

# A Review and Comparison of Diagnostic Instruments to Identify Students' Misconceptions in Science

Derya Kaltakci Gurel  
*Kocaeli University, TURKEY*

Ali Eryilmaz  
*Middle East Technical University, TURKEY*

Lillian Christie McDermott  
*University of Washington, USA*

•Received 10 February 2015•Revised 18 March 2015 •Accepted 20 March 2015

Different diagnostic tools have been developed and used by researchers to identify students' conceptions. The present study aimed to provide an overview of the common diagnostic instruments in science to assess students' misconceptions. Also the study provides a brief comparison of these common diagnostic instruments with their strengths and weaknesses. A total of 273 articles published (from the year 1980 to 2014) in main journals were investigated thoroughly through document analysis method. The study reveals interviews (53%), open-ended tests (34%), multiple-choice tests (32%) and multiple tier tests (13%) as the most commonly used diagnostic tools. However, each tool has some advantages as well as disadvantages over the others that should be kept in mind in their usages. A careful user of a diagnostic instrument such as a classroom teacher or a researcher would be aware of the diagnostic instruments and selects the most effective one for his/her purposes.

*Keywords:* diagnostic instruments, misconceptions, science education

## INTRODUCTION

As students learn about the world around them formally through school education or informally through their everyday experiences, they often tend to form their own views. Because of this concern, several studies have been conducted to depict students' understanding. The different forms of student understandings have been called by a number of different terms such as "alternative conceptions" (Klammer, 1998; Wandersee, Mintzes, & Novak, 1994), "misconceptions" (Clement, Brown, & Zietsman, 1989; Driver & Easley, 1978; Helm, 1980), "naïve beliefs" (Caramazza, McCloskey, & Green, 1980), "children's ideas" (Osborne, Black, Meadows, & Smith, 1993), "conceptual difficulties" (McDermott, 1993), "phenomenological primitives" (diSessa, 1993), "mental models" (Greca & Moreire, 2002) and so forth. Despite variations, all the terms stress differences between the ideas that students bring to instruction and the concepts by the current scientific theories. Whatever it is called, in almost all of these studies, the main aim is the

Correspondence: Derya Kaltakci Gurel,  
Department of Science Education, Kocaeli University, Umuttepe Campus, 41380,  
Kocaeli, Turkey.  
E-mail: kaderya@kocaeli.edu.tr  
doi: 10.12973/eurasia.2015.1369a

understanding of wrong and flawed conceptions that impedes learning or the identification of productive components of these flawed conceptions for other contexts. Therefore, the identification of these conceptions in a valid and reliable way becomes a prominent first step. In the present study, the term "misconception" is going to be used for those conceptions that contradict the scientifically accepted theories because of its common usage in the literature.

Effective test development requires a systematic, well-organized approach to ensure sufficient validity and reliability evidence to support the proposed inferences from the test scores (Downing, 2006). Hammer (1996) makes an analogy between a researcher exploring knowledge structure of individuals and a doctor diagnosing diseases. In this analogy, Hammer emphasizes the importance of studies in education that explore individuals' conceptions. According to him, a doctor who knows only one or two diseases would have only one or two options for diagnosing an ailment, regardless of the technical resources available. When the diagnosis is correct, the prescribed treatment may be effective; however, when the diagnosis is not correct, the treatment may not only be ineffective, it may be damaging. With this analogy, it is clear that studies focusing on conceptual understanding and methods to diagnose misconceptions in a valid and reliable way have great importance in science education research. Diagnostic tests are assessment tools which are concerned with the persistent or recurring learning difficulties that are left unresolved and are the causes of learning difficulties (Gronlund, 1981). In other words, these instruments bring to light the disparity between what we want our students know or learn and what they really know or learn.

This article addresses the importance of diagnostic assessment in science and presents an overview of the diagnostic instruments to assess misconceptions in the science education research literature since the 1980s in a comparative manner. The significance of this study lies in its contribution to the literature with the overview of the common diagnostic instruments in science to assess misconceptions. Also, the study provides a brief comparison of these common diagnostic instruments with their strengths and weaknesses. A careful user of a diagnostic instrument such as a classroom teacher or a researcher would be aware of these instruments and selects the most effective one for his/her purposes.

## METHOD

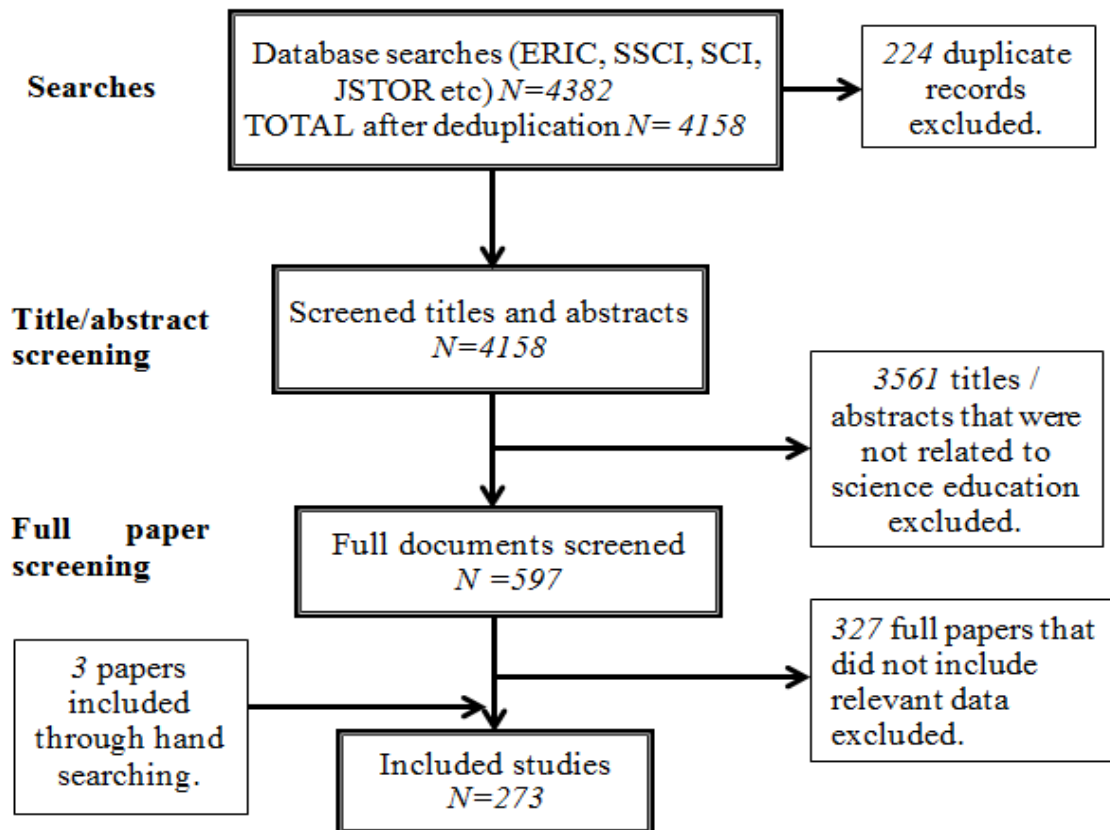
In this study, articles published in major journals in the field of science education and indexed in main databases were investigated thoroughly to gain data about the diagnostic instruments on misconception assessment. To identify relevant studies in the literature we conducted a systematic search of 9 databases with document analysis method. We limited the search to studies in English which were published

### **State of the literature**

- Studies on students' conceptions and reasoning in science have gained impetus over the last four decades. The main aim of these studies is the understanding of wrong and flawed conceptions that impedes learning or the identification of productive components of these flawed conceptions for other contexts.
- Identification of student conceptions, which widely called as "misconceptions", in a valid and reliable way becomes a prominent first step to deal with them.
- Diagnostic tests are assessment tools which are concerned with the persistent or recurring learning difficulties that are left unresolved and are the causes of learning difficulties.

### **Contribution of this paper to the literature**

- The significance of the present study lies in its contribution to the literature with the critical overview of the common diagnostic instruments in science education to assess students' misconceptions.
- According to the study, all diagnostic assessment methods were found to have their own strengths and limitations. Therefore, a combination of many methods is considered to be better than a single method.
- The results of this study shows that more emphasis should be given on three and four tier tests in all fields of science.



**Figure 1.** Flow of studies for inclusion in the review

between 1980 and 2014 in order to obtain most of the studies on misconception diagnosis in science. A multistage process was followed whereby each article was read and information from the articles were identified and discussed between two researchers. After narrowing from 4382 articles originally identified with an abstract keyword search, the present study included a total of 273 research articles whose abstracts revealed a focus on diagnosis of students' misconceptions in science.

In the literature searching, an iterative process was followed. Each found article's reference list was used as a source of new references. Obtained articles were investigated thoroughly in terms of their methods of misconception diagnosis. General discussions about diagnostic instruments in the articles were used to compare their strengths and weaknesses against the other methods. Figure 1 illustrates the flow of studies through the search and selection process. In documenting the references, the authors gave special attention to the ordinary multiple-choice tests and multiple tier multiple-choice tests because of the inadequacy of such a documenting for these instruments. This does not mean the other instruments (such as interviews or open-ended tests) deceive their effectiveness nowadays, yet they are still influential.

## RESULTS

In order to measure students' conceptions on several concepts, different diagnostic tools have been developed and used. Among them interviews, open-ended tests and multiple-choice tests are found to be the ones commonly used in science education research in order to identify misconceptions. However, each tool has some advantages as well as disadvantages over the others as discussed in several studies.

**Table 1.** Percentages of diagnostic tools used in examined studies to identify misconceptions (N = 273)

| Method of Diagnosis   | Percentage (%) |
|-----------------------|----------------|
| Interviews            | 53             |
| Open-ended tests      | 34             |
| Multiple-choice tests | 32             |
| Multiple-tier tests   |                |
| --Two-tier            | 9              |
| --Three-tier          | 3              |
| --Four-tier           | 1              |
| Others                | 9              |

Among 273 studies included in this study, the most common diagnostic tool was found to be interviews (53 %). Table 1 shows the percentages of diagnostic tools used in the examined papers in this study. The total percentages do not add up to 100 per cent since part of the studies use multiple diagnostic methods. Of the all analyzed studies, 42 % used a single diagnostic method, while 58 % used a combination of two or more of the diagnostic methods. The diagnostic tools labeled as 'others' include concept maps, word association, drawings, essays, etc.

In the following sections interviews, open-ended tests, ordinary multiple-choice tests, and multiple-tier tests are discussed in detail as the most frequently (above 10 %) used methods for diagnosing students' misconceptions in science according to reviewed articles. Sample items of misconception diagnostic instruments and brief explanation of their analysis are given in the Appendix.

## Interviews

Among various methods of diagnosing misconceptions, interviews have the crucial role because of their in-depth inquiry and possibility of elaboration to obtain detailed descriptions of a student's cognitive structures. In fact, interviews have been found to be one of the best (Franklin, 1992; Osborne & Gilbert, 1980b), and the most common (Wandersee et al., 1994) approach used in uncovering students' views and possible misconceptions. Several interviewing techniques have been used in the literature such as Piagetian Clinical Interviews (PCI) (Piaget, 1969; Ross & Munby, 1991), Interview-About-Instances (IAI) (Osborne & Gilbert, 1979), Interviews-About-Events (IAE) (Bau-Jaoude, 1991; Osborne & Freyberg, 1987; Osborne & Gilbert, 1980a), Prediction-Observation-Explanation (POE) (White & Gunstone, 1992), Individual Demonstration Interview (IDI) (Goldberg & McDermott, 1986; 1987), Teaching Experiment (TE) (Komorek & Duit, 2004). Interviews may be conducted with individuals or with groups (Eshach, 2003; Galili & Goldberg, 1993; La Rosa, Mayer, Paqtirxi, & Vincentini, 1984; Olivieri, Totosantucci & Vincentini, 1988; Van Zee, Hammer, Bell, Roy, & Peter, 2005). Duit, Treagust and Mansfield (1996) stated that the group interviews have the strength of studying the development of ideas in the interaction process between students.

The purpose of interviewing was stated by Frankel and Wallen (2000) as finding out what is on people's mind, what they think or how they feel about something. As stated by Hestenes, Wells and Swackhamer (1992) when skillfully done, interviewing is one of the most effective means of dealing with misconceptions. Although interview strategies have the advantages such as gaining in-depth information and flexibility, a large amount of time is required to interview a large number of people in order to obtain greater generalizability. Also training in interviewing is required for the researcher. In addition, interviewer bias may taint the findings. The analysis of data is a little bit difficult and cumbersome (Adadan & Savasci, 2012; Rollnick & Mahooana, 1999; Sadler, 1998; Tongchai et al., 2009).

## Open-ended tests

In order to investigate students' understanding, open-ended free-response tests were also used commonly in science education. This method gives test takers more

time to think and write about their own ideas, but it is difficult to evaluate the results (Al-Rubayea, 1996). Also because of language problems, identification of students' misconceptions becomes difficult (Bouvens as cited in Al-Rubayea, 1996) since students are generally less eager to write their answers in full sentences. Andersson and Karrqvist (1983), Colin, Chauvet and Viennot (2002), Langley, Ronen and Eylon (1997), Palacios, Cazorla and Cervantes (1989), Ronen and Eylon (1993), Wittman (1998) investigated misconceptions of students with open-ended questions or tests as a diagnostic instrument.

### **Ordinary multiple-choice tests**

In order to overcome the difficulties encountered in interviewing and open-ended testing processes, diagnostic multiple-choice tests, which can be immediately scored and applied to a large number of subjects, have been used to ascertain students' conceptions. These tests have been used either following in-depth interviews or alone as a broad investigative measure.

The development of multiple-choice tests on students' misconceptions makes a valuable contribution to the body of work in misconceptions research, assists in the process of helping science teachers more readily use the findings of research in their classrooms (Treagust, 1986). Results from diagnostic multiple-choice tests have been reported frequently in misconception literature. The validity evidence for this format is strong (Downing, 2006). From the point of view of teachers' usage, valid and reliable, easy-to-score, easy-to-administer, paper-and-pencil instruments enable teachers to effectively assess students' understanding of science. A science teacher can get information about students' knowledge and misconceptions by use of the diagnostic instruments. Once the student misconceptions are identified, teachers can work to remedy the faulty conceptions with appropriate instructional approaches. Advantages of using multiple-choice tests over other methods have been discussed by several authors (Çataloğlu & Robinett, 2002; Caleon & Subramaniam, 2010a; Iona, 1982; Tamir, 1990). To sum up, some of the advantages of multiple-choice tests are: (1) They permit coverage of a wide range of topics in a relatively short time. (2) They are versatile, and can be used to measure different levels of learning and cognitive skills. (3) They are objective in terms of scoring and therefore more reliable. (4) They are easily and quickly scored. (5) They are good for students who know their subject matter but are poor writers. (6) They are suitable for item analysis by which various attributes can be determined. (7) They provide valuable diagnostic information and are viable alternatives to interviews and other qualitative tools in gauging students' understanding and in determining the prevalence and distribution of misconceptions across a population.

The chief difficulty in these tests, however, is in interpreting students' responses if the items have not been constructed carefully (Duit et al., 1996). Researchers developed test items with distracters based on students' answers to essay questions or on other open-ended tests or interviews. Beichner (1994) suggested the combination of the strengths of interviewing technique and multiple-choice exams as an ideal course of action in order to investigate students' understanding in physics.

In spite of the advantages of multiple-choice tests mentioned above, there are some criticisms of them. Chang, Yeh and Barufaldi (2010) and Bork (1984) stated certain limitations and drawbacks of multiple-choice questions, such as: (1) Student guessing contributes to the error variance and reduces the reliability of the test. (2) Selected choices do not provide deep insights into student ideas or conceptual understanding. (3) Students being forced to choose each answer from among a very limited list of options, which is preventing them from constructing, organizing and

**Table 2.** Ordinary multiple-choice conceptual tests in science

| Field     | Conceptual Tests                                       | References  |
|-----------|--|---|
| Physics   | Force Concept Inventory (FCI)                          | (Hestenes, Wells & Swackhamer, 1992)                          |
|           | Force & Motion Conceptual Evaluation (FMCE)            | (Thornton & Sokoloff, 1998)                                   |
|           | Mechanic Baseline Test (MBT)                           | (Hestenes & Wells, 1992)                                      |
|           | Energy & Momentum Conceptual Survey (EMCS)             | (Singh & Rosegrant, 2003)                                     |
|           | Test of Understanding Graphs in Kinematics (TUG-K)     | (Beichner, 1994)  |
|           | Electric Circuits Conceptual Evaluation (ECCE)         | (Sokoloff, 1993)  |
|           | Diagnosing and Interpreting Resistor Circuits (DIRECT) | (Engelhardt & Beichner, 2004)                                 |
|           | Conceptual Survey in Electricity & Magnetism (CSEM)    | (Maloney, O’Kuma, Heiggelke & Van Heuvelen, 2001)             |
|           | Brief Electricity & Magnetism Assessment Tool (BEMA)   | (Ding, Chabay, Sherwood & Beichner, 2006)                     |
|           | Light & Optics Conceptual Evaluation (LOCE)            | (Sokoloff, 1997)  |
| Chemistry | Light and Spectroscopy Concept Inventory (LSCI)        | (Bardar, Prather, Brecher & Slater, 2007)                     |
|           | Quantum Mechanical Visualization Inventory (QMVI)      | (Cataloglu & Robinett, 2002)                                  |
|           | Mechanical Waves Conceptual Survey                     | (Tongchai, Sharma, Johnston, Arayathanitkul & Soankwan, 2009) |
| Chemistry | Chemistry Concept Inventory (CCI)                      | (Mulford & Robinson, 2002)                                    |
|           | Solution Concept Test                                  | (Uzuntiryaki & Geban, 2005)                                   |
|           | Heat and Temperature Concepts Test (HTCT)              | (Baser & Geban, 2007)   |
| Biology   | Conceptual Inventory of Natural Selection (CINS)       | (Anderson, Fisher & Norman, 2002)                             |
|           | Biology Concept Inventory (BCI)                        | (Klymkowsky & Garvin-Doxas, 2008)                             |
|           | Developmental Biology Content Survey                   | (Knight & Wood, 2005)   |
|           | Introductory Molecular Cell Biology Assessment (IMCBA) | (Shi, Wood, Martin, Guild, Vicens & Knight, 2010)             |
|           | Molecular Life Sciences Concept Inventory              | (Howitt, Anderson, Hamilton & Wright, 2008)                   |
| Biology   | Meiosis Concept Inventory                              | (Kalas, O’Neil, Pollock & Birol, 2013)                        |

presenting their own answers. (4) It is extremely difficult to write good multiple-choice questions.

Another very common criticism that was described by Rollnick and Mahooana (1999) is that multiple-choice tests do not provide deep enough insight into students’ ideas on the topic and students may give correct answers for wrong reasons. In other words, ordinary multiple-choice tests cannot differentiate correct answers due to correct reasoning from those that are due to incorrect reasoning (Caleon & Subramaniam, 2010a; Eryılmaz, 2010), so they may overestimate student scores (Kaltakçı, 2012; Peşman & Eryılmaz, 2010). Hestenes et al. (1992) proposed “false positive” and “false negative” concepts in order to emphasize the importance of accuracy of measures in a multiple-choice test. False positive is defined as a Newtonian answer chosen with non-Newtonian reasoning; whereas false negative is a non-Newtonian answer with Newtonian reasoning. False negatives are considered unproblematic and are attributed to carelessness or inattention. The minimization of false positive answers, on the other hand, is difficult. The authors stated that the major problem in a multiple-choice test development is to minimize false positives and negatives. Regarding this concern, Tamir (1990) conducted a study requiring students to justify their answers to the multiple-choice test. The results of the study revealed that students who chose the correct answer were not necessarily able to provide correct justifications. To overcome the limitations of multiple-choice tests, tests with multiple-tiers have been developed.

There exist several examples of multiple-choice tests in the literature which are usually called ‘conceptual tests’ or ‘inventories’. These tests were specifically designed as a result of thorough research in the field, and usually each distracter gives evidence about specific student misconceptions. Table 2 gives the references for most often used conceptual tests in science on a variety of topics.

Researchers recognized the difficulty in uncovering misconceptions by ordinary multiple-choice tests since the reason behind a students' selection is not evident (Griffard & Wandersee, 2001). Therefore, they extended multiple-choice tests into tests with two, three, or four tiers in order to compensate for the limitations of the ordinary multiple-choice tests used in diagnosing students' conceptions.

### **Two-tier multiple-choice tests**

In order to gather data from more students than is possible by interviews, justifications to multiple-choice items were used (Hrepic, 2004; Tamir, 1989) in which students were required to justify their selection of answers in multiple-choice items in the form of short answers. These justifications were recommended to be used as raw material for the construction of two-tier tests.

Generally, the two-tier tests were described as diagnostic instruments with first tier, including multiple-choice content questions, and second tier, including multiple-choice set of reasons for the answer to the first tier (Adadan & Savasci, 2012; Chen, Lin & Lin, 2002; Griffard & Wandersee, 2001; Treagust, 1986). Students' answers to each item were considered correct when both the correct choice and reason are given. Distracters were derived from students' misconceptions gathered from the literature, interviews, and open-ended response tests. Two-tier tests were considered a great improvement over the previous approaches in that these tests consider students' reasoning or interpretation behind their selected response and link their choices to misconceptions of the target concept (Wang, 2004). Also, as stated by Adadan and Savasci (2012), two-tier diagnostic instruments are relatively convenient for students to respond to and more practical and valuable for teachers to use in terms of reducing guesswork, allowing for large-scale administration and easy scoring, and offering insights into students' reasoning.

Since Treagust (1986) published his seminal work on the development of two-tier test, large number of researchers have developed and administered two-tier diagnostic tests in biology, chemistry, and physics. Table 3 summarizes the two-tier tests published in science with their references.

The study which criticized two-tier tests was done by Griffard and Wandersee (2001) in the discipline of biology. In order to examine the effectiveness of two-tier tests they used an instrument developed by Haslam and Treagust (1987) on photosynthesis. The study was conducted on six college students by paper and pencil instrument designed to detect alternative conceptions, and the participants responded to the two-tier instrument in a think-aloud task. The findings of the study raised concerns about the validity of using two tier tests for diagnosing alternative conceptions, since they claimed that the two-tier tests may diagnose alternative conceptions invalidly. It is not certain whether the students' mistakes were due to misconceptions or unnecessary wording of the test. Another concern about two-tier tests that was expressed by Tamir (1989) was that the forced-choice items in two-tier tests provide clues to correct answers that participants would not have had in an open-ended survey or interview. For instance, a student can choose an answer in the second tier on the basis of whether it logically followed from their responses to the first tier (Griffard & Wandersee, 2001; Chang et al., 2007), or the content of each choice of the second tier seems partially correct to a responder, but this partially correct response may attract the responder (Chang et al., 2007). Caleon and Subramaniam (2010a) and Hasan, Bagayoko and Kelley (1999) called the attention to significant limitations of two-tier tests in that, those tests cannot differentiate mistakes due to lack of knowledge from mistakes due to existence of alternative conceptions; and they cannot differentiate correct responses due to scientific knowledge and those due to guessing. Thus, two-tier tests might overestimate or underestimate students' scientific conceptions (Chang et al., 2007) or overestimate

**Table 3.** Two-tier multiple-choice conceptual tests in science

| Field           | Two-Tier Conceptual Tests  | References  |
|-----------------|--|---|
| Physics         | The Test of Image Formation by Optical Reflection (TIFOR)  | (Chen, Lin & Lin, 2002)   |
|                 | Student Understanding of Light and Its Properties  | (Fetherstonaugh & Treagust, 1992)   |
|                 | Light Propagation Diagnostic Instrument (LPDI)   | (Chu, Treagust & Chandrasegaran, 2009)  |
|                 | Two-tier Physics Questionnaire (on mechanics, electricity and magnetism, heat, sound and wave, and optics) | (Chang, Chen, Guo, Chen, Chang, Lin, Su, Lain, Hsu, Lin, Chen, Cheng, Wang & Tseng, 2007) |
| Chemistry       | Test to Identify Student Conceptualizations (TISC) (in chemical equilibrium)                               | (Voska & Heikkinen, 2000)   |
|                 | Qualitative Analysis Diagnostic Instrument (QADI) (inorganic chemistry)                                    | (Tan, Goh, Chia & Treagust, 2002)   |
|                 | Covalent Bonding and Structure Diagnostic Test   | (Peterson, Treagust & Garnett, 1986) (Treagust, 1986)                                     |
|                 | Nature of Solutions and Solubility Diagnostic Instrument (NSS-DI)  | (Adadan & Savasci, 2012)  |
|                 | Boiling Concept Test   | (Coştu, Ayas, Niaz, Ünal & Çalık, 2007)   |
|                 | Two-tier Chemical Equilibrium Test   | (Akkus, Kadayifci & Atasoy, 2011)   |
|                 | Two-tier Separation of Matter Test   | (Tuysuz, 2009)  |
|                 | Two-tier Chemical Concept Tests  | (Chiu, 2007)  |
|                 | Acid-Base Diagnostic Test (ABDT)   | (Artdej, Ratanaroutai, Coll & Thongpanchang, 2010)  |
|                 | Representational Systems and Chemical Reactions Diagnostic Instruments (RSCRDI)                            | (Chandrasegaran, Treagust & Mocerino, 2007)   |
| Biology         | Two-tier Diagnostic Instrument in Genetics   | (Tsui & Treagust, 2010)   |
|                 | What do you know about photosynthesis and respiration in plants? Diagnostic Test                           | (Haslam, 1986) (Treagust, 1986)<br>(Haslam & Treagust, 1987)                              |
|                 | Two-tier Diagnostic Malaria Test   | (Cheong, Treagust, Kyeleve & Oh, 2010)  |
|                 | Flowering Plant Growth and Development Diagnostic Test   | (Lin, 2004)   |
|                 | Two-tier Instrument on Photosynthesis  | (Griffard & Wandersee, 2001)  |
|                 | Two-tier Diagnostic Instrument in Plants and Human Circulatory System (ITP&HCS)                            | (Wang, 2004)  |
|                 | Diffusion and Osmosis Diagnostic Test (DODT)   | (Odom & Barrow, 1995)   |
|                 | Two-tier Diagnostic Instrument for Cell Division and Reproduction  | (Sesli & Kara, 2012)  |
|                 | Two-tier Genetics Concepts Test  | (Kılıç & Sağlam, 2009)  |
|                 | Breathing and Respiration Test   | (Mann & Treagust, 1998)   |
| Mineral Concept | (Monteiro, Nobrega, Abrantes & Gomes, 2012)  |   |

the proportions of the misconceptions since the gap in knowledge could not be determined by two tier tests (Aydın, 2007; Caleon & Subramaniam, 2010a, 2010b; Kutluay, 2005; Peşman & Eryılmaz, 2010; Türker, 2005). Chang et al. (2007) also mentioned that since the choices in the second tier constructed from the results of interviews, open-ended questionnaires and the literature review, students are likely to construct their own conceptions out of these and may tend to choose any item of the second tier arbitrarily. In order to eliminate this problem, a blank alternative was included with the multiple-choice items so that responders could write an answer that is not provided (Aydın, 2007; Eryılmaz, 2010; Kaltakçı, 2012; Peşman & Eryılmaz, 2010; Türker, 2005).

To sum up, two-tier tests have advantages over ordinary multiple-choice tests. The most important of them is that those tests provide students' reasoning or interpretation behind their selected response. However, these tests have some limitations in discriminating lack of knowledge from misconceptions, mistakes, or scientific knowledge. For this reason, three-tier tests become crucial in order to



determine whether the answers given to the first two tiers are due to a misconception or a mistake due to lack of knowledge.

### Three-tier multiple-choice tests

The limitations mentioned for the two-tier tests were intended to be compensated by incorporating a third tier to each item of the test asking for the confidence in the answers given in the first two tiers (Aydın, 2007; Caleon & Subramaniam, 2010a; Eryılmaz, 2010; Kutluay, 2005; Peşman & Eryılmaz, 2010; Türker, 2005). In three-tier tests, researchers constructed a multiple-choice test; the first tier of which included an ordinary multiple-choice test, the second tier of which was a multiple-choice test question asking for the reasoning, and the third tier of which was a scale asking for the students' confidence level for the given answers for the above two. Students' answers to each item were considered correct when both the correct choice and reason are given with a high confidence. Similarly, students' answers were considered as misconceptions when a wrong answer choice is selected with an accompanied wrong reasoning and with a high confidence. Three tier tests are considered to be more accurately eliciting the student misconceptions, since they can detect lack of knowledge percentages by means of the confidence tiers. This helps the test users such that the obtained percentage of misconception is free from false positives, false negatives and lack of knowledge, since each requires a different remediation and treatment.

In many of the three-tier test development processes, the researchers benefited from diverse methods of diagnosis of misconceptions (interviews, open-ended tests, concept maps). The diversity in the data collection methods enabled the researchers to gain valuable information about the students' misconceptions as well as providing a good foundation for developing a valid and reliable diagnostic assessment tool. Table 4 summarizes the three-tier tests published in science with their references.

Consequently, three tier tests had the advantage of discriminating the students' lack of knowledge from their misconceptions. Hence, they were considered to assess student misconceptions in a more valid and reliable way compared to ordinary multiple-choice tests and two-tier tests (Aydın, 2007; Eryılmaz, 2010; Kutluay, 2005; Peşman & Eryılmaz, 2010; Türker, 2005). However, since in three-tier tests, students were asked for their confidence for the choices in the first two tiers covertly, this might underestimate proportions of lack of knowledge and overestimate student scores. For this reason, four-tier tests in which confidence

**Table 4.** Three-tier multiple-choice conceptual tests in science

| Field     | Three-Tier Conceptual Tests   | References                          |
|-----------|---|-------------------------------------|
| Physics   | Three Tier Heat & Temperature Test                                  | (Eryılmaz, 2010)                    |
|           | Simple Electric Circuit Diagnostic Test (SEC DT)                    | (Peşman & Eryılmaz, 2010)           |
|           | The Wave Diagnostic Instrument (WADI)                               | (Caleon & Subramaniam, 2010a)       |
|           | Three Tier Circular Motion Test                                     | (Kızılıcık & Güneş, 2011)           |
|           | Gravity Concept Test  | (Kaltakci & Didis, 2007)            |
|           | Electricity Concept Test  | (Aykutlu & Şen, 2012)               |
| Chemistry | States of Matter Diagnostic Test (SMDT)                             | (Kirbulut & Geban, 2014)            |
|           | Three-tier Acids and Bases Test                                     | (Cetin-Dindar & Geban, 2011)        |
| Biology   | Atmosphere-Related Environmental Problems Diagnostic Test (AREPDiT) | (Arslan, Cigdemoglu & Moseley 2012) |

ratings were asked for the content and reasoning tiers separately are introduced more recently.

### Four-tier multiple-choice tests

Even though three-tier tests were thought to be measuring misconceptions free from errors and lack of knowledge in a valid way, they still have some limitations due to the covert rating of the confidence for the first and second tiers in those tests. This situation may result in two problems: one is the underestimation of the lack of knowledge proportions, and the other one is the overestimation of the students' misconception scores and the correct scores. To explain these problems in three-tier tests, one can look at Table 5 and Table 6 below. Table 5 provides the comparison of decisions for four-tier and three-tier tests in determining the lack of knowledge based on the possible student rating of confidence in four-tier tests. For example, if a student is "sure" about his answer in the main question tier and "not sure" about his answer in the reasoning tier in a four-tier test, the researcher can decide "lack of knowledge" for that item. However, in the corresponding three-tier form of the same item the student may indicate his confidence for the main and reasoning tiers either as "sure" or "not sure". As a result, depending on the rating of confidence, the researcher may have a decision of "lack of knowledge" if he is "not sure"; or "no lack of knowledge" if he is "sure". Hence, proportion of lack of knowledge may be underestimated in three-tier tests.

Similarly, in the decision of misconception scores and correct scores, three-tier tests overestimate the proportions of those scores compared to the four-tier tests. Table 6 compares the decisions for three and four-tier tests. For instance, in a four tier test, if a student gives a correct answer to the main question in the first tier and is sure about his answer for this tier, then gives a correct answer to the reasoning question in the third tier but is not sure about his answer for this tier, then the researcher's decision about the student's answer for this item is "lack of knowledge" because there is doubt about at least one tier of the student's answer. However in a parallel three-tier test in which the confidence rating is asked for two tiers together, the same student may select "sure" or "not sure" since he is not sure for at least one of the tiers. If he chooses "not sure" the researcher's decision would be that student has a "lack of knowledge", but if the student chooses "sure" then the researchers' decision for that student's answer for this item would be he has a "scientific knowledge" on this item. Hence his correct score would be overestimated. In the science education literature, there exist a limited number of four-tier misconception tests which are summarized in Table 7.

Even though four tier multiple-choice tests seem to eliminate many problems of the aforementioned instruments, they still possess several limitations such as: requiring a longer testing time, not advisable for using in achievement purposes (Caleon & Subramaniam, 2010b), and the possibility of students' choice of response in the first tier can influence their choice of response in the reasoning tier (Sreenivasulu & Subramaniam, 2013).

**Table 5.** Comparison of four-tier tests and three-tier tests in terms of determining lack of knowledge

| Four-tier Test              |                             |   | Three-tier test   |  |
|-----------------------------|-----------------------------|---|---|--|
| Confidence for the 1st tier | Confidence for the 3rd tier | Decision of researcher for LK in four-tier test | Corresponding possible student selection in three-tier test | Decision of researcher for LK in three-tier test |
| Sure                        | Sure                        | No LK   | Sure  | No LK  |
| Sure                        | Not sure                    | LK  | Sure  | No LK if "sure"                                  |
|                             |                             |   | Not sure  | LK if "not sure"                                 |
| Not sure                    | Sure                        | LK  | Sure  | No LK if "sure"                                  |
|                             |                             |   | Not sure  | LK if "not sure"                                 |
| Not sure                    | Not sure                    | LK  | Not Sure  | LK   |

LK: Lack of Knowledge

**Table 6.** Comparison of decisions in four-tier tests and three-tier tests

| 1st tier | 2nd tier | 3rd tier | 4th tier | Decision for four-tier test | Decision for three-tier test                              |
|----------|----------|----------|----------|-----------------------------|---|
| Correct  | Sure     | Correct  | Sure     | SC                          | SC  |
| Correct  | Sure     | Correct  | Not sure | LK                          | SC if "sure"<br>LK if "not sure"                          |
| Correct  | Not sure | Correct  | Sure     | LK                          | SC if "sure"<br>LK if "not sure"                          |
| Correct  | Not sure | Correct  | Not sure | LK                          | LK  |
| Correct  | Sure     | Wrong    | Sure     | FP<br>Rarely MSC            | FP<br>Rarely MSC  |
| Correct  | Sure     | Wrong    | Not sure | LK                          | FP if "sure"<br>Rarely MSC if "sure"<br>LK if "not sure"  |
| Correct  | Not sure | Wrong    | Sure     | LK                          | FP if "sure"<br>Rarely MSC if "sure"<br>LK if "not sure"  |
| Correct  | Not sure | Wrong    | Not sure | LK                          | LK  |
| Wrong    | Sure     | Correct  | Sure     | FN                          | FN  |
| Wrong    | Sure     | Correct  | Not sure | LK                          | FN if "sure"<br>LK if "not sure"                          |
| Wrong    | Not sure | Correct  | Sure     | LK                          | FN if "sure"<br>LK if "not sure"                          |
| Wrong    | Not sure | Correct  | Not sure | LK                          | LK  |
| Wrong    | Sure     | Wrong    | Sure     | MSC<br>Rarely MTK           | MSC<br>Rarely MTK   |
| Wrong    | Sure     | Wrong    | Not sure | LK                          | MSC if "sure"<br>Rarely MTK if "sure"<br>LK if "not sure" |
| Wrong    | Not sure | Wrong    | Sure     | LK                          | MSC if "sure"<br>Rarely MTK if "sure"<br>LK if "not sure" |
| Wrong    | Not sure | Wrong    | Not sure | LK                          | LK  |

SC: Scientific Conception; LK: Lack of Knowledge; FP: False Positive; FN: False Negative; MSC: Misconception; MTK: Mistake

**Table 7.** Four-tier multiple-choice conceptual tests in science.

| Field     | Four-Tier Conceptual Tests                   | References                         |
|-----------|--|------------------------------------|
| Physics   | Four Tier Wave Diagnostic Instrument (4WADI) | (Caleon & Subramaniam, 2010b)      |
|           | Four Tier Geometrical Optics Test (FTGOT)    | (Kaltakçı, 2012)                   |
| Chemistry | Thermodynamics Diagnostic Instrument (THEDI) | (Sreenivasulu & Subramaniam, 2013) |
| Biology   | -----  | -----                              |

## **DISCUSSION, CONCLUSION AND SUGGESTIONS**

Based on the comprehensive search of the literature related to misconceptions research in science education, researchers have reported a variety of methods for diagnosing misconceptions. However, they have not reached a consensus regarding the best method for this purpose. It depends on the context of the topic to be investigated, the characteristics of the intended subjects to be investigated, and the ability and resources of the researcher or the teacher. However, it is well known that a combination of many methods is better than a single method (Beichner, 1994; Schmidt, 1997). Therefore, in order to make valid inferences about students' misconceptions, several diagnostic tools used together and yielded particularly valuable results. Oral and written instruments have different nature of inquiries and combining them strengthen the inferences made based on the obtained data and eliminate the probable weaknesses coming from the nature of a single instrument.

In a previous study, among 103 misconception studies examined by Wandersee et al. (1994), 46 % used interviews, 20 % used multiple-choice tests, 19 % used sorting tasks, 8 % used questionnaire, and 6 % used open-ended tests. Comparing them with the results of the current study shows that the diagnostic tool trends in identifying misconceptions does not change a lot. Interviews with their in-depth inquiry are still among the most widely used diagnostic instruments in science. In some studies interviews were used alone (Eshach, 2003; Kirbulut & Beeth, 2013; Osborne & Gilbert, 1979), whereas in a numerous number of other studies they were used prior to written tests (Goldberg & McDermott, 1987), after the written test (Caleon & Subramaniam, 2010a; Hestenes et al., 1992; Schmidt, 1997), or during the test development process to construct the items of the written tests (Griffard & Wandersee, 2001; Pesman & Eryilmaz, 2010). On the other hand, the use of multiple-choice tests (ordinary or multiple-tier) increased in a deal recently compared to the aforementioned study. The usage of multiple-tier tests have gained impetus since 1990s and they are still under interest to get the most benefit. However, the number of ordinary multiple-choice tests in chemistry is found to be small compared to the other two fields. The number of three- and four-tier multiple choice tests in all of the three fields are still small and needs to be increased.

Since misconceptions are very resistant to change and problematic for further scientific knowledge, it is crucial to determine them correctly. Incorrect reasoning on multiple-tier multiple-choice test items provide a rich source of students' misconceptions. Addition of confidence rating in three and four-tier tests gives opportunity to assess the nature and strength of those misconceptions. They assess misconceptions which are free of errors and lack of knowledge in an easy manner. This helps the implementers of the tests (teachers or reserachers in science) because misconceptions and lack of knowledge in a subject require different interventions and discrimination of them from each other is crucial and important for this reason. The current study provides lists of references of a collection of common diagnostic instruments for the interested readers of teachers or researchers and it is obvious that the number of three and four tier tests is still not adequate in all fields of science.

To conclude, there are several ways to diagnose students' misconceptions in science, but all diagnostic assessment methods have their own strengths and limitations. Table 8 summarizes these methods with their strengths and weaknesses based on the analysis of several research articles in the scope of this study. Researchers or teachers who aim to use them should be cautious about these concerns and use the more appropriate method or a method serving best for their purposes.

**Table 8.** Comparison of strengths and weaknesses of diagnostic methods

|                   | Interview  | Open-ended Test   | Methods to Diagnose Misconceptions  |   |   | Four-tier MCT  |
|-------------------|--|---|---|---|---|--|
|                   |  |   | Ordinary MCT  | Two-tier MCT  | Three-tier MCT  |  |
| <b>Strengths</b>  | -Provides in-depth information.<br>-Flexibility of questioning.  | -Gives responders the chance to write their answers in their own words.<br>-Responders may give new and valuable answers which are not thought by the researcher before.  | -Time efficient in administering.<br>-Immediately scored.<br>-Objectively scored.<br>-Validity evidence is strong.<br>-Applied to a large number of subjects.   | -Holds all the strengths provided with Ordinary MCT.<br>-Gives an opportunity to decide the proportions of false positives and false negatives. | -Holds all the strengths provided with Two-tier MCT.<br>-Determines the answers given to the first two tiers are due to misconception or a mistake due to lack of knowledge.  | -Holds all the strengths provided with Three-tier MCT.<br>-Truly assesses misconceptions which are free of errors and lack of knowledge. |
| <b>Weaknesses</b> | -Requires a large amount of time to obtain and analyze the data.<br>-Requires a large number of people to obtain a greater generalizability.<br>-Requires training in interviewing.<br>-Analysis of data is difficult and subjective.<br>-Responders not respond freely if not built trust to the interviewer. | -Takes time to analyze responses.<br>-Scoring is a problem.<br>-Response-rate is relatively small (students are resistant to write their response and reasoning clearly). | -Less usable for the measurement of psychomotor skills.<br>-Do not provide deep enough investigation into the students' ideas.<br>-Students may give correct answers for wrong reasons (False Positive) or wrong answer with correct reasons. (False Negative)<br>-Wrongly interpreting students' responses if the items have not been constructed carefully.<br>-Guessing.<br>-Difficult to develop a well-constructed item. | -Overestimates the proportions of the misconceptions since the lack of knowledge cannot be determined.  | -Underestimates the proportions of lack of knowledge since cannot decide whether the responder is sure for his/her answer in the first tier, in the second tier or in both tiers.<br>-Overestimates students' scores. | -Requires a longer testing time.<br>-Usefulness may be limited to diagnostic purposes.   |

MCT: Multiple Choice Test

### AUTHORS' NOTE

This study was completed as part of the first author's doctoral dissertation. The preliminary version of this study was presented at İSER 2014 World Conference.

### ACKNOWLEDGEMENT

The authors appreciate the financial support of the Scientific and Technological Research Council of Turkey (TÜBİTAK) and the Faculty Development Programme at Middle East Technical University (ÖYP) for this study. We would like to thank the Physics Education Group at the University of Washington for the opportunities they provide during the first author's doctoral dissertation process.

### REFERENCES

- Al-Rubayea, A. M. (1996). *An Analysis of Saudi Arabian High School Students' Misconceptions About Physics Concepts*. Unpublished doctoral dissertation, Kansas State University, Manhattan, Kansas.

- Adadan, E., & Savasci F. (2012). An analysis of 16-17-year-old students' understanding of solution chemistry concepts using a two-tier diagnostic instrument. *International Journal of Science Education*, 34(4), 513-544.
- Akkus, H., Kadayifci H., & Atasoy B. (2011). Development and application of a two-tier diagnostic test to assess secondary students' understanding of chemical equilibrium concepts. *Journal of Baltic Science Education*, 10(3), 146-155.
- Andersson, B., & Karrqvist, C. (1983). How Swedish peoples, aged 12-15 years understand light and its properties. *International Journal of Science Education*, 5(4), 387-402.
- Anderson, D. L., Fisher, K. M., & Norman, J. G. (2002). Development and validation of the conceptual inventory of natural selection. *Journal of Research in Science Teaching*, 39, 952-978.
- Arslan, H. O., Cigdemoglu, C., & Moseley, C. (2012). A three-tier diagnostic test to assess pre-service teachers' misconceptions about global warming, greenhouse effect, ozone layer depletion, and acid rain. *International Journal of Science Education*, 34(11), 1667-1686.
- Artdej, R., Ratanaroutai, T., Coll, R. K., & Thongpanchang, T. (2010). Thai grade 11 students' alternative conceptions for acid-base chemistry. *Research in Science & Technological Education*, 28(2), 167-183.
- Aydın, Ö. (2007). *Assessing tenth grade students' difficulties about kinematics graphs by a three-tier test*. Unpublished master thesis, Middle East Technical University, Ankara.
- Aykutlu, I., & Şen, A. İ. (2012). Üç aşamalı test, kavram haritası ve analogi kullanılarak lise öğrencilerinin elektrik akımı konusundaki kavram yanlışlarının belirlenmesi. *Education and Science*, 37(166), 275-288.
- Bardar, E. M., Prather, E. E., Slater, T. F., & Brecher, K. (2007). Development and validation of the light and spectroscopy concept inventory. *Astronomy Education Review*, 5(2), 103-113.
- Baser, M. & Geban, O. (2007). Effectiveness of conceptual change instruction on heat and temperature concepts. *Research in Science & Technological Education*, 25(1), 115-133.
- Bau-Jaoude, S. B. (1991). A study of the nature of students' understandings about the concept of burning. *Journal of Research in Science Teaching*, 28, 689-704.
- Beichner, R. J. (1994). Testing student interpretation of kinematics graphs. *American Journal of Physics*, 62(8), 750-762.
- Bork, A. (1984). Letters to the editor. *American Journal of Physics*, 52(10), 873-874.
- Caleon, I. S. & Subramaniam, R. (2010a). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, 32(7), 939-961.
- Caleon, I. S. & Subramaniam, R. (2010b). Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education*, 40, 313-337.
- Caramazza, A., McCloskey, M., & Green, B. (1980). Curvilinear motion in the absence of external forces: naïve beliefs about the motion of objects. *Science*, 210(4474), 1139-1141.
- Cataloglu, E. & Robinett, R. W. (2002). Testing the development of student conceptual and visualization understanding in quantum mechanics through the undergraduate career. *American Journal of Physics*, 70(3), 238-251.
- Cetin-Dindar, A. & Geban, Ö. (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.
- Chandrasegaran, A. L., Treagust, D. F. & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice*, 8(3), 293-307.
- Chang, C. Y., Yeh, T. K., & Barufaldi, J. P. (2010). The positive and negative effects of science concept tests on student conceptual understanding. *International Journal of Science Education*, 32(2), 265-282.
- Chang, H. P., Chen, J. Y., Guo, C. J., Chen, C. C., Chang, C. Y., Lin, S. Y., Su, W. J., Lain, K. D., Hsu, S. Y., Lin, J. L., Chen, C. C., Cheng, Y. T., Wang, L. S., Tseng, & Y. T. (2007). Investigating primary and secondary students' learning of physics concepts in Taiwan. *International Journal of Science Education*, 29(4), 465-482.

- Chen, S. M. (2009). Shadows: young Taiwanese children's views and understanding. *International Journal of Science Education*, 31(1), 59-79.
- Chen, C. C., Lin, H. S., & Lin, M. L. (2002). Developing a two-tier diagnostic instrument to assess high school students' understanding- the formation of images by plane mirror. *Proc. Natl. Sci. Counc. ROC(D)*, 12(3), 106-121.
- Cheong, I. P. A., Treagust, D., Kyeleve, I. J., & Oh, P. Y. (2010). Evaluation of students' conceptual understanding of malaria. *International Journal of Science Education*, 32(18), 2497-2519.
- Chiu, M. H. (2007). A national survey of students' conceptions of chemistry in Taiwan. *International Journal of Science Education*, 29(4), 421-452.
- Chu, H. E., Treagust, D. F., & Chandrasegaran, A. L. (2009). A stratified study of students' understanding of basic optics concepts in different contexts using a two-tier multiple-choice items. *Research in Science & Technological Education*, 27(3), 253-265.
- Clement, J., Brown, D. E., & Zietsman, A. (1989). Not all preconceptions are misconceptions: finding 'anchoring conceptions' for grounding instruction on students' intuitions. *International Journal of Science Education*, 11, 554-565.
- Colin, P., Chauvet, F., Viennot, L. (2002). Reading images in optics: students' difficulties and teachers' views. *International Journal of Science Education*, 24(3) 313-332.
- Coştu, B., Ayas, A., Niaz, M., Ünal, S., & Çalık, M. (2007). Facilitating conceptual change in students' understanding of boiling concept. *Journal of Science Educational Technology*, 16, 524-536.
- Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an assessment tool: brief electricity and magnetism assessment. *Physical Review Special Topics-Physics Education Research*, 2(1), 10105-1-10105-7.
- diSessa, A. A. (1993). Towards an epistemology of physics. *Cognition and Instruction*, 10(2&3), 105-225.
- Downing, S. M. (2006). *Twelve steps for effective test development*. In S.M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25), New Jersey: Lawrence Erlbaum Associates, Inc.
- Driver, R. & Easley, J. (1978). Pupils and paradigms: a review of literature related to concept development in adolescent science students. *Studies in Science Education*, 5, 61-84.
- Duit, R., Treagust, D. F., & Mansfield, H. (1996). *Investigating student understanding as a prerequisite to improving teaching and learning in science and mathematics*. In D. F. Treagust, R. Duit, & B. J. Fraser (Eds.), *Improving teaching and learning in science and mathematics* (pp. 17-31). New York: Teachers College Press.
- Engelhardt, P. V. & Beichner, R. J. (2004). Students' understanding of direct current resistive electric circuits. *American Journal of Physics*, 72(1), 98-115.
- Eryilmaz, A. (2010). Development and application of three-tier heat and temperature test: Sample of bachelor and graduate students. *Eurasian Journal of Educational Research*, 40, 53-76.
- Eshach, H. (2003). Small-group interview-based discussions about diffused shadows. *Journal of Science Education and Technology*, 12(3), 261-275.
- Fetherstonhaugh, A. & Treagust, D. F. (1992). Students' understanding of light and its properties: teaching to engender conceptual change. *Science Education*, 76(6), 653-672.
- Frankel, J. R. & Wallen, N. E. (2000). *How to design and evaluate research in education* (4th ed.). US: McGraw-Hill Comp.
- Franklin, B. J. (1992). *The development, validation, and application of a two-tier diagnostic instrument to detect misconceptions in the areas of force, heat, light and electricity*. Unpublished PhD Thesis, The Louisiana State University.
- Galili, I. & Goldberg, F. (1993). Left-right conversions in a plane mirror. *The Physics Teacher*, 31(8), 463-466.
- Goldberg, F. M. & McDermott, L. C. (1986). Student difficulties in understanding image formation by a plane mirror. *The Physics Teacher*, 24(8), 472-481.
- Goldberg, F. M. & McDermott, L. C. (1987). An investigation of student understanding of real image formed by a converging lens or concave mirror. *American Journal of Physics*, 55(2), 108-119.
- Greca, I. M. & Moreire, M. A. (2002). Mental, physical and mathematical models in the teaching and learning of physics. *Science Education*, 86 (1), 106-121.

- Griffard, P. B. & Wandersee, J. H. (2001). The two-tier instrument on photosynthesis: what does it diagnose? *International Journal of Science Education*, 23(10), 1039-1052.
- Gronlund, N. E. (1981). *Measurement and evaluation in teaching*. NY: McMillan Pub. Co. Inc.
- Hammer, D. (1996). More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for educational research. *American Journal of Physics*, 64(10), 1316-1325.
- Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education*, 34(5), 294-299.
- Haslam, F. (1986). *Secondary students' understanding of photosynthesis and respiration in plants*. Unpublished manuscript. Science and Mathematics Education Centre, Western Australian Institute of Technology, Perth, WA.
- Haslam, F. & Treagust, D. F. (1987). Diagnosing secondary students' misconceptions about photosynthesis and respiration in plants using a two-tier multiple-choice instrument. *Journal of Biological Education*, 21(3), 203-211.
- Helm, H. (1980). Misconceptions in physics amongst South African students. *Physics Education*, 15, 92-105.
- Hestenes, D. & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher*, 30, 159-166.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, 30, 141-158.
- Howitt, S. T., Anderson, M., Hamilton, S., & Wright, T. (2008). A concept inventory for molecular life sciences: How will it help your teaching practice? *Australian Biochemist*, 39, 14-17.
- Hrepic, Z. (2004). *Development of a real-time assessment of students' mental models of sound propagation*. Unpublished PhD thesis, Kansas State University, Manhattan, Kansas.
- Iona, M. (1982). Virtual mirrors. *The Physics Teacher*, 20, 278.
- Kalas, P., O'Neil, A., Pollock, C. & Birol, G. (2013). Development of a meiosis concept inventory. *CBE Life Sci Educ.*, 12 (4), 655-664.
- Kaltakci, D. & Didis, D. (2007). Identification of pre-service physics teachers' misconceptions on gravity concept: A study with a 3-tier misconception test. In S. A. Çetin, & İ. Hikmet (Eds.), *Proceedings of the American Institute of Physics*, USA, 899, 499-500.
- Kaltakçı, D. (2012). *Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics*. Unpublished PhD Thesis, Middle East Technical University, Ankara, Turkey.
- Kılıç, D. & Sağlam, H. (2009). Development of a two-tier diagnostic test to determine students' understanding of concepts in genetics. *Eurasian Journal of Educational Research*, 36, 227-244.
- Kızılcık, H. S. & Güneş, B. (2011). Developing three-tier misconception test about regular circular motion. *Hacettepe University Journal of Education*, 41, 278-292.
- Kirbulut, Z. D. & Beeth, M. E. (2013). Representations of fundamental chemistry concepts in relation to the particulate nature of matter. *International Journal of Education in Mathematics, Science and Technology*, 1 (2), 96-106.
- Kirbulut, Z. D. & Geban, O. (2014). Using three-tier test to assess students' misconceptions of states of matter. *Eurasia Journal of Mathematics, Science & Technology Education*, 10 (5), 509-521.
- Klammer, J. (1998). *An overview of techniques for identifying, acknowledging and overcoming alternative conceptions in physics education*. (Report no: ED423121). Columbia University. Retrieved from <http://www.eric.ed.gov/PDFS/ED423121.pdf>
- Klymkowsky, M. W. & Garvin-Doxas, K. (2008). Recognizing students' misconceptions through Ed's tools and the Biology Concept Inventory. *PLoS Biology*, 6, e3.
- Knight, J. K. & Wood, W. B. (2005). Teaching more by lecturing less. *Cell Biology Education*, 4, 298-310.
- Komorek, M. & Duit, R. (2004). The teaching experiment as a powerful method to develop and evaluate teaching and learning sequences in the domain of non-linear systems. *International Journal of Science Education*, 26(5), 619-633.
- Kutluay, Y. (2005). *Diagnosis of eleventh grade students' misconceptions about geometric optic by a three-tier test*. Unpublished master thesis, Middle East Technical University, Ankara.

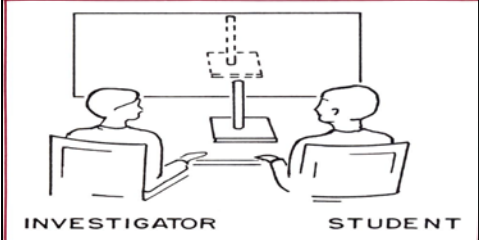


- Langley, D., Ronen, M., & Eylon, B. S. (1997). Light propagation and visual patterns: preinstruction learners' conceptions. *Journal of Research in Science Teaching*, 34(4), 399-424.
- La Rosa, C., Mayer, M., Patrizi, P., & Vicentini-Missoni, M. (1984). Commonsense knowledge in optics: preliminary results of an investigation into properties of light. *European Journal of Science Education*, 6(4), 387-397.
- Lin, S. W. (2004). Development and application of a two-tier diagnostic test for high school students' understanding of flower plant growth and development. *International Journal of Science and Mathematics Education*, 2, 175-199.
- Maloney, D. P., O'Kuma, T. L., Heiggelke, C. J., & Van Heuvelen, A. (2001). Surveying students' conceptual knowledge of electricity and magnetism. *American Journal of Physics*, 69(7), 12-23.
- Mann, M. & Treagust, D. F. (1998). A pencil and paper instrument to diagnose students' conceptions of breathing, gas exchange and respiration. *Australian Science Teachers Journal*, 44, 55-59.
- McDermott, L. C. (1993). How we teach and how students' learn-A mismatch?. *American Journal of Physics*, 61(4), 295-298.
- Monteiro, A., Nobrega, C., Abrantes, I. & Gomes, C. (2012). Diagnosing Portuguese students' misconceptions about the mineral concept. *International Journal of Science Education*, 34 (17), 2705-2726.
- Mulford, D. R. & Robinson, W. R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education*, 79(6),739-744.
- Odom, A. L. & Barrow, L. H. (1995). Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction. *Journal of Research in Science Teaching*, 32(1), 45-61.
- Olivieri, G., Torosantucci, G., & Vincentini, M. (1988). Coloured shadows. *International Journal of Science Education*, 10(5), 561-569.
- Osborne, J. F., Black, P., Meadows, J., & Smith, M. (1993). Young children's (7-11) ideas about light and their development. *International Journal of Science Education*, 15(1), 83-93.
- Osborne, R. & Freyberg, P. (1987). *Learning in science: the implications of children's science*. Auckland: Heinemann.
- Osborne, R. J. & Gilbert, J. K. (1979). Investigating student understanding of basic physics concepts using an interview-about-instances approach. *Research in Science Education*, 16, 40-48.
- Osborne, R. J. & Gilbert, J. K. (1980a). A method for investigating concept understanding in science. *European Journal of Science Education*, 2, 311-321.
- Osborne, R. J. & Gilbert, J. K. (1980b). A technique for exploring students' views of the world. *Physics Education*, 15, 376-379.
- Peşman, H., & Eryilmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *The Journal of Educational Research*, 103, 208-222.
- Peterson, R. F., Treagust, D. F., & Garnett, P. (1986). Identification of secondary students' misconceptions of covalent bonding and structure concepts using a diagnostic test instrument. *Research in Science Education*, 16, 40-48.
- Piaget, J. (1969). *The child's conception of physical causality*. New Jersey: Littlefield & Adams, Co
- Rollnick, M. & Mahooana, P. P. (1999). A quick and effective way of diagnosing student difficulties: two tier from simple multiple-choice questions. *South African Journal of Chemistry*, 52(4), 161-164.
- Ross, B. & Munby, H. (1991). Concept mapping and misconceptions: a study of high school students understanding of acids and bases. *International Journal of Science Education*, 13, 11-23.
- Sadler, P. M. (1998). Psychometric models of student conceptions in science: reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, 35(3), 265-296.
- Schmidt, H. J. (1997). Students' misconceptions- looking for a pattern. *Science Education*, 81(2), 123-135.
- Sesli, E. & Kara, Y. (2012). Development and application of a two-tier multiple choice diagnostic test for high school students' understanding of cell division and reproduction. *Journal of Biological Education*, 46(4), 214-225.

- Shi, J., Wood, W. B., Martin, J. M., Guild, N. A., Vicens, Q., & Knight, J. K. (2010). Diagnostic assessment for introductory molecular and cell biology. *CBE-Life Sciences Education*, 9, 453-461.
- Singh, C. & Rosengrant, D. (2003). Multiple-choice test of energy and momentum concepts. *American Journal of Physics*, 71(6), 607-617.
- Sokoloff, D. (1993). *Electric Circuit Concept Tests*. Unpublished study.
- Sokoloff, D. (1997). *Light and Optics Conceptual Evaluation*. Unpublished study.
- Sreenivasulu, B. & Subramaniam, R. (2013). University students' understanding of chemical thermodynamics. *International Journal of Science Education*, 35(4), 601-635.
- Tamir, P. (1989). Some issues related to the use of justifications to multiple-choice answers. *Journal of Biological Education*, 23, 285-292.
- Tamir, P. (1990). Justifying the selection of answers in multiple-choice items. *International Journal of Science Education*, 12(5), 563-573.
- Tan, K. C. D., Goh, N. K., Chia, L. S., & Treagust, D. F. (2002). Development and application of a two-tier multiple-choice diagnostic instrument to assess high school students' understanding of inorganic chemistry qualitative analysis. *Journal of Research in Science Teaching*, 39(4), 283-301.
- Thornton, R. K. & Sokoloff, D. R. (1998). Assessing student learning of Newton's Laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics*, 66(4), 338-352.
- Tongchai, A., Sharma, M. D., Johnston, I. D., Arayathanikul, K., & Soankwan, C. (2009). Developing, evaluating and demonstrating the use of a conceptual survey in mechanical waves. *International Journal of Science Education*, 31(18), 2437-2457.
- Treagust, D. (1986). Evaluating students' misconceptions by means of diagnostic multiple choice items. *Research in Science Education*, 16, 199-207.
- Tsui, C. Y. & Treagust, D. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education*, 32(8), 1073-1098.
- Tuysuz, C. (2009). Development of two-tier diagnostic instrument and assess students' understanding in chemistry. *Scientific Research and Essays*, 46, 629-631.
- Türker, F. (2005). *Developing a three tier test to assess high school students' misconceptions concerning force and motion*. Unpublished master thesis, Middle East Technical University, Ankara.
- Uzuntiryaki, E. & Geban, Ö. (2005). Effect of conceptual change approach accompanied with concept mapping on understanding of solution concepts. *Instructional Science*, 33, 311-339.
- Van Zee, E. H., Hammer, D., Bell, M., Roy, P., & Peter, J. (2005). Learning and teaching science as inquiry: a case study of elementary school teachers' investigations of light. *Science Education*, 89(6), 1007-1042.
- Voska, K. W. & Heikkien, H. W. (2000). Identification and analysis of student conceptions used to solve chemical equilibrium problems. *Journal of Research in Science Teaching*, 37(2) 160-176.
- Wandersee, J. H., Mintzes, J. J., & Novak, J. D. (1994). *Research on alternative conceptions in science*. In D. L. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp.177-210). New York: Macmillan.
- Wang, J. R. (2004). Development and validation of a two-tier instrument to examine understanding of internal transport in plants and the human circulatory system. *International Journal of Science and Mathematics Education*, 2, 131-157.
- White, R. & Gunstone, R. (1992). *Probing understanding*. London: The Falmer Press.
- Wittman, M. C. (1998). *Making sense of how students come to an understanding of physics: an example from mechanical waves*. Unpublished PhD thesis, University of Maryland



Appendix: Sample items of misconception diagnostic methods and their analysis

| Sample items of misconception diagnostic methods   | Method of analysis   |
|--|--|
| <p><b>Individual demonstration interview item</b><br/>                     Predict where the image would be located if the student were to move toward the investigator.</p>  <p style="text-align: right;"><i>(Goldberg &amp; McDermott, 1986)</i></p>   | <p>The main aim of the method is to find out how the student is thinking. Interviewer tries to find out what is on students' mind orally.</p>  |
| <p><b>Open-ended test item</b><br/>                     In a room perfectly sealed to external light there are some flowers in a vase. When a candle is lit in the room, one can see that the vase is white and that there is a red flower, a yellow flower, a purple flower, a pink flower, and some green leaves. What will we see after the candle is extinguished? Explain.<br/> <i>(Langley, Ronen &amp; Eylon, 1996)</i></p>   | <p>Students are allowed to write their own answers to the free-response test items.</p>  |
| <p><b>Ordinary multiple-choice test item</b><br/>                     Two-metal balls are the same size, but one weighs twice as much as the other. The balls are dropped from the top of a two story building at the same instant of time. The time it takes the balls to reach the ground below will be:<br/>                     a. About half as long for the heavier ball.<br/>                     b. About half as long for the lighter ball.<br/>                     c. About the same time for both balls.<br/>                     d. Considerably less for the heavier ball, but not necessarily half as long.<br/>                     e. Considerably less for the lighter ball, but not necessarily half as long.<br/> <i>(Hestenes, Wells, &amp; Swackhamer, 1992)</i></p>   | <p>Each distractor (wrong alternative) on the multiple-choice test item corresponds one of the pre-defined misconceptions. Someone who chooses a specific distractor is considered to have a corresponding specific misconception.</p> |
| <p><b>Two-tier multiple-choice test item</b><br/>                     1. The trait, curly hair, is dominant to straight hair. If we use "C" to represent the dominant allele (gene) for curly hair and "c" for the recessive allele, would a person with genotype Cc have curly hair?<br/>                     a. Yes   b. No   c. Don't know<br/>                     2. Reason for the above:<br/>                     a) The person needs to have CC for curly hair.<br/>                     b) The dominant allele C is expressed in a Cc condition.<br/>                     c) The person may or may not have curly hair.<br/>                     d) The recessive allele c is expressed.<br/> <i>(Tsui &amp; Treagust, 2010)</i></p>  | <p>Someone who chooses a wrong alternative in the first tier and chooses a related wrong reason in the second tier is considered to have a specific misconception.</p>   |
| <p><b>Three-tier multiple-choice test item</b><br/>                     1. The ozone layer,<br/>                     a) Protects the Earth from acid rain.<br/>                     b) Filters the ultraviolet (UV) rays of the sun.<br/>                     c) Helps to keep the Earth's temperature stable to make it livable.<br/>                     2. Which one of the following is the reason for your answer to the previous question?<br/>                     a) The ozone layer absorbs the sun's UV rays which is potentially damaging to life on the Earth.<br/>                     b) The ozone layer prevents sun rays to exit from the atmosphere, consequently keep it warm enough to live.<br/>                     c) The ozone layer works as a kind of shield, so it does not let acid rain to reach the Earth's surface.<br/>                     d) .....<br/>                     3. Are you sure about your answer given to the previous two questions?<br/>                     a. Yes   b. No   <i>(Arslan, Cigdemoglu, &amp; Moseley, 2012)</i></p> | <p>Someone who chooses a wrong alternative in the first tier, chooses a related wrong reason in the second tier and confident in answers given to the both tiers is considered to have a specific misconception.</p>                   |

**Four-tier multiple-choice test item**

1. Consider a real gas placed in a container. If inter-molecular attractions were to disappear suddenly, which one of the following is likely to occur?
  - a) The pressure decreases
  - b) The pressure increases
  - c) The pressure remains the same
  - d) The gas expands
2. Confidence rating for answer:
  - a. Just guessing
  - b. Very confident
  - c. Unconfident
  - d. Confident
  - e. Very confident
  - f. Absolutely Confident
3. Reason:
  - a. The gas molecules are further away from each other.
  - b. The frequency of collisions of the molecules with the walls of the container increases.
  - c. The gas molecules will have more freedom to move around.
  - d. It will behave more like an ideal gas.
4. Confidence rating for answer:
  - a. Just guessing
  - b. Very confident
  - c. Unconfident
  - d. Confident
  - e. Very confident
  - f. Absolutely Confident

*(Sreenivasulu & Subramaniam, 2013)*

Someone who chooses a wrong alternative in the first tier with high confidence and chooses a related wrong reason with high confidence is considered to have a specific misconception.