partial. Unfortunately, ‘partial’ has a wide range
43
44 (more than 0% to less than 100%) and the nature of the acid also matters.
45
46
47
48
49

50 *pH*

51
52 Almost two-thirds (65%) of the students possess the strong AC that pH ranges from 1 to 14
53
54 (AC16, SAC) which agreed with Zoller’s (1990) findings. Illustrations and diagrams in
55
56 textbooks as well as teachers’ instructional materials could be the source. This could also
57
58 account for the high strength of the AC as students normally do not question what they see in
59
60

1
2
3 instructional materials. Without clear emphasis that the range shown is for illustration
4 purposes only, students could mistake it as the absolute range that pH values would have. In
5
6 view of the high strength and high proportion of students having this AC, it is one that must
7
8 be addressed or, better still, be preempted while teaching the topic to deter its formation.
9
10
11 Inadequate teaching may have led to incomplete learning.
12
13
14
15
16

17
18 For students with AC17 (SpAC), it would be expected that they would avoid choosing an
19 unfamiliar answer if they are not aware of the self-ionization of pure water (in the reason tier).
20
21 However, the more important issue is that the students should not have the idea that pH of
22 pure water would be increasing when exposed to air (in the answer tier). With their
23
24 knowledge of the presence of carbon dioxide in air and/or acid rain, they should deduce that
25
26 the water should be getting more acidic, accompanied by a decrease in pH. One possible
27
28 explanation could be that they were aware that the water was getting acidic but got confused
29
30 over the definition of pH and thought that an increase in acidity should be followed by a
31
32 corresponding decrease in pH instead of an increase. Similar findings from others
33
34 (Demircioğlu et al., 2005; Özmen, 2009; Sheppard, 2006) indicated that students associated
35
36 pH with acidity only. Either owing to a lack of understanding of the mathematics or the
37
38 “more” mentality, students could have thought more hydrogen ions (assuming it means a
39
40 higher hydrogen ion concentration to them) leads to a higher (instead of lower) pH. This type
41
42 of thinking was also exhibited by some Turkish students, with the thinking “As the value of
43
44 pH increases, acidity increases.” (Demircioğlu et al., 2005)
45
46
47
48
49
50
51
52
53
54

55
56 A partial understanding that more than one factor contributes to the concentration of
57
58 hydrogen ions in an aqueous solution is responsible for the thinking that the pH of an aqueous
59
60 acid solution depends only on either the concentration of the acid (AC19, 17%, MAC) or on

1
2
3 the strength of the acid (AC20, 14%, SAC). Sheppard (2006) also found that students thought
4 that pH measures strength (of an acid). These medium to strong ACs are likely to be the
5 result of partial understanding of the concepts related to strengths of acids and pH. They can
6 be expected to improve once the ACs associated with strengths of acids are addressed.
7
8
9
10
11

12
13
14
15 It was found that about a quarter of the students (27%) thought that pH is a measure of the
16 total number of moles of hydrogen ions (AC21, MAC). This is closely related to “pH is a
17 measure of acidity” (Demircioğlu et al., 2005; Demircioğlu et al., 2009; Sheppard, 2006).
18 Another 14% thought that the pH of an aqueous acid is independent of its concentration
19 (AC22, SAC), which is rather illogical. The prevalence and strength of these two ACs can be
20 reduced if students are formally introduced to the mathematical definition of pH and its
21 implications rather than a cursory introduction.
22
23
24
25
26
27
28
29
30
31
32

33 34 *Neutralization*

35
36 AC29 is very prevalent (found in items 21 to 23, ranging from 12% to 41%, and from SpAC
37 to SAC). Other studies found the same AC (Demircioğlu et al., 2005; Özmen et al., 2009) or
38 a closely related one such as: the solution formed after neutralization is neutral (Zoller, 1990,
39 Schmidt, 1991, Huang, 2004). Students are well aware that neutralisation is between H^+ and
40 OH^- and might have treated it as a kind of ‘cancellation reaction’ where all the H^+ are
41 ‘cancelled out’ by OH^- to form water molecules.
42
43
44
45
46
47
48
49
50
51
52

53 AC30 (SAC) is analogous to AC27 (which has probably been ‘extrapolated’ to AC30), in that
54 the cations and anions of the salt formed after complete neutralization are paired up. This is a
55 strong AC that needs to be addressed. The same ‘explanation’ regarding electrostatic
56 attraction as before (as for AC27) should be applicable to this AC. Students need to
57
58
59
60

1
2
3 understand the difference in environment between the solid (ionic lattice) and the aqueous
4
5 solution for ionic compounds.
6
7

8 9 10 *Indicators*

11
12 That there are no H^+ left to react with an indicator at the end point and so the indicator
13
14 changes color (AC15), was the belief of a substantial proportion of the students (44%). There
15
16 are actually two issues concerning this moderate AC. Firstly, there is the supposition that an
17
18 indicator changes color when there are no hydrogen ions to react with, which implies the
19
20 notion of “something happens when nothing is present”. This is quite an unusual finding as a
21
22 reasonably good chemistry student would understand that no change would happen if there is
23
24 no reaction. The actual cause of this problem probably has something to do with the
25
26 fundamental understanding of chemical reactions, rather than a specific issue with acids and
27
28 bases. Secondly, it is linked to AC30 (to be discussed later), where students thought that there
29
30 was neither hydrogen ions nor hydroxide ions left at the end-point (Demircioğlu et al., 2005;
31
32 Özmen et al., 2009, Schmidt, 1991).
33
34
35
36
37
38
39
40

41 *Sub-microscopic views of acids and bases*

42
43 There seems to be apparently quite widespread confusion in terms of the ions present in pure
44
45 ethanoic acid. When pure ethanoic acid is compared with an equimolar amount of aqueous
46
47 ethanoic acid, students thought that
48
49

- 50
51 (a) pure ethanoic acid produces the same number of moles of H^+ as an equimolar amount
52
53 of aqueous ethanoic acid. (AC 23, 16%, SAC)
54
55 (b) pure ethanoic acid produces more H^+ than aqueous ethanoic acid. (AC24, 20%, SAC)
56
57 (c) pure ethanoic acid (in the liquid state) produces fewer H^+ than aqueous ethanoic acid.
58
59 (AC25, 15%, SpAC)
60

1
2
3 There were takers for all the possible (same, more and fewer) wrong answers that were also
4 fairly evenly distributed. It seems that there is a fundamental lack of understanding that pure
5 ethanoic acid, in the absence of water, is not expected to ionize to produce ions. Students
6 have been taught that an aqueous medium is needed for the manifestations of acidic and basic
7 behaviors.
8
9

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Compartmentalized thinking involving exclusive association of hydroxide ions with alkalis, and hydrogen ions with acids could have accounted for students' thinking that there are no hydrogen ions present in an alkaline solution (AC26, SAC). Banerjee (1991) also found the same AC in his sample of students.

Students thought that the sodium and hydroxide ions in aqueous sodium hydroxide are paired up (AC27) – this is attributed to the failure to understand that although the ions are held together by electrostatic attraction in the ionic lattice (solid state), the process of dissolution would require them to separate from each other. This spurious to moderate AC is consistent with other studies (Nyachwaya et al., 2011; Rosenthal and Sanger, 2012; Tien, et al., 2007), where students thought that ions from an ionic compound are paired up in aqueous solution. In addition, students have been taught that bases dissociate to form hydroxide ions when dissolved in water. The understanding that electrostatic attraction is strong might have led to the thinking that the attraction is still holding the ions in pairs, even after dissolution.

For AC28 (SpAC), students (15%) recognized that an equilibrium is established between ammonia and ammonium ions. However, they concluded that there are more ammonium ions than ammonia molecules (answer tier), and that the stabilization of the ammonium ions by the water molecules contributed to a shift in the equilibrium to favor the formation of

1
2
3 ammonium ions (reason tier). The students should reasonably have got the answer tier correct.
4
5 Since ammonia is a weak base (partial dissociation), it should convey an idea that only a
6
7 small amount of ammonia would react with water to form ammonium ions, leaving most of
8
9 the ammonia molecules undissociated. A possible remedy suitable at this level would have to
10
11 come in the form of informing students about the 'facts' and/or reasoning using the concept
12
13 of 'weak' base. This should be effective as the strength of the AC is low.
14
15
16
17
18
19

20 **Conclusion**

21
22 The identification of several ACs in this study on acid-base chemistry suggests that the
23
24 traditional teaching the students in this study have undergone has not been very effective in
25
26 building a solid understanding on acid-base chemistry. It is most likely that instruction has
27
28 focused on the syllabus content, which is understandable from the point of view of
29
30 examinations. A more effective way would be by being proactive, that is, by ascertaining the
31
32 ACs on the topic and addressing these also during traditional teaching.
33
34
35
36
37
38

39 A total of 30 ACs were identified in the main study using the 25-item 4TMC test. Some
40
41 examples of ACs that were found in this study but not found in the reviewed literature on
42
43 acid-base chemistry are:
44
45
46
47

- 48 (a) The higher the basicity of an acid, the more basic it is (11%).
49
50 (b) pH of an aqueous acid is independent of its concentration (14%).
51
52 (c) Pure ethanoic acid is a stronger acid than aqueous ethanoic acid (17%).
53
54 (d) Pure ethanoic acid (in the liquid state) produces more H^+ than aqueous ethanoic acid
55
56 (20%).
57
58
59
60

- 1
2
3 (e) Equimolar amounts of aqueous and pure ethanoic acid produce the same number of
4 moles of H^+ (16%).
5
6
7
8 (f) There are more NH_4^+ than NH_3 molecules in aqueous ammonia (15%).
9
10

11
12 With the exception of the ACs concerning strengths of acids (and bases) and ions present in
13 aqueous solutions, we believe that some of the remaining ACs could be easily addressed pre-
14 emptively during teaching. A good example is the AC concerning the range of pH, where the
15 teacher could point out that the pH charts in textbooks just serve as illustrations and that the
16 range of pH can go beyond those values shown.
17
18
19
20
21
22
23
24
25
26

27 **Implications**

28
29
30 The study has some implications:
31

- 32
33 (a) The 25 4TMC instrument has helped to identify 30 ACs. This is considered to be
34 effective when compared to other methods of data acquisition (such as MCQs and
35 interviews) in terms of time needed for testing, scoring and extraction of relevant
36 information. Teachers can use the diagnostic instrument straightaway.
37
38
39
40
41
42 (b) The 4TMC test also provides more information beyond the traditional one-tier MCQ
43 and two-tier MCQs as the extra tiers of confidence ratings provide metacognitive
44 information that helps to surface qualitatively students' levels of certainty for their
45 responses and, to a first approximation, how deep-seated these ACs are
46
47
48
49
50
51 (c) Computations of indices such as CAQ and CB provide additional insights into the
52 students' perceived confidence of their performance. These also provide, in addition
53 to percentages of ACs, additional indications of how self-aware students are of their
54 own performances.
55
56
57
58
59
60 (d) The use of a 'trick' question in this study where all the stated options are incorrect and

1
2
3 where students need to provide their own answer and reason in the blank space
4 provided has implications for test developers. For example, it is the norm, especially
5 in MCQs, two-tier MCQs, three tier MCQs and four-tier MCQs that the correct
6 answer is in one of the responses. Students can get this correct either through correct
7 content knowledge, partial knowledge or guesswork. It is possible to enhance the
8 robustness of the test by having perhaps one or two questions where the given options
9 are all wrong and where students would need to provide their own answers. Of course,
10 space needs to be provided for each question in such tests for them. This can lengthen
11 the format of the test but it has the additional advantage of penalizing guessing. The
12 ‘trick’ question is a novelty that would be of interest for future research. In this study,
13 no student answered this question correctly.

- 14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30 (e) The use of practitioner knowledge to craft distracters for the MCQs is also a useful
31 approach. Though this approach is not without its drawbacks, it has been shown in
32 this study that in place of interviews (which are time-consuming), this is also an
33 approach worth considering. The first author’s teaching experience has come in useful
34 in this regard.

35
36
37
38
39
40
41
42 Overall, this study has provided further support for the use of the 4-tier format in diagnostic
43 studies on ACs.

44 45 46 47 48 49 **Contribution to the literature**

50
51
52 The contribution of this study to the literature can be summarized as follows:

- 53
54
55 (a) While previous studies on acid-base chemistry have documented several ACs, a
56 number of these may not be true ACs – some of these could be due to lack of
57 knowledge, which are not apparent when traditional MCQs, two-tier MCQs or certain
58
59
60

1
2
3 other approaches are used. In the present study, the use of confidence ratings has
4
5 helped significantly to discriminate ACs from responses due to guessing or lack of
6
7 knowledge.
8
9

10 (b) The use of a ‘trick’ question in the test items has shed some useful light – elaborated
11
12 in item (d) in the section on implications. It is worth exploring this aspect further in a
13
14 separate study.
15
16
17
18
19
20

21 **Limitations**

22
23 The relatively small sample size does not represent the entire chemistry student population in
24
25 the country. There could also be some subjectivity on the part of students in expressing their
26
27 confidence to their responses to the questions.
28
29
30
31
32

33 **Acknowledgement**

34
35 Our thanks go to the schools and students which participated in this study. We also thank the
36
37 reviewers for their constructive comments on an earlier version of this manuscript.
38
39
40
41

42 **References**

- 43
44 Artdej, R., Ratanaroutai, T., Coll, R. K., & Thongpanchang, T. (2010). Thai Grade 11
45
46 students’ alternative conceptions for acid–base chemistry. *Research in Science &*
47
48 *Technological Education, 28*(2), 167–183
49
50
51
52 Banerjee, A. C. (1991). Misconceptions of students and teachers in chemical equilibrium.
53
54 *International Journal of Science Education, 13*(4), 487–494.
55
56
57 Caleon I, & Subramaniam, R. (2010). Do students know what they know and what they don’t
58
59 know? Using a four-tier diagnostic test to assess the nature of students’ alternative
60
conceptions. *Research in Science Education, 40*(3), 313–337.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
- Cetin-Dindar, A., & Geban, O. (2011). Development of a three-tier test to assess high school students' understanding of acids and bases. *Procedia-Social and Behavioral Sciences*, 15, 600–604.
- Chiu, M. H. (2004). An investigation of exploring mental models and causes of secondary school students' misconceptions in acids-bases, particle theory, and chemical equilibrium. Annual Report to the National Science Council in Taiwan (in Chinese) Taiwan: National Science Council. Quoted in Chiu, M. H. (2007). A national survey of students' conceptions of chemistry in Taiwan. *International Journal of Science Education*, 29(4), 421–452.
- Chiu, M. H. (2007). A national survey of students' conceptions of Chemistry in Taiwan. *International Journal of Science Education*, 29(4), 421–452.
- Cros D., Maurin, M., Amouroux R., Chastrette M., Leber J., & Fayol M. (1986), Conceptions of first-year university students of the constituents of matter and the notions of acids and bases. *European Journal of Science Education*, 8, 305–313.
- Demircioğlu, G., Ayas, A., & Demircioğlu, H. (2005). Conceptual change achieved through a new teaching program on acids and bases. *Chemistry Education Research and Practice*, 6(1), 36–51.
- Drechsler M., & Schmidt H. J. (2005). Textbooks' and teachers' understanding of acid-base models used in chemistry teaching. *Chemistry Education Research and Practice*, 6, 19–35.
- Gabel D. L., & Bunce D. M. (1994). Research on problem solving: chemistry. In D. L. Gabel (Eds.), *Handbook of research on science teaching and learning*, Macmillan, New York, pp. 301–326.
- Gabel, D. L., Samuel, K. V., & Hunn D. (1987). Understanding the particulate nature of matter. *Journal of Chemical Education*, 64(8), 695–697.

- 1
2
3 Glassman, S. (1967). High school students' ideas with respect to certain concepts related to
4
5 chemical formulas and equations. *Science Education*, 51, 84–103.
6
7
8 Griffard, P. B., & Wandersee, J. H. (2001). The two-tier instrument on photosynthesis: What
9
10 does it diagnose? *International Journal of Science Education*, 23(10), 1039–52.
11
12
13 Hand, B. M., & Treagust, D. F. (1988). Application of a conceptual conflict strategy to
14
15 enhance student learning of acids and bases. *Research in Science Education*, 18, 53–63.
16
17
18 Heitjan, D. F.; Basu, S. (1996). Distinguishing "missing at random" and "missing completely
19
20 at random". *The American Statistician*, 50(3), 207–213.
21
22
23 Herron, J. D. (1975). Piaget for chemists. Explaining what 'good' students cannot understand.
24
25 *Journal of Chemical Education*, 52(3), 146–150.
26
27
28 Howell, D. C. (2007). *The analysis of missing data*. In Outhwaite, W., & Turner, S.
29
30 Handbook of Social Science Methodology. London: Sage.
31
32
33 Huang, W. C. (2004). The types and causes of misconceptions of elementary students on
34
35 acids–bases. Annual Report to the National Science Council in Taiwan (in Chinese)
36
37 Taiwan: National Science Council. Quoted in Chiu, M. H. (2007). A national survey of
38
39 students' conceptions of chemistry in Taiwan. *International Journal of Science*
40
41 *Education*, 29(4), 421–452.
42
43
44 Juang, C. S. (2004). *The misconceptions of secondary school students on material science*
45
46 *and organic compounds*. Annual Report to the National Science Council in Taiwan (in
47
48 Chinese) Taiwan: National Science Council.
49
50
51
52
53
54
55
56
57
58
59
60
Kavanaugh R. H., & Moomaw W. R. (1981). Including formal thought in introductory
chemistry students. *Journal of Chemical Education*, 58, 263–265.
Kozma, R., & Russell, J. (1997). Multimedia and understanding: expert and novice responses
to different representations of chemical phenomena. *Journal of Research in Science*
Teaching, 34, 949–968.

- 1
2
3 Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York,
4
5 Wiley.
6
7
8 Lundeberg, M. A., Fox, P. W., Brown, A. C., & Elbedour, S. (2000). Cultural influences on
9
10 confidence: country and gender. *Journal of Educational Psychology*, 92(1), 152–159.
11
12
13 McClary, L. M., & Bretz, S. L. (2012). Development and assessment of a diagnostic tool to
14
15 identify organic chemistry students' alternative conceptions related to acid strength.
16
17 *International Journal of Science Education*, 34(15), 2317-2341.
18
19
20 Nakhleh, M. B., & Krajcik J. S. (1994). Influence of levels of understanding as presented by
21
22 different technologies on students' understanding of acid, base and pH concepts.
23
24 *Journal of Research in Science Teaching*, 31, 1077–1096.
25
26
27 Nakhleh, M. B. (1992). Why some students don't learn chemistry: Chemistry misconceptions.
28
29 *Journal of Chemical Education*, 69(3), 191–196.
30
31
32 Nyachwaya J. M., Mohamed A. R., Roehrig G. H., Wood N. B., Kern A. L., & Schneider J. L.
33
34 (2011). The development of an open-ended drawing tool: An alternative diagnostic
35
36 tool for assessing students' understanding of the particulate nature of matter.
37
38 *Chemistry Education Research and Practice*, 12, 121–132.
39
40
41 Özmen, H., Demircioğlu, G., & Coll, R. K. (2009). A comparative study of the effects of a
42
43 concept mapping enhanced laboratory experience on Turkish high school students'
44
45 understanding of acid-base chemistry. *International Journal of Science and*
46
47 *Mathematics Education*, 7, 1–24.
48
49
50
51 Potgieter, M., & Davidowitz, B. (2011). Preparedness for tertiary chemistry: Multiple
52
53 applications of the Chemistry Competence Test for diagnostic and prediction purposes.
54
55 *Chemistry Education Research and Practice*, 12, 193–204.
56
57
58 Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement.
59
60 *Journal of Educational Measurement*, 6(1), 1–9.

- 1
2
3 Ross, B. H. (1989). *High-school students' concepts of acids and bases*. Unpublished master's
4 thesis, Queen's University, Kingston, Ontario.
5
6
7
8 Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley &
9 Sons.
10
11
12 Saglam, Y., Karaaslan, E. H., & Ayas, A. (2011). The Impact of Contextual Factors On The
13 Use Of Students' Conceptions. *International Journal of Science and Mathematics*
14 *Education*, 9(6), 1391–1413.
15
16
17
18 Schmidt H. J. (1991). A label as a hidden persuader: chemists. neutralization concept.
19 *International Journal of Science Education*, 13, 459–471.
20
21
22
23 Sheppard K. (2006). High school students' understanding of titrations and related acid-base
24 phenomena. *Chemistry Education Research and Practice*, 7(1), 32–45.
25
26
27
28 Sreenivasulu, B., & Subramaniam, R. (2013). University students' understanding of chemical
29 thermodynamics. *International Journal of Science Education*, 35(4), 601–635.
30
31
32
33 Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on test of cognitive
34 abilities. *Intelligence*, 25(2), 93–109.
35
36
37
38 Treagust, D. F. (1985). *Diagnostic test to evaluate students' misconceptions in science*. Paper
39 presented at the 58th annual meeting of the National Association for Research in
40 Science Teaching, French Lick Springs, Indiana, April 15 –18, 1985.
41
42
43
44 Treagust D. F. (1988). Development and use of diagnostic tests to evaluate students'
45 misconceptions in science. *International Journal of Science Education*. 10, 159-169.
46
47
48
49 Tsai, C. C., & Chou, C. (2002). Diagnosing students' alternative conceptions in science.
50 *Journal of Computer Assisted Learning*, 18, 157–165.
51
52
53
54 Zoller, U. (1990). Student's misunderstandings and misconceptions in college freshmen
55 chemistry (general and organic). *Journal of Research in Science Teaching*, 27, 1053–
56
57
58
59
60 1065.

Table 1 Summary of studies on alternative conceptions on acids and bases

S/N	Country/ Reference	Sample	Instrument	Nature of ACs/Remarks*
1	France (Cros, Amou ro, Chastrette, Fayol, Leber, & Maurin, 1986)	400 (overall) university year 1 students	Free interviews followed by semi- structured interviews and then a questionnaire- based enquiry	<ul style="list-style-type: none"> No heat is evolved during a reaction between an acid and a base A solution with $\text{pH } 7 \pm 2$ is drinkable.
2	Australia (Hand & Treagust, 1988)	60 16 year old students	Interviews	<ul style="list-style-type: none"> An acid is something which eats material away or which can burn you. Testing for acids can only be done by trying to eat something away. Neutralisation is the breakdown of an acid or something changing from an acid. The difference between a strong and a weak acid is that strong acids eat material away faster than a weak acid/ A base is something which makes up an acid. <p>(p55)</p>
3	Canada (Ross, 1989)	34 grade 12 students who had completed grade 11 advanced chemistry program.	25 items single tier 4-options MCQ. 8 selected students underwent 40 mins clinical interviews.	<ul style="list-style-type: none"> "Acids contain hydroxide ions." "All acids are strong acids." "Concentrated is the same as strong." "Acids are poisonous." "Acid rain is formed from water and chlorine or hydrogen gas." "Acids contain hydrogen in the gaseous state." "Acids and bases react to form a solution." "A strong acid has a higher pH than a weak acid." "A gas is released when an acid and a metal reacts because heat changes the liquid to a vapor." "When hydrochloric acid and magnesium react, more gas is released than (when) acetic acid reacts with magnesium because the reaction is more violent." "When hydrochloric acid and magnesium react, more gas is released (than) when acetic acid reacts with magnesium (as) more hydrogen bonds need to be broken." "A strong acid reacts more slowly than a weak acid." <p>(pp 105 – 107)</p>
4	Israel (Zoller, 1990)	University Freshmen, Number unknown. Based on 15	–	<ul style="list-style-type: none"> "How come an aqueous solution may have a $\text{pH} < 0$ or $\text{pH} > 14$?" (p.1058) "Salts, which are formed in a neutralization of acids with bases, are "neutral" species. As such, their aqueous solutions must be neutral (i.e., the pH is definitely 7)." (p.1059)

S/N	Country/ Reference	Sample	Instrument	Nature of ACs/Remarks*
		years of teaching experience.		
5	India (Banerjee, 1991)	162 undergraduate chemistry students and 69 school teachers	21 item test instrument (7 single tier MCQ, 8 short answer response, 3 on problem solving and 3 on application) on chemical equilibrium	<ul style="list-style-type: none"> • “No hydrogen ions in an aqueous solution of sodium hydroxide or in distilled water.”. p490) • “Rainwater is neutral.” (71% of teachers, 79% of students) (p. 491) • Teachers (76%) and students (46%) believe that “for the same concentration, the pH of acetic acid will be less than or equal to that of hydrochloric acid solution in water.” (pp 490 – 491)
6	Germany (Schmidt, 1991)	7500 grammar-school students (Gymnasium). Test – 154 items from 10 topics.	10 MCQ items on acids and bases.	<ul style="list-style-type: none"> • “In any neutralization reaction, a neutral solution is formed, even if a weak acid or base takes part in the reaction.” • “Neutralization is an irreversible reaction.” (p459)
7	USA (Nakhleh, 1992)	Grade 11 chemistry students	Interview	<ul style="list-style-type: none"> • 20% of the students drew non-particulate representations when asked how an acid or base solution would “appear under a very powerful magnifying glass”,
8	USA (Nakhleh & Krajcik, 1994)	Fifteen grade 11 senior high-school students	Semi-structured Interviews	<ul style="list-style-type: none"> • “pH is inversely related to harmful” and “bases are not harmful”. • “Bubbles or bubbling is a sign of chemical reaction or strength.” • “Acids and bases have their own particular color or color intensity (bases are colored blue, acids are colored pink), and even different pH solutions have different colors.” • “The molecules fight and combine, and phenolphthalein helps with neutralization.” • “Acids melt metals, acids are strong and bases are not strong.” • “pH was regarded as a compound called phenolphthalein, a chemical reaction and a number related to intensity.” (pp 1087 – 1090)
9	Sweden (Drechsler & Schmidt, 2005)	7 upper secondary students – high achievers	Interviews	<p>When applying the Bronsted theory, students showed some problems:</p> <ul style="list-style-type: none"> • Many represented the reaction between dilute hydrochloric acid and aqueous sodium hydroxide without the formation of water. • Majority of students did not consider water as the acid in the dissociation of ammonia in water.

S/N	Country/ Reference	Sample	Instrument	Nature of ACs/Remarks*
				<ul style="list-style-type: none"> In the reaction of nitric acid with copper, nitric behaves as an acid.
10	Turkey Demircioğlu, Ayas, & Demircioğlu, (2005)	88 tenth grade secondary school students	Worksheets: 3 sections 20 item one- tier MCQ 'Concept Achievement Test'.	<ul style="list-style-type: none"> "Acids burn and melt everything" "All acids and bases are harmful and poisonous" "As pH increases, acids become harmless and bases are not harmful" "Different pH solutions have different colors" "pH is a measure of acidity" "A strong acid doesn't dissociate in water solution, because its intra-molecular bonds are very strong" "The only way to test a sample whether it is an acid or a base is to see if it eats something away, for example metal, plastic, animal, and us" "All salts are neutral" "Salts don't have a pH value" "In all neutralization reactions, acid and base consume each other completely" "At the end of all neutralization reactions, there are neither H^+ nor OH^- in the resulting solutions" "A strong acid is always a concentrated acid" "Bubbles or bubbling is a sign of chemical reaction or strength of an acid or a base" "Indicators help with neutralization" "As the value of pH increases, acidity increases" "While bases turn blue litmus paper into red, acids turns red litmus paper into blue" "As the number of hydrogen atoms increases in the formula of an acid, its acidity becomes stronger" "Species having formulas with hydrogen are acids and those having formulas with hydroxyl are bases" <p>(p46)</p>
11	USA (Sheppard, 2006)	16 students from three high school chemistry classes.	4 Interview tasks: <ul style="list-style-type: none"> Introductory pH event Neutralization Questions about the models Acid-base titration 	<ul style="list-style-type: none"> "pH measures acidity only" (p36) pH "measured the 'strength' of an acid or base or the amount of acid or base present." (p36) Students had considerable difficulty explaining "how the pH values related to the actual substances in terms of the particles present." (p36) "All indicators changed color at the same pH value and this was invariably at pH 7." (p36) "pH (is) a linear scale." "Acids were inherently more 'powerful' than bases." (p38) "Neutralization as a process of dominance of acids over bases." (p38) "The process of neutralization as the physical mixing of an acid with a base and named no products, drew no equations, and represented

S/N	Country/ Reference	Sample	Instrument	Nature of ACs/Remarks*
				<p><i>the process diagrammatically with unreacted chemical species.</i>" (p37)</p> <ul style="list-style-type: none"> Students' "<i>representations of sub-microscopic events</i>" of neutralization "<i>simply showed base particles attached to acid particles.</i>" (p37) Titration curve: for initial non-changing value of pH, about half of the students explained that "<i>despite the acid having been added, the reaction had not yet started.</i>" One quarter said "<i>no reaction was occurring.</i>" (p40) For the sudden drop in pH value near the endpoint, approximately one third of the students described it as "<i>the reaction suddenly starting to occur.</i>" (p40) Half the students described the leveling of the pH after endpoint "<i>as resulting from an excess of acid particles.</i>" (p40)
12	Taiwan Chiu (2007)	A National Survey of Students' Conceptions of Chemistry in Taiwan This was a 6-year study (Pilot: August 2000 to March 2003.) For the first 2 years of the study, researchers developed two-tier diagnostic tests. After pilot testing, revising, and validating the original tests, the two-tier diagnostic items were arranged into three sets of formal paper-and-pencil tests. A nationwide test was carried out in April 2003.		
	Taiwan (Huang, 2004)	400 Elementary school pupils	As part of a 20 item two-tier MCQ instrument.	<p>Elementary School:</p> <ul style="list-style-type: none"> 44% thought that "<i>soapsuds are neutral because neutral materials do not harm skins and clothes.</i>" 36% thought that "<i>the solution of sodium bicarbonate and acetic acid to be neutral because it is a neutralisation reaction.</i>" 27% thought that "<i>all the acid/bases are toxic—for example, detergent and hydrochloric acid are detrimental to our health.</i>" <p>(.434)</p>
	Taiwan (Chiu, 2004)	430 junior high school students and 240 senior high school student	<p>As part of a 17 item two-tier MCQ instrument.</p> <p>5 Two-tier MCQ (Junior High)</p> <p>8 Two-tier MCQ (Senior High)</p>	<p>Secondary school students:</p> <ul style="list-style-type: none"> 13% of the junior high and 34% of the high school students thought that "<i>weak electrolytes are in a molecular state in a solution. They are in an ionic state after electrification.</i>" "<i>When the test items dealt with sub-microscopic viewpoints of the acid/base, we did not see cognitive development in student performance. For instance, the correct answers for both-tiers were 11% junior high and 9% senior high school students when the test item asked about the distribution of the HCl particles in the water.</i>"

S/N	Country/ Reference	Sample	Instrument	Nature of ACs/Remarks*
				<ul style="list-style-type: none"> 25% of junior high school students thought that “<i>the mixed solution of equal concentrations and volumes of acid (CH₃COOH) and sodium hydroxide (NaOH) is neutral because they react with each other completely.</i>” “<i>Whereas the pressure increases, the degree of dissolution of gas decreases; therefore, the CO₂ is released from water to increase pH value (21% junior and 7% senior); or, the degree of dissolution of gas does not change with pressure (12% junior and 9% senior).</i>” “<i>Weak electrolyte exists as a molecule in water because some molecules decompose to ions, then positive and negative ions attract with each other to combine as molecules again (19% junior and 9% senior); or, weak electrolyte exists as a molecule or ions in water because weak electrolyte can just partially decompose (13% junior and 34% senior).</i>” <p>(pp 434 – 435)</p>
	Taiwan (Hsu, 2004)	512 elementary school pupils, 207 junior high school students and 251 senior high school students	As part of an 18 item instrument on Categorisation of matter’	<ul style="list-style-type: none"> “<i>Acetic acid and water can dissolve because water has the ability to dilute things (36% elementary, 46% junior, and 40% senior).</i>” “<i>Acetic acid and water cannot dissolve because acetic acid decomposes instead of dissolving in water (17% junior and 30% senior).</i>” <p>(p436)</p>
	Taiwan (Juang, 2004)	As part of ‘Material Science’: 288 junior high school students and 256 senior high school students	As part of a ‘Material Science’ test (14 items for junior high school and 19 items for senior high school)	<p>For junior high school students</p> <ul style="list-style-type: none"> 27% thought that “<i>C₂H₅OH is base because it contains OH in its molecular formulas.</i>” 26% thought that “<i>C₂H₅OH is an acid because it has the highest number of hydrogen molecules.</i>” 16% thought that “<i>C₂H₅OH is base because all of the organic compounds are neutral.</i>” <p>(p436)</p> <p>For senior high school students</p> <ul style="list-style-type: none"> 29% thought that “<i>C₂H₅OH to be a base given the functional group (OH).</i>” 10% thought that “<i>ethyl alcohol is acidic because it has the highest number of hydrogen molecules.</i>” 58% understood that “<i>ethyl alcohol is neutral</i>”but attributed this to different reasons. 29% thought that “<i>C₂H₅OH is neutral because all of the organic compounds are neutral.</i>”

S/N	Country/ Reference	Sample	Instrument	Nature of ACs/Remarks*
				(p36)
13	Thailand (Artdej, Ratanaroutai, Coll & Thongpanchan, 2010)	55 Grade 11 students	18 Two-tier MCQ	<ul style="list-style-type: none"> • “CH_3COONa is a non-electrolyte.” (p175) • “NH_4OH is a non-electrolyte.” (p175) • “A strong acid could produce more bubbles upon reaction with metal than a weak acid.” (p175) • “A strong acid cannot dissociate in water, because of strong intramolecular hydrogen bonding.” (p175) • “All bases are ionic compounds.” (p175) • “All bases are corrosive.” (p175) • “CO_3^{2-} and NH_3 and $HCOO^-$ are Brønsted-Lowry acids.” (p176) • “BH_3 is a Lewis base.” (p176) • “H_2SO_4 is the conjugate acid of SO_4^{2-}.” (p176) • “HPO_4^{2-} and NH_3 are a conjugate acid–base pair in the reaction $HPO_4^{2-}(aq) + NH_3(aq) \rightarrow PO_4^{3-}(aq) + NH_4^+(aq)$” (p176) • “The concentration and proton attraction of H_3O^+ influences the strength of acids.” (p177) • “Aqueous solution of H_2SO_4 is more acidic than that of HCl.” (p177) • “A weak acid can dissociate completely in solution, and that, in a weak acid solution, water is a significant contributor to the H^+ concentration of the HF solution.” (p177) • “Concentrations of AOH and BOH are equal and the base dissociation constant (K_b) of AOH is less than BOH. At equilibrium, concentration of BOH is more than, or equal to, the concentration of AOH.” (p178) • For a base BOH (K_b of $BOH = 2.6 \times 10^{-8}$ mol/dm³, concentration of 0.1 mol/dm³), “concentration of OH^- was equal to 1.0×10^{-1} mol/dm³.” (p178) • “Pure water is a good conductor.” (p178) • “K_w is equal to 1.0×10^{-7}.” (p178) • “The addition of an acidic solution to water did not affect the concentration of OH^- or H_3O^+ in the system, because the ionization of water depended on temperature.” (p179) • “When 0.1 mol of KOH was added into 1.0L of water, the solution would be neutral.” (p179)
14	Turkey (Özmen, Demircioğlu & Coll, 2009)	59 high school students	25 MCQ (15 one-tier, 10 two-tier) followed by an intervention study.	<ul style="list-style-type: none"> • “The only way to test a sample whether it is an acid or a base is to see if it eats something away, for example metal, plastic, animal, and us.” • “Acids burn and melt everything.” • “All salts are neutral.” • “Salts don’t have a value of pH 10.”

S/N	Country/ Reference	Sample	Instrument	Nature of ACs/Remarks*
				<ul style="list-style-type: none"> • “All acids and bases are harmful and poisonous.” • “Strong acids can react with all metals to form H_2 gas.” • “Strength of an acid depends on the number of hydrogen atoms in an acid.” • “As the value of pH increases, acidity increases.” • “pH is only a measure of acidity.” • “A strong acid is always a concentrated acid.” • “A strong acid doesn’t dissociate in water solution, because its intra-molecular bonds are very strong.” • “In all neutralization reactions, acid and base consume each other completely.” • “At the end of all neutralization reactions, there is neither H^+ nor OH^- ions in the resulting solutions.” • “As concentration of H_3O^+ in an acid solution increases, pH of the solution increases.” • “After all the neutralization reactions, the pH of solution formed is always 7.” • “All acids and bases conduct electricity the same.” <p>(p14)</p>
15	Turkey (Cetin-Dindar & Geban, 2011)	3 groups of high schools students (12, 111, 156)	Interviews (12) Open-ended questions (111) 18 three-tier MCQs (156)	(Not specified) Study aimed at showing three-tier MCQ is more reliable in identifying alternative conceptions than conventional MCQs
16	Turkey (Saglam, Karaaslan & Ayas, 2011)	Two stages: Stage 1: 106 university students took a questionnaire. Stage 2: 16 students with proper understanding volunteered to be interviewed.	2 rounds of interviews.	<ul style="list-style-type: none"> • Most students “associated acids with red color and bases with blue color.” (1407) • “Formula X is an acid, when HCl is added, the equilibrium shifts right, and similarly because Formula Y is a base, when NaOH is added, the equilibrium shifts left in order to eliminate the stress.” (p1404)
17	USA (McClary & Bretz, 2012)	104 second semester (in 2 sections) college organic chemistry students.	9 Four-tier MCQs (in 3 sets of 3) Confidence scale: 0 – 100%	<ul style="list-style-type: none"> • “Functional group determines acid strength (e.g. “Acetic acid is more acidic than phenol and 2,4-pentadione because it is a carboxylic acid.”) • Stability determines acid strength (e.g. “2,4-pentadione is more acidic than acetone and acetaldehyde because 2,4-pentadione has two carbonyl groups.”)

S/N	Country/ Reference	Sample	Instrument	Nature of ACs/Remarks*
				(p14)

*Note: As far as possible, the phrasings in the original articles were used for most of the ACs.

Appendix A

Confidence ratings and strength of ACs

Consonant with approaches used in psychology (Lundeberg et al., 2000; Stankov & Crawford 1997), the following measures were also computed for each tier of every item, as well as for the entire test, using the confidence ratings provided by the students:

(a) CF : mean confidence rating.

$$\text{CF for a student} = \frac{1}{m} \sum_{i=1}^m c_i$$

$$\text{CF for an item} = \frac{1}{n} \sum_{j=1}^n c_j$$

$$\text{CF for the sample or group} = \frac{1}{n} \sum_{j=1}^n \left(\frac{1}{m} \sum_{i=1}^m c_i \right)_j$$

where m = number of items

n = number of students in the group

c = confidence rating for an item

(b) CFC : mean confidence ratings for an item with a correct response.

(c) CFW : mean confidence ratings for an item with an incorrect response.

(d) CAQ (Confidence Accuracy Quotient) provides an indication of whether the students can distinguish between what they know and what they do not know. It can be interpreted as a student's meta-cognitive ability (Potgieter & Davidowitz, 2011).

For an item,

$$\text{CAQ} = \frac{\text{CFC} - \text{CFW}}{\text{Standard deviation of confidence ratings of item}}$$

A negative CAQ value is an indication of over-confidence.

(e) CB (Confidence Bias)

This is a measure of how accurate an individual was in assessing his or her own performance. Schraw (2009) defined the calculation of confidence bias (termed bias index by Schraw) as

$$\text{Bias Index} = \left(\frac{1}{N} \sum_{i=1}^N c_i \right) - \left(\frac{1}{N} \sum_{i=1}^N p_i \right) = \frac{1}{N} \sum_{i=1}^N c_i - p_i$$

where N = number of items

c_i = confidence rating for an item

p_i = performance score for an item

The CB for an item is calculated as follows (Caleon & Subramaniam, 2010):

$$\text{CB} = \frac{1}{n} \sum_{i=1}^n c_i - p_i$$

where n = number of students in the group

c_i = a student's recoded confidence rating for the item

p_i = a student's score for an item

Since the scores of both-tiers were used as a measurement of performance, for the

1
2
3 whole group, p is equivalent to B , which is the proportion correct for both-tiers. B
4 ranges from 0 to 1. Since the confidence rating is an ordinal scale from 1 to 6, it needs
5
6
7
8 to be recoded to a value range that is same as the student's score p , that is, B .
9

10 This can be achieved by the conversion

$$c = CF - 15$$

11
12 Therefore, the whole expression becomes (Sreenivasulu & Subramaniam, 2013)

$$CB = CF - 15 - B$$

13
14
15
16
17
18
19
20
21 where B is the proportion correct for both-tiers.
22
23
24

25
26 Unbiased responses would generate CB values equal to or close to zero.
27
28
29

30
31 CB has a range from -1.00 to 1.00. It is used and interpreted as an index that reflects
32
33 on the matching / mismatching of the students perceived confidence in their
34
35 performance (in answering the questions) with their actual performance. The
36
37 underlying assumption is that, for self-aware students, their perceived confidence is
38
39 proportional to their actual performance, that is, higher confidence rating translates
40
41 directly into higher performance. Hence, a CB value of zero depicts the 'ideal
42
43 situation' where there is exact calibration between their confidence levels and their
44
45 performance scores. A highly positive CB value would occur when CF is high,
46
47 accompanied by a low B value. This means that the students could feel very confident
48
49 with their answers (and presumably should have scored well) but their actual
50
51 performances were low, that is, they were over confident!
52
53
54
55
56
57
58
59
60

1
2
3
4 For the purpose of identifying ACs, a combination of percentage of answer-reason
5 combination and associated confidence ratings were used to analyze responses for both-tiers.
6
7
8
9 The types of ACs are categorized following the approach of Caleon & Subramaniam (2010):
10
11

- 12
13 (a) Significant alternative conception (SiAC) – answer-response option selected by 10%
14 or higher of the respondents.
15
16
17 (b) Spurious alternative conception (SpAC) – AC with mean confidence rating of 3.5 and
18 below. This could be due to lack of knowledge and/or understanding as well as
19 guessing.
20
21
22
23 (c) Genuine alternative conception (GAC) –significant AC with an associated mean
24 confidence rating above 3.5.
25
26
27
28 (d) Moderate alternative conception (MAC) –genuine AC with an associated mean
29 confidence rating between 3.5 and 4.0.
30
31
32
33 (e) Strong alternative conception (SAC) – AC with an associated mean confidence rating
34 of 4.0 and above.
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Appendix B

Analyses of missing data

The prevalence of missing data is likely to be due to the nature of four-tier MCQs as students have to provide four responses for each question in the test paper, which runs into a number of pages. For a 24 item instrument comprising four-tier MCQs, the total number of responses to be provided by each student would be $24 \times 4 = 96$. It is very likely that some students would have missed out filling in the relevant information for a number of items. While reviewing studies on ACs using multi-tier MCQs, it was noted that very few authors mentioned about missing data. It is likely that samples with missing data were excluded from the analyses (for example, McClary & Bretz, 2012) or that missing data was not an issue. While it is possible to analyze the data based on intact data alone, some compromises would need to be made while interpreting the findings. An appropriate solution would be to get additional students to take the test. However, as it was already near the end of the school calendar year when the data was gathered, new samples could only be obtained in the following year to make up for the numbers. The time factor was thus not in favor.

A quick analysis was done to tabulate the distribution of missing data (Table 2).

Number of missing data	% missing out of maximum	Frequency	% of sample
0	0.00	92	65.25
1	1.09	27	19.15
2	2.17	8	5.67
3	3.26	6	4.26
4	4.35	3	2.13
7	7.61	1	0.71
10	10.87	1	0.71
13	14.13	1	0.71
28	30.43	1	0.71

30	32.61	1	0.71
----	-------	---	------

Table 2 Distribution of missing data

The two scripts with 28 and 30 missing data missing respectively (i.e. about one third of the responses were not provided) had to be discarded due to too many missing information.

An alternative approach of using imputing (Rubin, 1987) to address the issue was also considered – substituting the missing information. This would not work well with missing responses for the answer and reason tiers but may be possible for the confidence ratings. The use of isomorphic questions (there are a few in the instrument) may be able to address this issue to a limited extent in terms of determining the consistency with which students answered similar questions. Isomorphic questions test on the same concept but in different contexts and so the combinations of answer-reason pairs are expected to be somewhat similar to those with missing responses. Even so, responses from isomorphic questions cannot be unequivocally used to substitute data for those with missing responses as there is a possibility that students may perceive the set of isomorphic questions differently since the questions are set in different contexts to test their understanding and their application of the same concepts, and they may end up providing different responses, including confidence ratings.

The main study required detailed analyses of the data ranging from the correctness of the answer tier, reason tier, both tiers and associated confidence ratings to ascertain and confirm the ACs manifested in the students. The validity of substituting missing confidence ratings for the answer and/or reason tiers with other values requires closer examinations. Missing confidence ratings could possibly be addressed in two ways:

- (a) Using the assumption that the student is equally confident in choosing their answer and reason, one can consider using the other available confidence rating as a substitute. For example, if the confidence rating for the reason tier is missing, it can be assumed to be the same as the confidence rating for the answer tier. This, however, does not make sense if either or both of the answer and a reason response is/are missing as well.
- (b) Using the mid-point value of 3.5 as a substitute for the missing confidence rating. However, based on the findings from the pilot phase, the average confidence rating was slightly below the mid-point value. Hence, using 3.5 could have artificially inflated the confidence ratings, especially in cases where students resorted to guessing as they do not know the answer and/or reason.

For the 139 scripts with missing information, the Pearson correlation coefficients between the confidence rating for the answer tier and the confidence rating for the reason tier were computed (Table 3):

Item	1	2	3	4	5	6	7	8	9	
<i>r</i>	0.667	0.626	0.623	0.688	0.587	0.709	0.746	0.663	0.706	
Item	10	11	12	13	14	15	16	17	18	19
<i>r</i>	0.690	0.392	0.592	0.705	0.600	0.753	0.565	0.799	0.572	0.688
Item	21	22	23	25						
<i>r</i>	0.677	0.752	0.677	0.576						

Values were computed according to complete responses for each question

Table 3 Pearson Correlation Coefficients between Answer and Reason Tiers

The overall Pearson correlation coefficient (based on every complete pair of responses to every item) was $r(3116) = 0.68$, $p < 0.001$. All these values suggested positive correlations between the two confidence ratings. However, it was felt that the correlations were not strong enough to justify substituting one missing confidence rating with the corresponding

confidence rating from the other tier. Furthermore, students may also perceive the answer and reason tiers separately as two different multiple-choice questions (Griffard & Wandersee, 2001). To complicate matters further, as the knowledge level measured in each tier can be somewhat different (Tsai & Chou, 2002), students should reasonably indicate different levels of confidence, thus making it unsound to assume that the two tiers are of equivalent levels. This situation was observed in the study on ACs using 4TMC items (Caleon & Subramaniam, 2010), where half of the samples showed different confidence levels for the answer and reason tiers.

The means of the confidence ratings were also worked out for each item (Table 4):

Question	1	2	3	4	5	6	7	8	9	11
Mean CR(A)	3.68	3.99	3.31	3.73	3.66	4.19	3.73	3.60	2.97	4.17
Mean CR(R)	3.34	3.68	3.33	3.62	3.73	3.92	3.72	3.53	3.05	3.93
Question	12	13	14	15	16	17	18	19	20	21
Mean CR(A)	3.71	3.01	3.71	3.77	3.09	3.78	3.72	2.57	3.64	3.71
Mean CR(R)	3.71	2.94	3.36	3.71	2.80	3.65	3.38	2.27	3.41	3.48
Question	22	23	25							
Mean CR(A)	3.37	3.95	3.73							
Mean CR(R)	3.30	3.41	3.45							

Mean CR(A) – Mean confidence rating for the answer tier for an item.

Mean CR(R) – Mean confidence rating for the reason tier for an item.

Values were computed according to complete responses for each question

Table 4 Mean confidence ratings of answer and reason tiers

With respect to the answer tier, more items have mean confidence ratings greater than 3.5.

For the reason tier, fewer items have mean confidence ratings greater than 3. Hence, if the mid-point value of 3.5 is used to substitute the missing confidence ratings, it would result in

(a) deflation of confidence rating for answer tier, and

(b) inflation of confidence rating for reason tier

Item Number	Missing Responses	Item Number	Missing Responses	Item Number	Missing Responses
1	0	12	2	22	4
2	0	13	2	23	8
3	1	14	1	25	7
4	8	15	15		
5	8	16	3		
6	1	17	2		
7	5	18	1		
8	8	19	9		
9	8	20	3		
11	4	21	3		

Table 5 Number of missing responses for each item

There were only 92 scripts without any data omission. For the remaining 47 scripts (more than one third of the population), 21 out of the 23 items were affected. In view of the big number of computations (FI, DI, and confidence ratings, etc) necessary, the analyses would be rendered unnecessarily complicated and excessively time consuming.

After all these considerations, the idea of concentrating the analyses on the 92 'perfect' scripts was explored. To justify this approach, the respective means and standard deviations of the mean cognitive scores of the 139 'total' scripts versus the 92 'perfect' scripts were computed. Imputation for missing items in the cognitive options was done by using the standard approach, whereby items with no answer or reason indicated are scored zero. This was followed by two-tailed t-tests on the mean total scores of these two sets of scripts to see whether there were any statistically significant differences between the two approaches.

	Answer		Reason		Both	
	N=139*	N=92	N=139*	N=92	N=139*	N=92
Mean	30.65	30.15	27.81	26.32	16.80	16.16
Std Dev	11.27	10.41	11.68	10.96	10.68	9.59

<i>t</i>-test	$t(229) = 0.34, p = 0.73$	$t(229) = 0.97, p = 0.33$	$t(229) = 0.46, p = 0.65$
----------------------	---------------------------	---------------------------	---------------------------

* Items with missing answer or reason are scored zero.

Table 6 Comparison of the means, standard deviations and *t*-tests of the ‘Total’ and ‘Perfect’ samples’ cognitive scores

The mean scores and standard deviations of the ‘different sample sizes’ shown in Table 6 are very close to each other for the answer tier, reason tier and both tiers (Table 6). In addition, all the *p*-values are much greater than 0.05, and was quite close to 1 for both-tiers. Hence, there was no statistically significant difference in student performance based on the 139 scripts versus the subset of the 92 scripts; that is, the impact of omitting the scripts with missing information would appear to be minimal as far as the performance of the students was concerned.

A more robust form of analysis was also done to explore whether the data are missing completely at random (MCAR) (Little & Rubin, 1987; Howell, 2007). If data are MCAR, then it is acceptable to use list-wise deletion followed by complete case analysis (Howell, 2007). For this, the distribution of data in the two samples (one with 92 scripts with no missing data and another with 47 scripts with missing data) for one of the variables need to be examined. Two items, (items 1 and 2) were selected for analysis, and in both samples, these items had no missing data. Only the confidence ratings were chosen for analysis because the values for each tier can be assumed to span a continuum (1 to 6). The summated cognitive scores for each question were not used as for each question, the score is dichotomous (0 or 1) in nature, and this prevents the use of parametric statistics for analysis. The results are summarized in Table 7, which presents a similar picture as that shown in Table 4. The mean confidence ratings and standard deviations were reasonably close to each other for both items and all tiers. More importantly, all the *p*-values were again greater than 0.05. It can therefore be concluded that there is no statistically significant differences in

confidence ratings, based on the 47 scripts with missing data and the 92 ‘perfect’ scripts for these two items.

	Answer-tier		Reason-tier		Both-tiers	
Item 1	N=47	N=92	N=47	N=92	N=47	N=92
Mean	3.68	3.66	3.34	3.33	3.51	3.49
Std Dev	1.00	1.12	1.01	1.15	0.93	1.03
t-test	$t(137) = 0.09, p = 0.93$		$t(137) = 0.07, p = 0.94$		$t(137) = 0.09, p = 0.93$	
Item 2	N=47	N=92	N=47	N=92	N=47	N=92
Mean	4.00	4.00	3.57	3.75	3.79	3.88
Std Dev	0.98	1.15	1.14	1.11	0.98	1.01
t-test	$t(137) = 0.00, p = 1.00$		$t(137) = 0.88, p = 0.38$		$t(137) = 0.49, p = 0.62$	

Table 7 Comparison of means, standard deviations and t-tests of ‘missing data’ sample and ‘perfect’ sample’s confidence rating for items with no missing response

In other words, the sample with 92 scripts can be considered to be a random sample of the larger sample with 139 scripts. This further supports the notion that the impact of omitting the scripts with missing data would be minimal. It can therefore be reasonably concluded that the data are indeed predominantly missing completely at random (MCAR).

Appendix C

Qn	DI ^a	Proportion Correct				A tier					R tier					B tiers				
	B	A	R	B	CF	CFC	CFW	CAQ	CB	CF	CFC	CFW	CAQ	CB	CF	CFC	CFW	CAQ	CB	
1	0.43	0.35	0.33	0.27	3.66	3.63	3.68	-0.05	0.26	3.33	3.63	3.18	0.40	0.19	3.49	3.86	3.36	0.49	0.23	
2	0.39	0.24	0.33	0.22	4.00	3.82	4.06	-0.21	0.38	3.75	4.03	3.61	0.38	0.33	3.88	3.93	3.86	0.06	0.36	
3	0.70	0.30	0.40	0.30	3.42	3.75	3.28	0.49	0.18	3.46	3.92	3.15	0.76	0.19	3.44	3.93	3.23	0.81	0.18	
4	0.39	0.21	0.24	0.20	3.67	3.58	3.70	-0.12	0.34	3.62	3.68	3.60	0.08	0.33	3.65	3.75	3.62	0.14	0.33	
5	0.43	0.53	0.59	0.34	3.72	4.02	3.37	0.47	0.21	3.75	4.11	3.24	0.69	0.21	3.73	4.13	3.53	0.51	0.21	
6	0.09	0.23	0.25	0.07	4.25	4.29	4.24	0.04	0.58	3.97	3.65	4.07	-0.35	0.53	4.11	3.58	4.15	-0.52	0.56	
7	0.52	0.61	0.51	0.45	3.83	4.11	3.39	0.52	0.12	3.79	4.17	3.40	0.61	0.11	3.81	4.29	3.42	0.71	0.12	
8	0.26	0.70	0.30	0.20	3.59	3.78	3.14	0.48	0.32	3.49	3.93	3.30	0.48	0.30	3.54	4.11	3.40	0.58	0.31	
9	0.39	0.55	0.33	0.17	2.83	3.04	2.56	0.36	0.19	2.90	3.07	2.82	0.19	0.21	2.86	3.56	2.72	0.70	0.20	
11	0.26	0.36	0.30	0.12	4.26	4.12	4.34	-0.22	0.53	4.01	3.82	4.09	-0.26	0.48	4.14	3.95	4.16	-0.25	0.51	
12	0.26	0.15	0.40	0.12	3.79	3.79	3.79	-0.01	0.44	3.88	4.16	3.69	0.47	0.46	3.84	4.14	3.80	0.37	0.45	
13	0.17	0.35	0.24	0.16	3.08	3.22	3.00	0.15	0.25	2.93	2.95	2.93	0.02	0.22	3.01	3.13	2.98	0.12	0.24	
14	0.09	0.25	0.26	0.14	3.77	3.65	3.81	-0.14	0.41	3.35	3.50	3.29	0.18	0.33	3.56	3.69	3.54	0.16	0.37	
15	0.13	0.15	0.07	0.05	3.78	3.36	3.86	-0.44	0.50	3.78	3.17	3.83	-0.49	0.50	3.78	3.40	3.80	-0.34	0.50	
16	0.09	0.18	0.22	0.13	3.12	3.35	3.07	0.21	0.29	2.68	3.10	2.57	0.39	0.21	2.90	3.29	2.84	0.38	0.25	
17	0.09	0.03	0.08	0.03	3.84	4.33	3.82	0.46	0.53	3.66	3.57	3.67	-0.08	0.50	3.75	4.33	3.73	0.54	0.52	
18	0.00	0.34	0.16	0.11	3.80	3.81	3.80	0.00	0.45	3.39	2.67	3.53	-0.78	0.37	3.60	2.95	3.68	-0.78	0.41	
19	0.17	0.23	0.22	0.12	2.45	2.33	2.48	-0.11	0.17	2.17	2.10	2.19	-0.08	0.12	2.31	2.41	2.30	0.10	0.14	
20	0.57	0.65	0.43	0.36	3.59	3.73	3.31	0.39	0.16	3.36	3.48	3.27	0.19	0.11	3.47	3.56	3.42	0.13	0.14	
21	0.17	0.38	0.20	0.15	3.78	3.66	3.86	-0.19	0.40	3.49	2.83	3.65	-0.67	0.35	3.64	3.00	3.75	-0.73	0.38	
22	0.00	0.13	0.07	0.01	3.42	3.42	3.43	-0.01	0.47	3.26	2.67	3.30	-0.51	0.44	3.34	3.50	3.34	0.14	0.46	
23	0.00	0.01	0.14	0.00	4.07	1.00	4.10	-2.82	0.61	3.51	3.38	3.53	-0.12	0.50	3.79	–	3.79	–	0.56	
25	0.00	0.00	0.00	0.00	3.74	–	3.74	–	0.55	3.42	–	3.42	–	0.48	3.58	–	3.58	–	0.52	
Mean		0.30	0.26	0.16	3.63	3.54	3.56	-0.03	0.36	3.43	3.44	3.36	0.07	0.33	3.53	3.64	3.48	0.16	0.34	
SD		0.20	0.15	0.12	0.43	0.72	0.49	0.68	0.15	0.43	0.56	0.44	0.44	0.14	0.42	0.49	0.44	0.46	0.15	

a : DI computed based on both-tiers (proportion of correct responses from top 25% – proportion of correct responses from bottom 25%)–

b : Items (such as 23 and 25) were omitted from the calculation in cases where no student got the correct answer.

Table 8 Proportion of students who gave correct responses and the values of relevant confidence measures per item of ABCDI (N=92)

Appendix D

Alternative Conceptions		Answer-Reason	% of sample with AC	Mean CFW	Std dev of CR	AC Type
Properties of acids/bases						
AC1	Acids are corrosive (concentration of acid is not considered).	Q1A(d)	15.22	3.36	1.28	SpAC
AC2	Acids are more dangerous than bases because acids are more reactive than bases.	Q3A(a)	35.87	3.30	0.88	SpAC
AC3	Rain water in an unpolluted area is neutral.	Q2D(a)	19.57	4.00	0.77	SAC
		Q2D(c)	32.61	4.23	1.25	SAC
AC4	A compound containing H will produce H^+ , and compound containing OH will produce OH^- .	Q5A(a)	11.96	3.82	1.47	MAC
		Q5B(b)	16.30	3.53	1.06	MAC
AC5	The higher the basicity of an acid, the more basic it is.	Q25A(d)	10.87	3.40	0.97	SpAC
AC6	An acid conducts electricity in solution because of free electrons.	Q8C(b)	11.96	3.64	1.29	MAC
Strength of acids/bases						
AC7	Pure ethanoic acid is a stronger acid than aqueous ethanoic acid.	Q4B(c)	17.39	3.88	1.02	MAC
AC8	An acid that produces a higher H^+ concentration is the stronger acid (initial acid concentration not considered).	Q6C(b)	15.22	4.64	1.15	SAC
		Q6D(b)	10.37	3.70	1.16	MAC
		Q11C(c)	11.96	4.18	0.87	SAC
AC9	An acid with a higher initial concentration is the stronger acid.	Q6C(c)	10.87	4.10	1.37	SAC
AC10	A dibasic acid is a stronger acid than a monobasic acid.	Q6D(a)	22.83	4.71	0.96	SAC
AC11	The higher the initial base concentration, the stronger the base.	Q7D(c)	10.87	3.10	1.20	SpAC

Alternative Conceptions		Answer-Reason	% of sample with AC	Mean CFW	Std dev of CR	AC Type
AC12	The higher the initial number of moles of the base, the stronger the base.	Q7D(d)	14.29	4.38	1.33	SAC
AC13	For two weak acids, the one that produces more H^+ is the stronger acid (without consideration of the fraction dissociated).	Q12A(d)	15.22	3.86	0.95	MAC
		Q12D(c)	14.13	3.69	1.55	MAC
		Q12D(d)	22.83	4.00	0.71	SAC
AC14	A dibasic weak acid produces twice the amount of H^+ as compared to a monobasic weak acid (without consideration of relative strengths).	Q13B(b)	11.96	3.73	1.79	MAC
Indicators						
AC15	At the end point, indicator changes colour as it has no H^+ to react with.	Q14C(a)	43.48	3.98	0.95	MAC
pH						
AC16	pH ranges only from 1 to 14.	Q15B(d)	65.22	4.03	1.07	SAC
AC17	pH of pure distilled water increases slowly with time as it is exposed to air due to self-ionization.	Q16D(b)	25.00	3.39	1.12	SpAC
AC18	Colas have pH of about 5.	Q17B(b)	55.43	4.14	1.00	SAC
AC19	pH of an aqueous acid solution depends on the concentration of the acid only.	Q25B(a)	17.39	3.69	1.01	MAC
AC20	The pH of an aqueous acid solution depends on the strength of the acid only.	Q25D(c)	14.13	4.31	1.32	SAC
AC21	pH is a measure of the total number of moles of H^+ .	Q25D(b)	28.26	3.54	1.03	MAC
AC22	pH of an aqueous acid is independent of its concentration.	Q25D(c)	14.13	4.31	1.32	SAC

Alternative Conceptions		Answer-Reason	% of sample with AC	Mean CFW	Std dev of CR	AC Type
Sub-microscopic views of acids/bases						
AC23	Equimolar amounts of aqueous and pure ethanoic acid produce the same number of moles of H^+ .	Q4A(e)	16.30	4.00	0.65	SAC
AC24	Pure ethanoic acid (in the liquid state) produces more H^+ than aqueous ethanoic acid.	Q4B(d)	19.57	4.06	0.80	SAC
AC25	Pure ethanoic acid (in the liquid state) produces fewer H^+ than aqueous ethanoic acid.	Q4C(b)	15.22	3.36	1.01	SpAC
AC26	An alkaline solution contains no H^+ .	Q18A(a)	13.04	4.08	1.00	SAC
AC27	In an aqueous solution of NaOH, the Na^+ are paired up with the OH^- .	Q18C(a)	20.65	3.79	0.71	MAC
		Q18C(b)	10.87	3.00	0.94	SpAC
AC28	There are more NH_4^+ than NH_3 molecules in aqueous ammonia because NH_3 reacts with water to form NH_4^+ , which are stabilized by the water molecules, causing the equilibrium to shift to form even more NH_4^+ .	Q19C(b)	15.22	2.86	1.35	SpAC
Neutralisation						
AC29	On complete neutralisation between an acid and a base, there is neither H^+ nor OH^- left.	Q21C(a)	41.30	4.00	1.04	SAC
		Q21D(a)	14.13	4.00	0.91	MAC
		Q22C(a)	11.96	3.64	0.67	MAC
		Q23A(a)	11.96	3.73	0.65	MAC
		Q23B(a)	39.13	4.19	0.89	SAC
		Q23B(d)	18.48	4.41	1.00	SAC
AC30	On complete neutralisation between an acid and a base, the cations and anions of the salt formed are paired up.	Q23B(a)	39.13	4.19	0.89	SAC
		Q23B(d)	18.48	4.41	1.00	SAC

Table 10 List of alternative conceptions identified in main study