

DEVELOPMENT OF AN INVENTORY FOR ALTERNATIVE CONCEPTION AMONG STUDENTS IN CHEMISTRY

Per-Odd Eggen, Jonas Persson, Elisabeth Egholm Jacobsen, Bjørn Hafskjold

Received: 2.1.2017

Accepted: 30.6.2017

Abstract A chemistry concept inventory (Chemical Concept Inventory 3.0/CCI 3.0) has been developed for assessing students learning and identifying the alternative conceptions that students may have in general chemistry. The conceptions in question are assumed to be mainly learned in school and to a less degree in student's daily life. The inventory presented here aims at functioning as a tool for adjusting teaching practices in chemistry and is mainly designed for assessing the learning outcome during university general chemistry courses. Used as a pre-test the inventory may also give information about student's starting point when entering university's first year chemistry courses. The inventory also aims at functioning as a tool for adjusting teaching practices in chemistry. It has been administered as a pre- and post-test in general chemistry courses at the Norwegian University of Science and Technology (NTNU), and evaluated using different statistical tests, focusing both on item analysis and the on the entire test. The results indicate that the concept inventory is a reliable and discriminating tool in the present context.

1. Introduction

During the last decades, studies aimed at describing the concepts held by students in the fields of Science, Technology, Engineering and Mathematics (STEM) have been performed. Concepts are developed from early age when children form intuitive ideas of natural phenomena. During the process of learning more about the natural world, by observations or a theoretical approach, they develop new or revised concepts based on their own interpretation of new information within their own context with existing ideas and beliefs. The concepts that are not consistent with the established consensus are sometimes called misconceptions (Smith et al 1994) or alternative conceptions.

Students in introductory courses in chemistry exhibit a number of alternative concepts concerning chemical behaviour. There has been published a number of reviews of common alternative conceptions in chemistry (Bowen & Bunce, 1997), (Stavy, Learning Science in the Schools, 1995) (Gabel & Bunce, 1994) (Wandersee, Mintzes, & Novak, 1994), as well as an extensive bibliography (Pfundt & Duit, 2000), dealing with these issues. This research has give valuable background knowledge for later development of assessment methods in chemistry education (Treagust & Chiu, 2011) (Cloonan & Hutchinson, 2010).

Some alternative conceptions influence the learning process in a deeper sense than just producing inadequate explanations to questions. (Nakleh, 1992), (Krajcik, 1991) The students reconstruct their conceptual understanding, in the face of new information, from their present conceptions. When encountering new information that contradicts their alternative conceptions, they might find it difficult to accept this information as it seems to be wrong. This may lead to a conflict that can be solved in different ways, the new information can be; ignored, rejected, disbelieved, deemed irrelevant, held for consideration later, reinterpreted to fit the students' current conceptions or accepted with only minor changes in the student's

conception. The information may also be accepted and the prior conception revised. Chemical concepts such as a solution may be learned in daily life, but many (maybe most) such concepts must be assumed to be learned in school. Tools for mapping and analysing students' conceptions may therefore be useful when adjusting teaching towards a practice that can facilitate deeper understanding.

A number of teaching techniques have emerged during the last decades, with the aim of increasing the learning outcome among students. The new methods have demonstrated that compared to a teacher-centred lecturing, a more student-centred model of education using more hands-on and inquiry-based approach, increases the student's knowledge and conceptual understanding of a subject (Taber, 2009). In order to compare different methods it is important to have assessment tools that are able to measure the students' conceptual understanding as well as to understand what conceptions and background limitations the students have when entering a class.

The Force Concept Inventory (FCI) is an assessment tool created by Hestenes et al. (Hestenes, Wells, & Swackhamer, 1992) for use in high school and college physics classes. A concept inventory consists of a series of multiple-choice questions, based on qualitative conceptual orientated problems. It aims to measure deep understanding and conceptual knowledge rather than a student's ability to solve problems. The results can be compared to the results of students in classes with different teaching methods in order to determine teaching efficiency. In addition, a concept inventory can sample the alternative conceptions in a student population, as well as being able to assess the progression during a course by administering pre- and post-test using the inventory.

Our aim was to develop a chemical concept inventory for use in general chemistry courses at Norwegian universities as well as in upper secondary school in Norway. The inventory serves two purposes: The primary purpose is to map students' understanding of concepts in chemistry. The secondary is to use the inventory as an independent tool for evaluation of learning activities. When comparing a chemical concept inventory with e.g. the Force Concept Inventory, one has to bear in mind that conceptions relating to forces to a large extent is learned in daily life, but most chemical conceptions can be assumed to result from teaching practices. In this article we describe the development of the inventory and some results from its application to students at university level. The applications assured the quality of the inventory, whereas evaluating students' learning activities will have to be based on more extensive and systematic tests. This will be the topic of future papers. Together with reviews of common alternative conceptions in chemistry, concept inventories may constitute a basis for understanding students learning difficulties and achievements. The inventories may also display differences in learning outcomes when comparing students from different learning institutions.

2. Development and validity

The test was developed from a number of existing validated concept inventories in chemistry (Krause, Birk, Bauer, Jenkins, & Pavelich, 2004) (Mulford & Robinson, 2002) in addition to questions written by the authors based on literature and personal experience. The main objective is to use the inventory in the compulsory general chemistry courses for mapping students' understanding of concepts and to use it for finding effects of teaching activities. The inventory is mainly intended for use in first year of university studies, but may have a wider application, for example in upper secondary school. The CCI is, however, only evaluated and validated in a university context. From a test bank of about 113 concept questions originating from the chemistry

inventories of Krause et al. (Krause, Birk, Bauer, Jenkins, & Pavelich, 2004), Mulford and Robinson (Mulford & Robinson, 2002) and about 30 questions made by ourselves, we excluded a number of questions as being out of context or not suitable for our purposes, leaving us with 70 remaining questions. These questions were compared with the curriculum of all the general chemistry courses given at NTNU, by us and other experts at NTNU, as to exclude questions not covered.

The number of questions was considered too large to administer as a single test, and a sub-set of 50 questions was used in a first version (1.0) of the inventory administered as a post-test in a course during the 2014 spring semester. The results from the test were analysed with respect to difficulty and discrimination. Nine questions were judged as unsatisfactory, either being too easy or not sufficiently discriminating. The test was also considered to be too time consuming and a number of questions of duplicate nature were removed. The next step was to construct two tests (2.0 and 2.1) with major overlap of questions while testing new questions. These tests consisting of 38 and 40 questions, respectively, were administered as post-tests in two different courses in the 2014 fall semester. The result judged three and four questions, respectively, to be unsatisfactory. Combining these results made it possible to construct an inventory (3.0) of 40 questions with an estimated completion time of 40-45 min. During the 2015 spring semester this inventory was administered to 60 students in general chemistry courses. 25 of these belonged to phys./math. master program, 18 to material technology, 13 to nano technology and four students to other programs. The chemistry course was compulsory only in the nano and phys./math. programs.

Soliciting expert opinions is a standard method of assessing the validity of a test. The term “validity”, which is not a statistical construct, refers to the extent which the test actually measures what it is supposed to measure. Validity, as such, can have different aspects (Kline, 1986). Face validity can be determined by a common sense reading of an instrument; a test would lack face validity if it tested concepts unrelated to the subject. Content validity reflects the coverage of the subject, i.e. does the test cover enough of a topic? Both of these are typically assessed by expert consensus, as has been done with this inventory. The experts assessing the test have all lectured general chemistry courses for several years at NTNU and partially also at upper secondary school, thus having insight in the relevant chemistry curricula.

The inventory described in this paper covers a wide range of topics generally introduced in an undergraduate general chemistry course, but cannot claim to cover the majority of chemical concepts. It should rather be seen as a starting point, at least in a Nordic context. The questions and the topics that are covered are given in Table 1.

Table 1. Topics covered in the Chemical Concept Inventory

Topic	Question #
Molecular geometry	28, 29
Atomic structure	4
Stoichiometry	24
Chemical bonding	2, 3, 26, 27, 32, 33
Gases	20, 21, 31
Chemical equilibrium	9, 10, 11, 12, 13, 14, 35
Redox reactions and electrochemistry	18, 19, 36, 38
Phase equilibrium	22, 25, 30

Thermochemistry	32, 37
Thermodynamics	1, 34, 39, 40
Intermolecular forces	5,6,7,8,
Acid-base equilibria	15, 16, 17

The sequence of these questions was in part random. We could not detect any effect of the sequence, but this may be investigated in later studies.

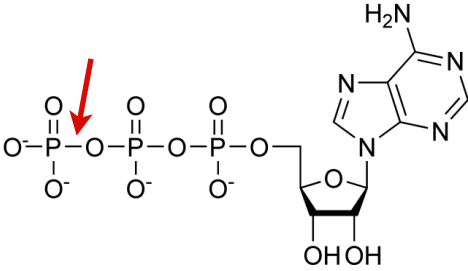
Two examples of questions from this CCI are given in Figures 1 and 2.

Which of the following statements describes a situation where the intermolecular forces are overcome?

- A) Decomposition of ammonia (NH_3) into nitrogen and hydrogen
- B) Decomposition of sodium chloride (NaCl) into chlorine and sodium ions
- C) Evaporation of methanol molecules from liquid methanol (CH_3OH).
- D) Separating hydrogen atoms from each other in hydrogen gas (H_2).
- E) All of the above.

Figure 1: Question 8 from the concept inventory examining the nature of intermolecular forces.

In the picture of an ATP-molecule, a specific bond is indicated with an arrow. Which statement on this bond is CORRECT?



- A) It always takes energy to break a bond. This is also true for this bond.
- B) It takes energy to break most bonds, but this is a high energy bond that will release energy when broken.
- C) All bonds release energy when broken.
- D) It takes energy to break covalent bonds, but this is an ionic bond, which release energy when broken.

Figure 2: Question 33 from the concept inventory examining the nature of intramolecular forces.

3. Results

3.1 Student group

At the Norwegian University of Science and Technology, all engineering masters are required to study at least one chemistry course. This gives us an opportunity, not only to study the results of chemistry masters, but also to study students with different interests in chemistry. The final inventory (3.0) was given in a general chemistry course (TMT4110) with students from three different master programs. In this course, the students are supposedly high-achieving, mastering in Physics, Nanotechnology and Materials Science and Engineering. The Physics and Nanotechnology students are generally regarded as high-achieving students as admission grades are higher compared with the other master programs at NTNU. The results may therefore be expected to be higher for this specific course compared to other courses, possibly with the exception of e.g. Chemistry master students. A study of results for individual questions in the different versions of the inventory through the development phase is consistent with this hypothesis. The test, given as a post-test, was voluntary with no extra credit given. 60 students performed the test during a 45 min exercise (problem solving)-session.

3.2 Analysis of results

In order to investigate the reliability and discrimination power of the inventory, a number of statistical tests focusing both on individual items and on the test as a whole, has been performed.

There exist two aspects with test reliability; consistency and discriminatory power. A test is reliable if it is consistent within itself and over time. If a test is shown to be reliable, one can assume that the same students would get the same score if they would take the test again. A large variance in the test score of a reliable test will then depend on a systematic variation in the student population, where different levels of understanding or mastery will give different scores on the test. Both these aspects of test reliability can be assessed statistically.

Using the results from the inventory we performed five statistical tests: three focusing on individual items (item difficulty index, item discrimination index, item point biserial coefficient) and two on the test as a whole (Kuder-Richardson test reliability and Ferguson's δ). (Kline, 1986) (Kuder & Richardson, 1937)

3.2.1 Item difficulty index

The item difficulty index (P) is a measure of the difficulty of each test item. It is defined as the ratio of the total number N_1 of correct answers to the total number N of students who answered the specific item:

$$P = \frac{N_1}{N} \quad (1)$$

The difficulty index is somewhat misnamed, since it is simply the proportion of correct answers to a particular item, where the name "easiness index" might be more appropriate. The greater the P value, the higher

percentage of correct answers and consequently the easier the item is for the population. The difficulty index will also depend on the population, which is the case in this study.

There are a number of different criteria for acceptable values of the difficulty index for a test (Doran 1980). The optimum value for an item would be $P = 0.5$, while it is useful to have a sensible range. A widely adopted criterion requires the difficulty index to be between 0.3 and 0.9 for each question. For a test with a large number (M) of items, the test difficulty may instead be considered as the average difficulty index (\bar{P}) of all the items (P_i):

$$\bar{P} = \frac{1}{M} \sum_{i=1}^M P_i \quad (2)$$

The results from the test gave a range of [0.33, 0.92] with a mean of 0.66 and only one question over the acceptable limit. This we regard as satisfactory even though the mean is somewhat high, indicating that the questions are easy, however we do not expect any ceiling effects. One will also expect that difficulties will change between different student groups as found in the case of FCI (Persson, 2015).

3.2.2 Item discrimination index

The item discrimination index (D) is a measure of the discriminatory power for individual items in a test. That is, the extent to which an individual test item distinguishes a student who know the material well from those who do not. A high discrimination index will therefore indicate a higher probability for students with a robust knowledge to answer the item correctly, while those with less knowledge or misconceptions more probably will get the wrong answer.

The item discrimination index (D) is calculated by first dividing the sample into two groups of equal size, a high (H) score group and a low (L) score group based on their individual total scores on the test. For a specific item, one counts the number of correct answers in both the high and low groups: N_H and N_L . If the total number of students taking the test is N , the discrimination index for a specific item can be calculated as

$$D = \frac{N_H - N_L}{N / K} \quad (3)$$

where K is a numerical factor based on how the division in the high and low group is made. If we split the sample in two, using the median, the high and low groups consist each of 50% of the total sample, giving $K=2$. However, it is possible to use other groupings, for example taking the top 25% as the high group and the bottom 25% as the low group. The 50% - 50% grouping may underestimate the discrimination power, since it takes all students into account. To reduce the probability of underestimating the discrimination power we use a 25% - 25% grouping, which necessary means that we have to discard half of the available data. The discrimination index is then expressed as:

$$D = \frac{N_H - N_L}{N / 4} \quad (4)$$

The range of the item discrimination index D is $[-1, +1]$, where $+1$ is the best value and -1 the worst. If all students in the high score group and none in the low score group get the correct answer the discrimination

index would be +1. If none in the high score group and all in the low score group get the correct answer the discrimination index would be -1. These extremes are unlikely, but it is a good practice to remove items with negative discrimination index. A question is typically considered to provide a good discrimination if $D \geq 0.3$ (Doran 1980). In a test with a number of items it is possible to allow a few items with a lower discrimination index, but the majority should have high discrimination indices in order to ensure that the test can distinguish strong and weak mastery. It is possible to check this by calculating the averaged discrimination index (\bar{D}) for all items in the test with equation 5:

$$\bar{D} = \frac{1}{M} \sum_{i=1}^M D_i \quad (5)$$

We found that the average discrimination index where 0.45 with a range of [0.13, 0.73] for the individual questions. However, 10 questions in this students group were below the recommended value, though most of them slightly. These were questions 1: (D=0,13); 11 and 31 (D=0,2); 13, 19, 24, 26, 30, 33 and 37 (D=0,27). When comparing the results of these questions in our development tests we found that most of them were over the recommended limit. This indicates that that the questions are slightly less discriminating in this specific student group, which is believed to be high-achieving. As the overall result is satisfactory, the test as a whole is accepted as discriminating enough, though special attention must be paid in the case of high-achieving students.

3.2.3 Point biserial coefficient

The point biserial coefficient is a measure of the individual item reliability. It reflects the correlation between the total score and the score on individual items. A positive coefficient indicates that a student with a high total score is more likely to answer the item correctly than a student with a low total score. To calculate the point biserial coefficient for an item, one calculates the correlation between the score for a question and the total scores. The student's score on an item can have two values: 1 (correct) or 0 (wrong). If the number of items in the test is sufficiently large, >20 , the test can be viewed as continuous. The point biserial coefficient can then be defined as:

$$r_{pbc} = \frac{\bar{X}_1 - \bar{X}_0}{\sigma_x} \sqrt{\frac{P}{1-P}} \quad (6)$$

Where \bar{X}_1 is the average total score for those who answered a item correctly, \bar{X}_0 is the average total score for all, σ_x is the standard deviation of the total scores and P is the difficulty index for this item. A reliable item should be consistent with the whole test so a high correlation between individual questions score and the total score is desirable. A satisfactory point biserial coefficient is $r_{pbc} > 0.2$. Items with lower values may still be used, as long as the number is small, but the test as a whole should have an average higher than 0.2. The average point biserial coefficients for the inventory were found to be 0.41, with only one question below the recommended limit. The range in this case was [0.13, 0.57]. The results show that the test is reliable for our purposes.

3.3 Test analysis

The reliability of single items in the test is measured by the point biserial coefficient. In order to examine the reliability of the whole test, other methods have to be used. In this work we use two measures of the reliability for the test as a whole: Kuder-Richardson reliability index and Ferguson's delta (δ) (Kuder & Richardson, 1937) (Kline, 1986).

3.3.1 Kuder-Richardson reliability index

One way to evaluate the reliability of a test is to administer it twice to the same sample. In such a case we would expect a significant correlation between the two test scores, provided the students' performance is stable and the test conditions are the same. The correlation coefficient between the two sets of scores will be defining the reliability index of the test. It is obvious that this method is not practical to use, as persons in the sample will remember the questions.

In the case of a test that has been specifically designed for a certain knowledge domain with parallel questions, the Spearman-Brown formula (Ghiselli, Campbell & Zedeck 1981) can be used as an alternative to calculate the reliability index. This equation connects the reliability index with the correlation between any two parallel equally sized subsets of the test. Kuder and Richardson took this idea further, by dividing the test into the smallest possible subsets, individual items (Kuder & Richardson, 1937). That is, each item is considered as a single parallel test assuming that the means, variance and standard deviation is the same for all items. The result gives the reliability index as:

$$r_{\text{test}} = \frac{M}{M-1} \left(1 - \frac{\sum \sigma_{xi}^2}{\sigma_x^2} \right) \quad (7)$$

where M is the number of items in the test, σ_{xi} is the standard deviation for the i^{th} item score and σ_x is the standard deviation of the total test score.

This expression takes the different variances of the items into account, relaxing the assumption that all items must have the same means, variance and standard deviation. For multiple-choice tests the formula can be rewritten as:

$$r_{\text{test}} = \frac{M}{M-1} \left(1 - \frac{\sum P_i(1-P_i)}{\sigma_x^2} \right) \quad (8)$$

where M is the number of items in the test, P_i is the difficulty index for each item and σ_x is the standard deviation of the total test score. This is the Kuder-Richardson reliability formulas, often referred to as KR-20 and KR-21 as being formula 20 and 21 in Kuder and Richardson's original paper (Kuder & Richardson, 1937). The possible range of the Kuder-Richardson reliability index is between 0 and 1, where a value greater than 0.7 would make the test reliable for group measurements and a value over 0.8 for assessing individuals (Doran

1980). In this study the obtained Kuder-Richardson reliability index of 0.88. This value indicates that our inventory is also suitable for individual assessment.

3.3.2 Ferguson's delta

Ferguson's delta is another whole test statistic. It measures the discriminatory power of the whole test by investigating how the students' scores are distributed. One aims at a broad distribution in total scores, as this indicates a better discrimination.

The expression of Ferguson's delta can be written as:

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - (N^2 / (M + 1))} \quad (9)$$

where N is the number of students taking the test, M is the number of items in the test and f_i is the frequency of cases with the same score. If a test has a Ferguson's delta greater than 0.90, it is considered to provide a good discrimination among students (Kline 1986, p. 144 and 150). In our study the Ferguson's delta was 0.98.

4. Discussion

There is little knowledge on the effectiveness of different teaching methods in chemistry, and as demonstrated by Mazur (Mazur, 2009) and others, university grades do not always reflect subject knowledge satisfactory. According to John Hattie, most reports from tests of teaching methods are positive, but only a fraction of these methods are in the "zone of desired effects" (Hattie, 2009). Hence, here seems to be a need for tools suitable for selecting teaching methods that can improve students' learning.

Our objective has been to create a concept inventory which can serve as a tool to assess different methods of teaching or interventions in the study of chemistry. Examples of such interventions can be to use videos as a supplement to text-books, laboratory activities or so called "inverted" or "flipped" classroom (Lage, Platt, & Treglia, 2000). The CCI may also be used for other purposes, such as studying connections between different misconceptions or mapping of which topics where the teaching should be reformed. Investigating connections between conceptions requires investigations beyond mapping scores of the complete inventory. This may nevertheless be achieved by comparing students answers to specific questions. Methods suited to measure learning efficacy can be used at universities for more than designing the chemistry study, for example to advice students about choosing the best study methods or influencing the content of chemistry teaching education. Other possible areas of use may be to test the knowledge developed during upper secondary chemistry classes. Concept inventory tests may also indicate quality differences in chemistry text-books, changes caused by altered curricula or other changes in the primary or secondary school system.

The developed chemistry concept inventory is designed to fit the current curricula in general chemistry courses at the Norwegian University of Science and Technology (NTNU). The curricula in these courses are to a high degree consistent with the content of widely used international editions of general chemistry textbooks for colleges and universities. Therefore, it seems probable that the inventory can be used at other teaching institutions, for similar purposes as those described in this article. The main aim has been oriented

at general chemistry courses where the inventory will be used as a post-test. However, the test might also work as a tool for investigating students' individual learning by administering it both as a pre- and post-test. The development of alternative concept inventories can give similar tools, designed for other studies, curricula or narrowed towards one or a few chemistry topics instead of the whole curriculum. As indicated by Mazur (Mazur, 2009) and others, current assessment systems often fail to test students understanding of science concepts. The described concept inventory may also be used to adjust and improve student assessments. Analyses of test results in ongoing studies will give information about the present "state of the art" at our university, and also about the differences between students at this university and other teaching institutions.

5. Conclusions

A chemistry concept inventory for assessing students learning and identifying their alternative conceptions has been evaluated and tested. This inventory, CCI 3.0, aims at functioning as an assessment tool in chemistry education. Used as a post-test, it may give information about the effect of changes in teaching practices in university general chemistry courses. Used as a pre-test in first year university courses, the CCI 3.0 gives information about the learning outcome from upper secondary general science and chemistry courses. A concept inventory used for these purposes must be reliable and have necessary discriminatory power within the context where the inventory is administered. This inventory has been administered and evaluated using statistical tests, and results indicate that the concept inventory is a reliable and discriminating tool in the present context. The results of the tests can be grouped in the main knowledge areas to identify strengths and weaknesses in the students understanding, thus helping students to focus on the areas in need for improvement. The statistical analysis of the reliability and discriminatory power of CCI 3.0 shows that it can be applied within the context of general chemistry courses at NTNU and possibly other universities. The present form of the chemistry concept inventory is by no means the final, but will serve as a template for future versions as well as an inventory suitable for longitudinal studies. Such studies are started both at NTNU and at the university of Jyväskylä and may serve as a tool for investigating and comparing student's conceptions during their first year courses.

Access to the CCI may be given on request to corresponding author.

REFERENCES

- Adams, W. K., Wieman, C. E., Perkins, K. K., & Barbara, J. (2008). Modifying and Validating the Colorado Learning Attitudes about Science Survey for Use in Chemistry. *Journal of Chemical Education*, 85(10), s. 1435.
- Bowen, D., & Bunce, D. M. (1997). Testing for Conceptual Understanding in General Chemistry 1. *The Chemical Educator*, 2, ss. 1-17.
- Cloonan, C. A., & Hutchinson, S. (2010). A Chemistry Concept Reasoning Test. *Chemistry Education Research and Practice*, ss. 205-209.
- Gabel, D. L., & Bunce, D. M. (1994). Research on problem solving: Chemistry. *Handbook of Research on Science Teaching and Learning*, 11, ss. 301-326.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *Phys. Teach.*, 30, s. 141.
- Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. Methuen.
- Krajcik, J. S. (1991). Developing Students' Understanding of Chemical Concepts. I S. M. Glynn, R. H. Yeany, B. K. Britton, & N. J. Lawrence Erlbaum (Red.), *The Psychology of Learning Science* (ss. 117-147).
- Krause, S., Birk, J., Bauer, R., Jenkins, B., & Pavelich, M. J. (2004). Development, testing and application of a chemistry concept inventory. *Annual Frontiers in Education Conference*. Savannah.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, ss. 151-160.
- Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *The Journal of Economic Education*, 31(1), ss. 30-43.
- Mayer, R. E. (2014). Principles for multimedia learning. California, USA: Harvard University.
- Mazur, E. (2009). Farewell, lecture. *Science*, ss. 191-196.
- Mulford, D. R., & Robinson, W. R. (2002). An Inventory for Alternate Conceptions among First-Semester General Chemistry Students. *J. Chem. Educ.*, 79(6), s. 739.
- Nakleh, M. B. (1992). Why some students don't learn chemistry: Chemical misconceptions. *J. Chem. Educ.*, 69, ss. 191-196.
- Persson, R. J. (2015). Evaluating the Force Concept Inventory for different student groups at the Norwegian University of Science and Technology. *arXiv preprint arXiv: 1504.06099*.
- Pfundt, H., & Duit, R. (2000). *Bibliography. Students' alternative frameworks and science education*. Kiel, Germany: University of Kiel Institute for Science Education.
- Smith, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: a Constructivist Analysis of Knowledge in Transition. *The Journal of the Learning Sciences*, 3(2), ss. 115-163.
- Stavy, R. (1991). Using analogy to overcome misconceptions about conservation of matter. *J. Res. Sci. Teach.*, 28, ss. 305-313.
- Stavy, R. (1995). Learning Science in the Schools. I L. Erlbaum:, *Research Informing Practice* (ss. 131-154). Hillsdale.
- Taber, K. S. (2009). *Progressing Science Education: Constructing the scientific research programme into the contingent nature of learning science*. Dordrecht: Springer.
- Treagust, D. F., & Chiu, M. (2011). Diagnostic assessment in chemistry. *Chemistry Education Research and Practice*, ss. 119-120.
- van Marion, P. (2015). Praktisk arbeid. I P. v. (edt), *Biologididaktikk* (s. 257). 7094: Cappelen Damm Akademisk.
- Wandersee, J. H., Mintzes, J. J., & Novak, J. D. (1994). Research on alternative conceptions in science. I D. Gabel (Red.), *Handbook of Research on Science Teaching and Learning*. New York: Macmillan.