4.1 Case study 1 (CS1 Greece)

Concept focus	Features of acids and bases pH and indicators Understanding salts
Activities implemented	Activities A-G
Inquiry skills	Planning investigations (carrying out investigations) Developing hypotheses Forming coherent arguments Working collaboratively (teamwork)
Scientific reasoning and literacy	Scientific reasoning (drawing conclusions) Scientific literacy (everyday applications of acids and bases)
Assessment methods	Classroom dialogue Teacher observation Peer-assessment Self-assessment Worksheets Student devised materials (pH scale) Other assessment items (post-implementation test)
Student group	**Grade:** lower second level **Age:** 12 years **Group composition:** co-ed (16 girls, 7 boys), groups of 4-5 **Prior experience with inquiry:** None

This case study describes full implementation of the **Acids, bases, salts** SAILS inquiry and assessment unit. Each of the inquiry skills and competencies identified in the SAILS project were assessed, through a combination of classroom dialogue, peer-assessment, self-assessment, evaluation of student artefacts and a post-implementation test. The teacher prepared student rubrics for peer-assessment, and a self-assessment tool for evaluation of *developing hypotheses*.

(i) How was the learning sequence adapted?

The **Acids, bases and salts** SAILS unit was implemented in full; the learning sequence followed the steps described in the unit with no modifications. The steps of the learning sequence are outlined below.

Phase I: Formulating hypotheses

In the initial activity, students worked in groups to taste, smell, and make other general observations about six different samples (vinegar, lemon juice, orange juice, yoghurt, baking soda dissolved in water and toothpaste dissolved in water). Their observations were recorded in the worksheet. Then there was a discussion with the entire classroom guided by the teacher where students narrated possible previous experiences with the aforementioned substances, and also the knowledge of their scientific names was testified. Subsequently, each group wrote down a short composition in the worksheet that concerns their working hypotheses about: (i) which substances are similar (to each other), (ii) what are the common characteristics (among them) and (iii) if they are aware of any other substances which have similarities with those provided.

Phase II: Experimental

Students carried out the following scientific experiments in groups to test their hypotheses (they used a Labdisc instrument to make numerical measurements of the pH):

- Experiment 1: Qualitative classification of the substances using pH indicator extracted from red cabbage.
- Experiment 2: Students measured the numerical value of the pH of each sample substance using Labdisc. They then constructed a pH scale – each group used the recordings from experiments 1&2 in order to construct their own pH scale, calibrated both in numbers and colour with the help of the worksheets provided.
- Experiment 3: Students ascertain the existence of salts by using the red cabbage indicator.
- Experiment 4: Students now find out how salts can be identified through their property of dissolution in the presence of acids (acids dissolve salts, bases dissolve fats).

Phase III: Conclusions

Groups summarised their observations and results from their worksheets through classroom dialogue. In a new worksheet they wrote down: (i) the classification of substances in three major groups: acids, bases and salts, (ii) matching of substances to the constructed pH scale, and (iii) the property of acids to dissolve salts and the property of bases to dissolve fats. They then returned to their initial hypotheses, to make corrections with the help of the respective worksheets. The teacher worked supportively with each group, solving inquiries or disagreements.

Phase IV: Everyday applications

In this phase each group answered questions related to acids and bases in everyday life, using their worksheets and textbooks to support their answers. The answers were peer-assessed using a holistic rubric, which evaluated the accuracy and completeness of students' answers. The teacher provided the rubric in advance, and explained the assessment criteria and the weight factor of each criterion. Then there was a whole-class discussion, guided by the teacher, to facilitate the final correction of the answers. Each group put a final score on the worksheet and the peer-assessment finished. Afterwards, each student had to take an individual test. The teacher assessed the folders with the worksheets and peer-assessment forms for each group.

(ii) Which skills were to be assessed?

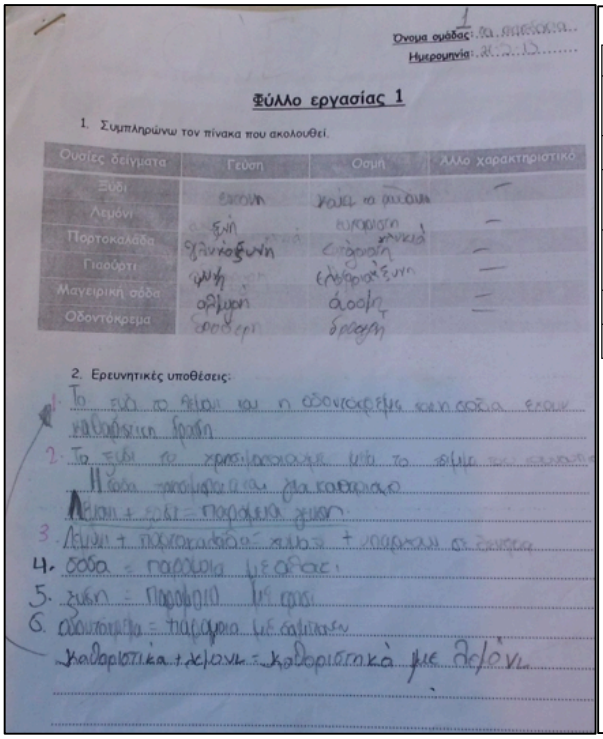
The following skills were assessed in this case study: *developing hypotheses*, *planning investigations* (carrying out an investigation), *forming coherent arguments* and *working collaboratively* (teamwork). Assessment opportunities included teacher observation and feedback, evaluation of student artefacts using rubrics, peer-assessment and self-assessment.

Developing hypotheses

During Phase I the groups were asked to write down their working hypotheses concerning: (i) which substances are similar (to each other), (ii) what are the common characteristics (among them), and (iii) if they are aware of any other substances which have similarities with those provided. Firstly, the groups tasted 6 different samples (vinegar, lemon juice, orange juice, yogurt, baking soda dissolved in water and toothpaste dissolved in water) in order to record the taste feeling, the smell, and other general observations for each of the substances in a worksheet. All groups showed that vinegar and lemon have a strong taste and smell while one group introduced the concept of causticity of acids because it observed that vinegar is so strong that "burns the nostrils." A similar observation made by 3 of the 4 groups for toothpaste, noting that the flavour is spicy.

The working hypothesis process did not produce satisfactory results. The students were not able to detect that the samples contain a common constituent to which they owe their common characteristic. Yet, students' observations had sufficient precision in ways of using these substances and they tried to make a first form of grouping. They observed that (a) lemon, toothpaste and soda have cleaning action, and (b) lemon and vinegar have similar taste (Figure 1). In addition 4 of the 5

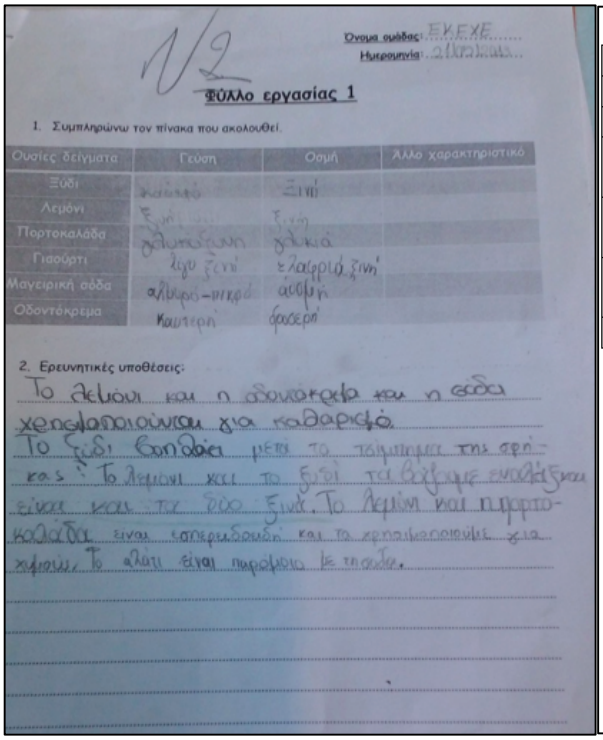
groups showed that soda is similar to salt, while two groups noticed that we use vinegar for itchy insect bites, etc. (Figure 1). However, the total assessment of the worksheets showed that students cannot autonomously detect regularities to these substances and therefore they cannot proceed towards experimental design without the leading role of the teacher.



Fill in the table

Samples	Taste	Smell	Other
Vinegar	*Sharp*	*Burns the nostrils*	-
Lemon juice	*Sour*	*Pleasant*	-
Orange juice	*Sweet and sour*	*Sweet and pleasant*	-
Yoghurt	*Sour*	*Lightweight sour*	-
Baking soda	*Salty*	*Odourless*	-
Tooth paste	*Refreshing*	*Refreshing*	-

Working hypotheses:
1. Vinegar, lemon, toothpaste and soda have cleaning ability. Cleaners + lemon = cleaners with lemon.
2. We can use vinegar after a mosquito bite.
Soda is used for cleaning.
Lemon and vinegar have a similar taste.
3. Lemon and orange juice: available bulk or on trees
4. Soda: similar with salt
5. Vinegar: similar with wine
6. Toothpaste: similar to shampoo



Fill in the table

Samples	Taste	Smell	Other
Vinegar	*Spicy*	*Sour*	-
Lemon juice	*Sour*	*Sour*	-
Orange juice	*Sweet and sour*	*Sweet*	-
Yoghurt	*Slightly sour*	*Light sour*	-
Baking soda	*Salty - bitter*	*Odourless*	-
Tooth paste	*Spicy*	*Refreshing*	-

Working hypotheses:
Lemon and toothpaste and soda are used for cleaning. Vinegar helps after a sphinx bite. We use by turns lemon and vinegar and both are sour. Lemon and orangeade are citrus and we use them for making juices.

Figure 1: Example of student working showing hypotheses developed

In phase III (conclusions) the groups summarised through discussion their observations and records from the previous worksheets. The groups went back to their initial hypotheses in order to make

corrections with the help of the respective working sheet (Figure 2). The instructor worked supportively with each group, solving inquiries or dissents that could arise.

Hypothesis no. ___

The mistake was:

.....

.....

.....

The final conclusion is:

.....

.....

.....

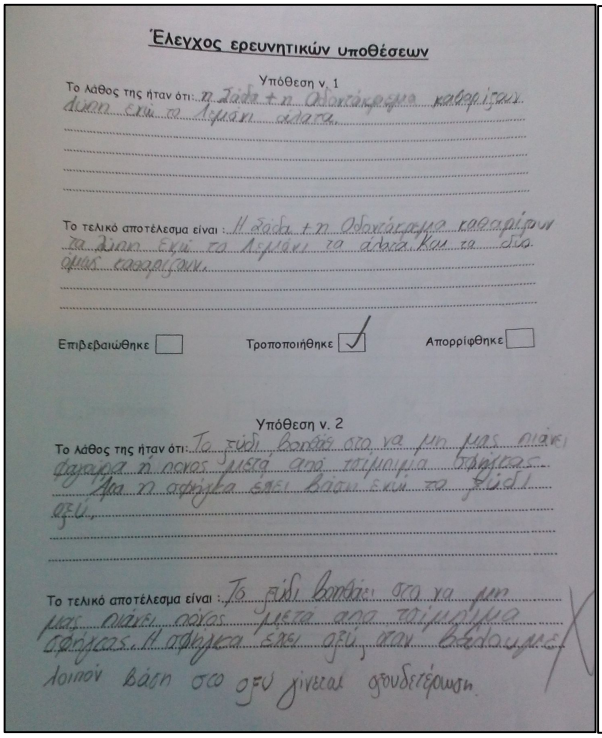
Verified ☐ Modified ☐ Rejected ☐

Figure 2: Self-assessment of hypothesis formulation

All groups completed the first hypothesis, which was an investigation of the cleaning effect of the substances that they tested during the initial activity. All groups commented that "acids dissolve salts, bases dissolve fats." Additionally, three groups that had recorded in their hypothesis that vinegar is good for itchy bites confirmed their assumptions. One of the groups completed the additional information about the mechanism of neutralisation.

Regarding the observation that lemon and vinegar have similar taste, all groups confirmed this hypothesis, including two teams who described that this is because they are both acids. Three groups rejected the hypothesis that milk, yoghurt and butter look alike, which they had initially supported. These groups justified the modification of their hypotheses saying that yoghurt is different because it is acidic, while one group identified the type of acid (lactic acid).

The initial hypothesis that the soda and salt look alike was also rejected, with students noting that the two substances belong to different groups (bases, salts). Finally a group rejected the initial assumption of similarity of floor cleaners with the lemon juice (because of its smell), noting that one belongs to bases while the other is acidic. Also, the same group confirmed the hypothesis that the toothpaste and shampoo look alike because both contain base.



Hypothesis testing

Hypothesis no. ___

The mistake was:

Soda and toothpaste are used for cleaning fats whereas lemon is used for cleaning salts.

The final conclusion is:

Soda and toothpaste are used for cleaning fats whereas lemon is used for cleaning salts. But both of them are cleaners.

Verified ☐ Modified ☑ Rejected ☐

Hypothesis no. ___

The mistake was:

Vinegar helps us to not itching after a sphinx bite. So sphinx has base and vinegar acid.

The final conclusion is:

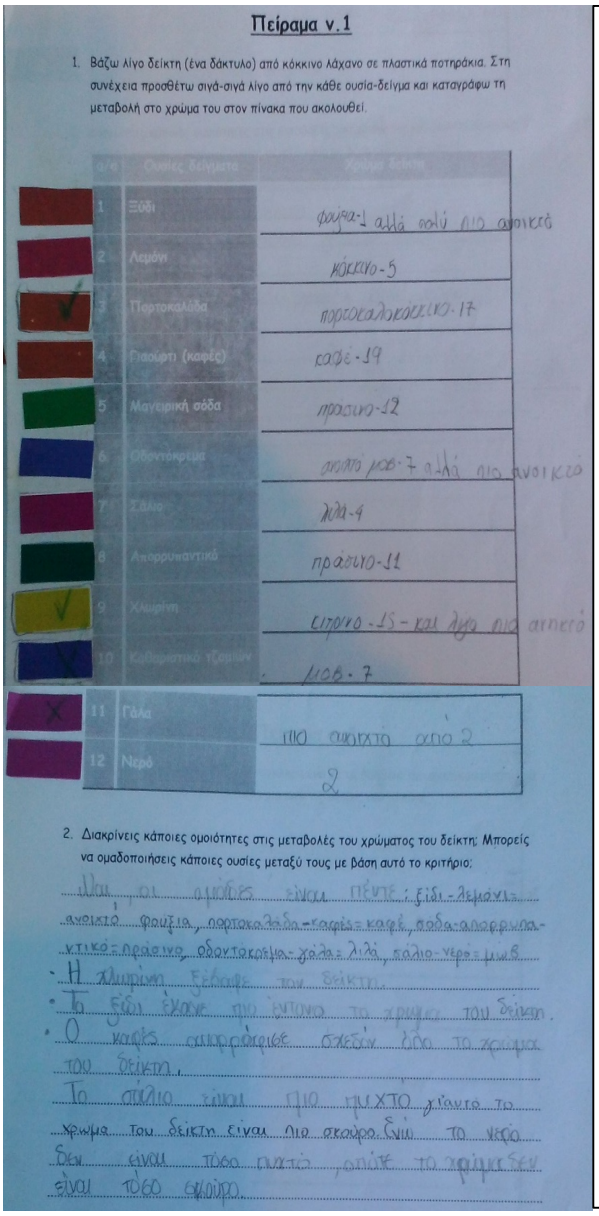
Vinegar helps us not to get pain after a sphinx bite. So sphinx has acid and when we put base on the acid neutralization is happened.

Verified ☐ Modified ☑ Rejected ☐

Planning investigations (carrying out an investigation)

During phase II (experimental), students conducted pre-designed experiments to identify common characteristics among the substances in order to formulate a first grouping form of the underlying substances. In the first experiment (qualitative classification of the substances using pH indicator extracted from red cabbage) the records of the groups were fairly accurate. Only two groups had 2 and 3 respectively mistakes (wrong shade selection). These errors were corrected by the intervention of the teacher. The groups showed variations in the way they formulated their observations (2 groups chose to record their observations in bullets, one group used a free text, another group used a combination of bullets and free text, while the last group wrote word pairs reflecting the identified grouping e.g. lemon-vinegar, soda-detergent). The groups recorded the following classifications: (a) similarity between soda and detergent (3 out of 4 groups), (b) similarity between lemon and vinegar (all 4 teams). Three out of four groups separately mentioned bleach liquor, which was considered unique and it was not grouped with any other substance. One group made a systematic recording with five groupings of substances, where each group corresponds to a specific shade indicator (Figure 3 and Figure 4).

In the second experiment, the students measured the numerical value of the pH of each sample using Labdisc's sensor. The recordings were precise because of the Labdisc. Four of the five groups chose to record to two decimal digits and one group used one decimal digit by applying rounding. The decision wasn't based on some reasoning. Then the groups proceeded to categorise substances on the basis of their results. The categories made by groups ranged from 2 (most poor) to 5 (the best performance). All groups identified the similarity between the lemon and the vinegar as well as between water, saliva and toothpaste. It was a surprise that only two groups identified the similarity between the detergent and bleach.



Experiment no. 2

1. Use the Labdisc or pH metre to measure the numerical value of pH for each of the samples

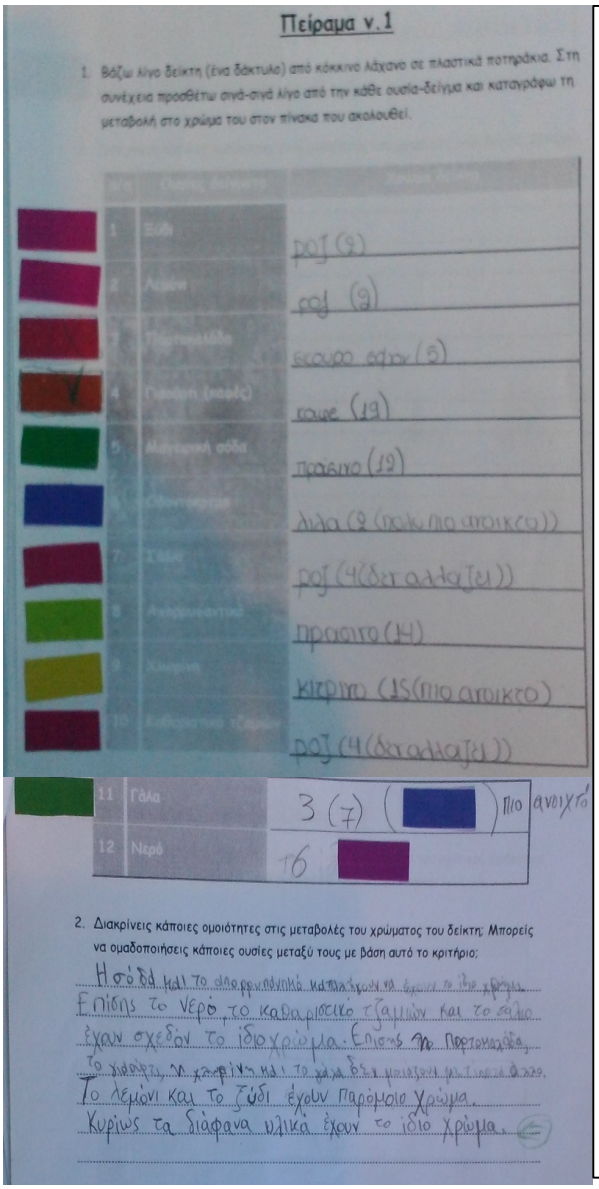
	Substance	pH value
1	Vinegar	***magenta but much lighter***
2	Lemon juice	***red-5***
3	Orange juice	***orange-red-17***
4	Yoghurt (coffee)	***brown-19***
5	Baking soda	***green-12***
6	Toothpaste	***light purple-7 but much more lighter***
7	Saliva	***lilac-4***
8	Detergent	***green-11***
9	Chlorine (cleaning bleach)	***yellow-15 but more lighter***
10	Detergent for windows	***purple-7***
11	Milk	***more lighter than 2***
12	Water	***2***

2. Can you detect any similarities concerning the pH value for these substances? Can you make a primary classification of the samples relying on the pH value of each one?

Yes ,the teams are five: vinegar-lemon; open fuchsia, orangeade-coffee; brown, soda-detergent; green, toothpaste-milk; lilac, saliva-water; purple

- ***Bleach discoloured the indicator.***
- ***Vinegar made the colour of the indicator more vivid.***
- ***Coffee took-up almost all the colour of the indicator. Saliva is thicker and for that the colour of the indicator is darker. Water isn't so thick and then indicator's colour isn't so dark.***

Figure 3: Example of student responses to experiments 2 and 3 (Phase II: Experimental)



Experiment no. 2

1. Use the Labdisc or pH metre to measure the numerical value of pH for each of the samples

	Substance	pH value
1	Vinegar	***pink (2)***
2	Lemon juice	***pink (2)***
3	Orange juice	***dark salmon (5)***
4	Yoghurt (coffee)	***brown (19)***
5	Baking soda	***green (12)***
6	Toothpaste	***lilac (2) much more lighter***
7	Saliva	***pink (4) (it doesn't change)***
8	Detergent	***green (14)***
9	Chlorine (cleaning bleach)	***yellow (15 more lighter)***
10	Detergent for windows	***pink (4 (it doesn't change)***
11	Milk	***3(7) more lighter***
12	Water	***6***

2. Can you detect any similarities concerning the pH value for these substances? Can you make a primary classification of the samples relying on the pH value of each one?

Soda and detergent, in the end have the same colour. Water, saliva and glass cleaner have nearly the same colour. Furthermore orangeade, yoghurt, bleach and milk are not similar to nothing else. Lemon and vinegar have similar colour. Mainly transparent materials have the same colour.

Figure 4: Example of student responses to experiments 2 and 3 (Phase II: Experimental)

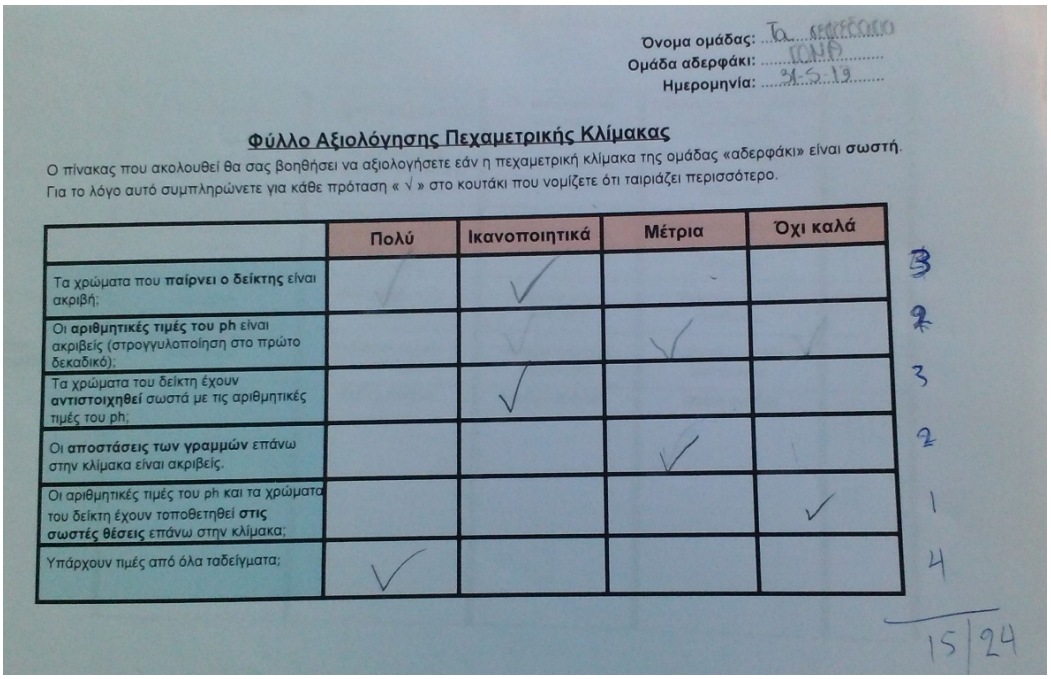
Assessment of the pH scale (peer assessment using a rubric)

The students used a rubric to assess the construction of the pH scale. Using Rubric 1 (Table 1), the groups mostly scored relatively high (3.3-3.5/4) with the exception of one group, which showed significantly lower performance than the rest (2.8/4). The groups showed relative harmony in their scores since the score attribute is similar. For example the average scores vary from 0.4 to 0.7 while the standard deviations of the scores vary from 0 to 0.4 indicating little dispersion in their choices. The criteria assessed with greater accuracy are those concerning the completeness of the samples used (Question: No sample is missing from the scale) and the accuracy of numerical measurements (Question: Numerical values of pH are precise). The criterion that shows the largest discrepancy in valuation is the distances of the lines on the scale (Question: Numerical calibration of the pH scale is precise). Here the groups gave different scores (scores' deviation 1.2 degrees). The groups did not experience difficulties understanding the criteria, but rather in the quantification of the 1st and 5th criterion (The colour of the indicator for each sample is precise - numerical values as well colour values have been placed in the right spots across the scale). The groups took the help of the teacher in order to be decided what is the appropriate score. For the first criterion the difficulty was due to the fact that different groups came through their experiments in slightly different shades indicators

for the same substances and the small differences in the proportions used. These small variations were interpreted initially by the students as errors and so it was needed the intervention of the teacher in order to put the correct score. For the 5th criterion the problem was that groups found difficulties in defining the correlation between the scores of the rubric and the errors of the pH scales. Examples of student artefacts are shown in Figure 6, Figure 7 and Figure 8.

Table 1: Rubric 1 – used to evaluate correctness of constructed pH scale

	Excellent (4)	Good (3)	Needs improvement (2)	Unacceptable (1)
1. The colour of the indicator for each sample is precise	All measurements are correct/No mistakes at all	Some mistakes/Most measurements are correct	Several mistakes/Some measurements are correct/It can be improved	A lot of mistakes/It needs a lot of work to be improved
2. Numerical values of pH are precise	All numerical values are precise/No mistakes at all	Some mistakes/Most numerical values are precise	Several mistakes/Some numerical values are precise/It can be improved	A lot of mistakes/It needs a lot of work to be improved
3. Each colour is matched with the right numerical value of the pH	All colours are matched with the right numerical value/No mistakes at all	Some mistakes/Most colours are matched with the right numerical value	Several mistakes/Some colours are matched with the right numerical value/It can be improved	A lot of mistakes/It needs a lot of work to be improved
4. Numerical calibration of the pH scale is precise	All numerical values are precise/No mistakes at all	Some mistakes/Most numerical values are precise	Several mistakes/Some numerical values are precise/It can be improved	A lot of mistakes/It needs a lot of work to be improved
5. Numerical values as well colour values have been placed in the right spots across the scale	All numerical values and colour values have been placed in the right spots/No mistakes at all	Some mistakes/Most numerical values and colour values have been placed in the right spots	Several mistakes/Some numerical values and colour values have been placed in the right spots/It can be improved	A lot of mistakes/It needs a lot of work to be improved
6. No sample is missing from the scale	No sample is missing from the scale/No mistakes at all	Some mistakes/Most samples are present in the scale	Several mistakes/Some samples are present in the scale/It can be improved	A lot of mistakes/It needs a lot of work to be improved



Όνομα ομάδας: Τα κόκκινα
Ομάδα αδερφάκι: τομάτα
Ημερομηνία: 31-5-19

Φύλλο Αξιολόγησης Πεχαμετρικής Κλίμακας

Ο πίνακας που ακολουθεί θα σας βοηθήσει να αξιολογήσετε εάν η πεχαμετρική κλίμακα της ομάδας «αδερφάκι» είναι **σωστή**. Για το λόγο αυτό συμπληρώνετε για κάθε πρόταση « √ » στο κουτάκι που νομίζετε ότι ταιριάζει περισσότερο.

	Πολύ	Ικανοποιητικά	Μέτρια	Όχι καλά	
Τα χρώματα που **παίρνει ο δείκτης** είναι ακριβή;		√			3
Οι **αριθμητικές τιμές του ph** είναι ακριβείς (στρογγυλοποίηση στο πρώτο δεκαδικό);			√		2
Τα χρώματα του δείκτη έχουν **αντιστοιχηθεί** σωστά με τις αριθμητικές τιμές του ph;		√			3
Οι **αποστάσεις των γραμμών** επάνω στην κλίμακα είναι ακριβείς.			√		2
Οι αριθμητικές τιμές του ph και τα χρώματα του δείκτη έχουν τοποθετηθεί **στις σωστές θέσεις** επάνω στην κλίμακα;				√	1
Υπάρχουν τιμές από όλα ταδείγματα;	√				4

15 | 24

Figure 5: Example of peer-assessment of correctness of constructed pH scale, using rubric 1

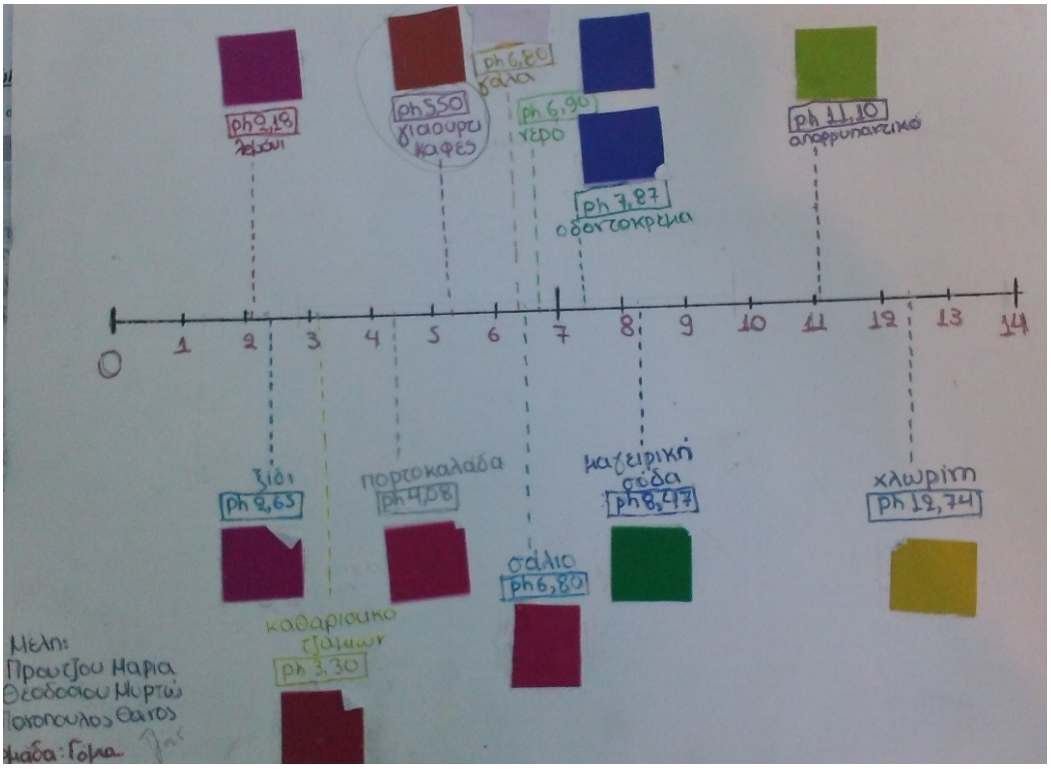


Figure 6: Student artefact for construction of pH scale (example 1)

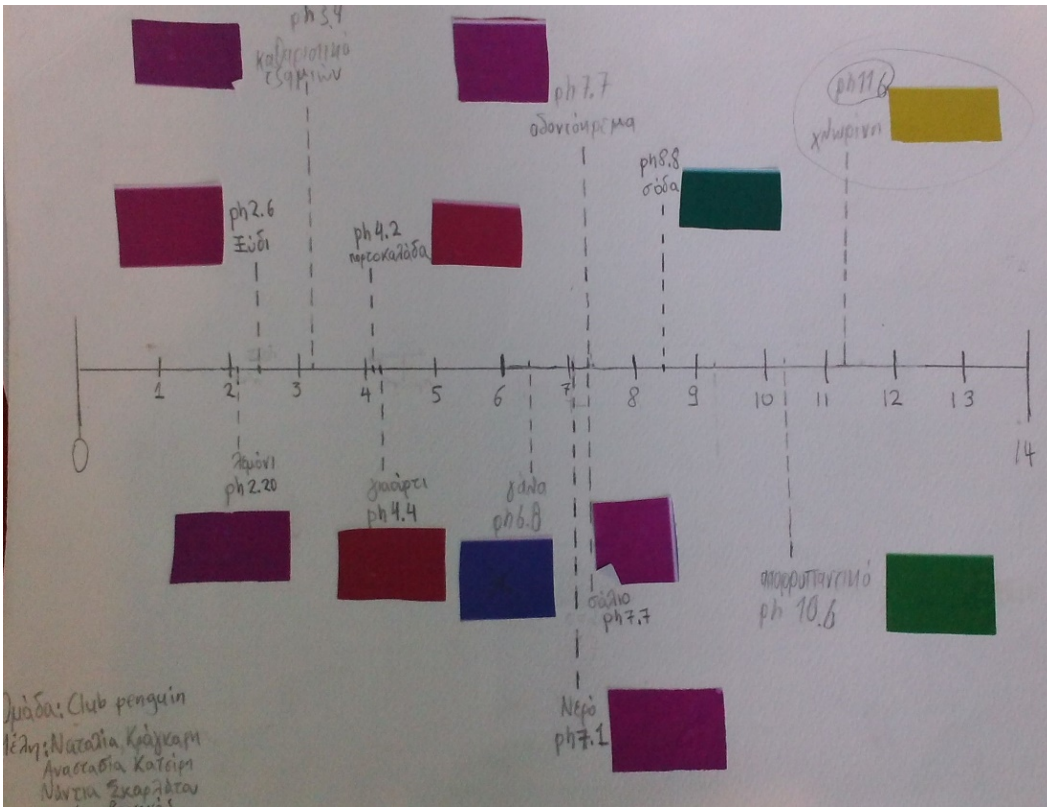


Figure 7: Student artefact for construction of pH scale (example 2)

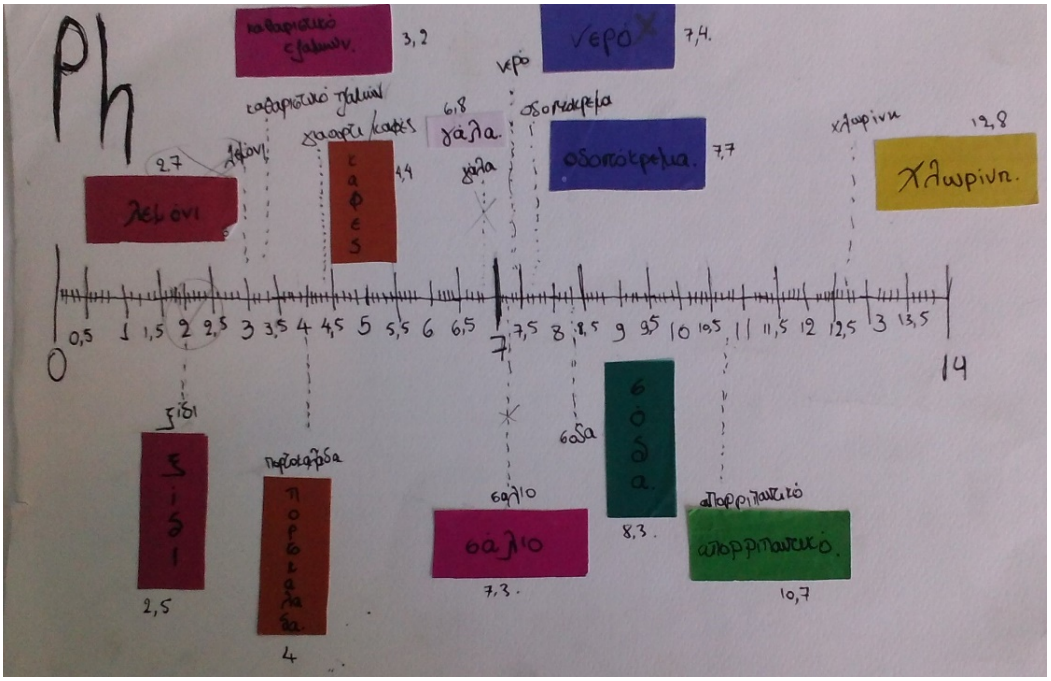


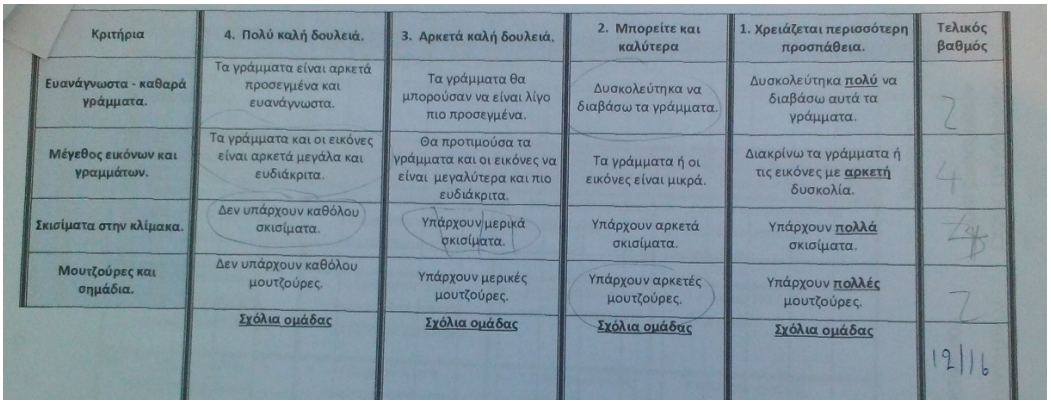
Figure 8: Student artefact for construction of pH scale (example 3)

A second peer-assessment tool, Rubric 2 (Table 2), was used to assess the appearance of the constructed pH scales. A first quantitative assessment of the results shows that the groups obtained relatively high score (3.1 to 3.8 out of 4). The average scores vary by 0.2 to 0.5 while the standard deviations of the scores vary from 0.1-0.7 indicating slightly greater dispersion in their choices than the assessment of correctness (Rubric 1, Table 1). The criterion with the greatest assessed accuracy is "Readability". In this question all the sibling groups gave the same score. The criterion with the

largest discrepancy in valuation was the "Size of icons and fonts". Here the sibling groups gave different scores (score deviation 0.6 degrees). The groups did not face difficulties in understanding the criteria and their quantification.

Table 2: Rubric 2 – used to evaluate appearance of constructed pH scale

	Excellent (4)	Good (3)	Needs improvement (2)	Unacceptable (1)
1. Readability	No difficulty reading at all	There were 2-3 spots that I found difficult to read	There were 4-5 spots that I found difficult to read	It was hard for me to read the text. There were over 6 spots that I found difficult to read
2. Size of icons and fonts	Icons and fonts are clear	Icons and fonts could be clearer (2-3 bad spots)	Icons and fonts should be clearer (4-5 bad spots)	Icons and fonts are so small that I hardly recognise them (>6 bad spots)
3. Rips on the pH scale	No rips	1-2 rips	3-4 rips	5-6 rips
4. Smudges and spots on the pH scale	0-1 smudges	2-3 smudges	4-5 smudges	6-7 smudges



Κριτήρια	4. Πολύ καλή δουλειά.	3. Αρκετά καλή δουλειά.	2. Μπορείτε και καλύτερα	1. Χρειάζεται περισσότερη προσπάθεια.	Τελικός βαθμός
Ευανάγνωστα - καθαρά γράμματα.	Τα γράμματα είναι αρκετά προσεγμένα και ευανάγνωστα.	Τα γράμματα θα μπορούσαν να είναι λίγο πιο προσεγμένα.	Δυσκολεύτηκα να διαβάσω τα γράμματα.	Δυσκολεύτηκα πολύ να διαβάσω αυτά τα γράμματα.	2
Μέγεθος εικόνων και γραμμάτων.	Τα γράμματα και οι εικόνες είναι αρκετά μεγάλα και ευδιάκριτα.	Θα προτιμούσα τα γράμματα και οι εικόνες να είναι μεγαλύτερα και πιο ευδιάκριτα.	Τα γράμματα ή οι εικόνες είναι μικρά.	Διακρίνω τα γράμματα ή τις εικόνες με αρκετή δυσκολία.	4
Σκισίματα στην κλίμακα.	Δεν υπάρχουν καθόλου σκισίματα.	Υπάρχουν μερικά σκισίματα.	Υπάρχουν αρκετά σκισίματα.	Υπάρχουν πολλά σκισίματα.	4
Μουτζούρες και σημάδια.	Δεν υπάρχουν καθόλου μουτζούρες.	Υπάρχουν μερικές μουτζούρες.	Υπάρχουν αρκετές μουτζούρες.	Υπάρχουν πολλές μουτζούρες.	2
	Σχόλια ομάδας	Σχόλια ομάδας	Σχόλια ομάδας	Σχόλια ομάδας	12/16

Figure 9: Example of peer-assessment of appearance of constructed pH scale, using rubric 2

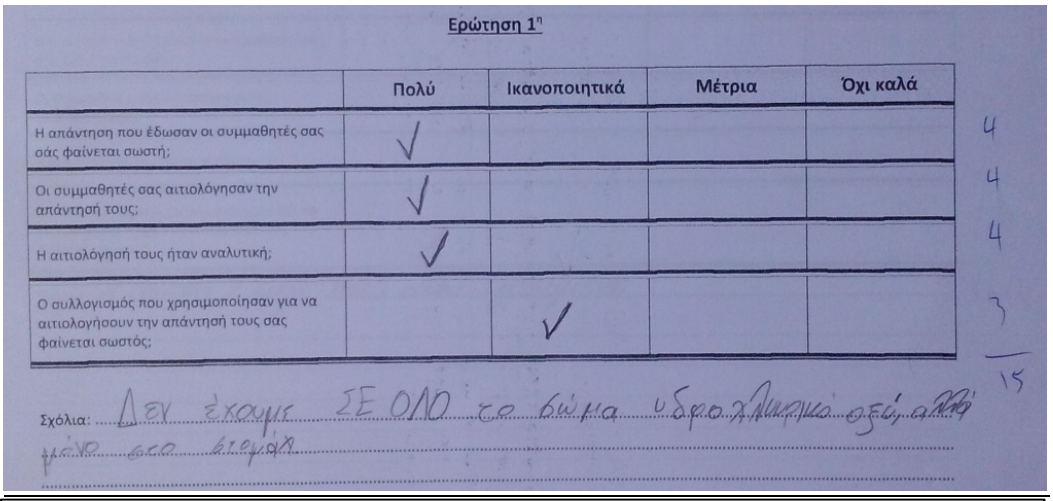
Forming coherent arguments

During the last phase (Phase IV: Everyday applications) each group had to answer some questions regarding acids and bases in everyday life. Some of these are: (i) what happens in milk so that it becomes yoghurt, (ii) what happens in wine so that it is altered to vinegar, and (iii) why does black tea change colour when lemon is added? These questions also revise earlier knowledge gained by the students in physics courses. In order to answer, groups were free to rely on their worksheets as well as textbooks. Peer-assessment was used to evaluate the responses, using a holistic rubric that assesses the accuracy and completeness of students' answers (Table 3). The rubric was discussed in advance with the groups in order to facilitate the final correction of the answers. Each group then put a final score on the worksheet and the peer-assessment finished.

Table 3: Rubric for peer-assessment of answers to worksheet for activity 7

	Excellent (4)	Good (3)	Needs Improvements (2)	Unacceptable (1)
Does the answer seem right?	All points seem right/No mistakes at all	Some mistakes/Most points seem right	Several mistakes/Some points seem right/It can be improved	The answer is unacceptable
Do they use arguments in order to convince you?	All arguments convinced me/No mistakes at all	Some mistakes/Most arguments convinced me	Several mistakes/Some arguments convinced me/It can be improved	The arguments are unacceptable
Is the argumentation being used complete?	The argumentation is complete/No mistakes at all	Some mistakes/Most arguments are complete	Several mistakes/Some arguments are complete/It can be improved	The argumentation is unacceptable
Does the argumentation being used feel right?	All points seem right/No mistakes at all	Some mistakes/Most points seem right	Several mistakes/Some points seem right/It can be improved	The answer is unacceptable

Figure 10: Students engaged in the unit activities



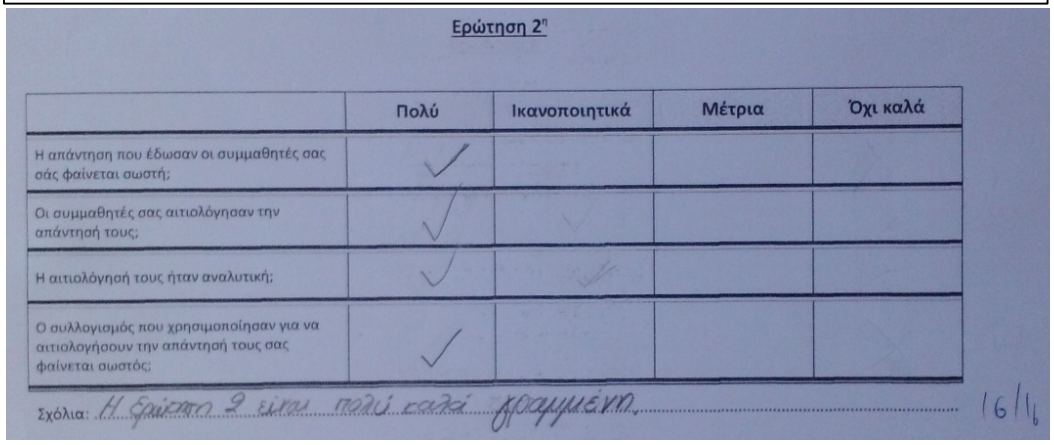
Ερώτηση 1η

	Πολύ	Ικανοποιητικά	Μέτρια	Όχι καλά
Η απάντηση που έδωσαν οι συμμαθητές σας σάς φαίνεται σωστή;	√			
Οι συμμαθητές σας αιτιολόγησαν την απάντησή τους;	√			
Η αιτιολόγησή τους ήταν αναλυτική;	√			
Ο συλλογισμός που χρησιμοποίησαν για να αιτιολογήσουν την απάντησή τους σας φαίνεται σωστός;		√		

4
4
4
3
15

Σχόλια: Δεν έχουμε σε όλο το σώμα υδροχλωρικό οξύ αλλά μόνο στο στομάχι

Comments: ***We don't have hydrochloric acid throughout the body but only in the stomach***



Ερώτηση 2η

	Πολύ	Ικανοποιητικά	Μέτρια	Όχι καλά
Η απάντηση που έδωσαν οι συμμαθητές σας σάς φαίνεται σωστή;	√			
Οι συμμαθητές σας αιτιολόγησαν την απάντησή τους;	√			
Η αιτιολόγησή τους ήταν αναλυτική;	√			
Ο συλλογισμός που χρησιμοποίησαν για να αιτιολογήσουν την απάντησή τους σας φαίνεται σωστός;	√			

Σχόλια: Η ερώτηση 2 είναι πολύ καλά γραμμένη.

16/16

Comments: ***Question number 2 is very well written***

Figure 11: Examples of peer-assessment of students' answers

Working collaboratively (teamwork)

There was no specific tool utilised to measure this skill. However, the teacher gave an extended report on how the students collaborated during the whole learning sequence. Students worked in groups of 4-5 people (Figure 10). They were familiar with this process because they work daily in the same way. In each group there were distinct roles, which were those of the registrar, deputy registrar/announcer and scientists (who carried out the experiments). These roles were changed cyclically so that all team members participated in each of the roles. The quality of collaboration between groups fluctuated during the activities. During phase I (Introduction – developing hypotheses), all members participated with enthusiasm even the weaker students because they felt safe to engage in something that was not scored and where there was no right or wrong answer. A positive attitude was observed in all groups. All students tested the substances of the samples and reported various characteristics of these substances, and they contributed equally with their knowledge of how to use these substances in everyday life.

During phase II: Experimental, the students continued to maintain a high degree of cooperation, with some variations. It is worth noting here the attitude of one of the groups in which participated a student with problems related to communication, concentration and motor coordination. The other members tried to help their classmate to participate in the process (both to perform experiments and to record the results).

During the construction of the pH scale, two groups had two different approaches (calibration per centimetre or per millimetre calibration for greater accuracy). Both approaches were discussed. One group adopted the most accurate approach but at the implementation there were several inaccuracies while the other group decided the other approach as it would require less work. A problem raised in this phase when the groups began to assess the construction of the PH scale of the sibling group. Some variations in scores and the identification of some errors and omissions brought some problems between the members of the groups because they were trying to identify the peer responsible of each error/omission. In other experimental activities collaboration evolved smoothly.

During the last phase (phase IV: Everyday applications), the smoothness of collaboration exhibited strong fluctuations. The participation of the weakest students decreased, as this activity was very demanding. The members discussed their answers by exchanging arguments. In these discussions tensions emerged when opposing arguments continued to collide without finding a solution. The teacher reported that in this case the students lacked the flexibility to return to the material and recordings to unravel the subject of dispute, but both sides continued to insist on relying on what they had memorised. A typical example was for the question ***"Can you guess some of the ingredients contained in ointment proper for bee/wasp sting?"*** The group members stubbornly insisted on their initial choice, instead of seeking the information they lacked.

In conclusion, the collaboration of groups was satisfactory. The evaluation process of peer groups seemed to cause reactions because it sparked some competition between students, reducing their positive attitude.

(iii) Evidence collected

Teacher opinion:

The unit implementation ran quite smoothly. However, there was significant difficulty in directing the thoughts of the students in the right direction to make purposeful research hypotheses. The research hypotheses of students were not sufficiently targeted and took the guidance of the teacher for a smooth transition in the experimental procedure.

Suggestions for improvement

The teacher reported the following suggestions for improvement:

- Fewer substances samples to be clearer colours coming out.
- The quantification of evaluation criteria makes it difficult for students. It would be better before the evaluations of work using rubrics to add an activity in which the students through classroom discussion would set their own criteria along with the underlying quantification