Developing a three-tier diagnostic instrument on chemical equilibrium (TT-DICE)

J. Jusniar[1], E. Effendy[2], Endang Budiasih[2] and S. Sutrisno[2]

Abstract

This study aims to develop a Three-tier Diagnostic Instrument on Chemical Equilibrium (TT-DICE) to diagnose high school students' misconceptions related to this topic. TT-DICE consists of thirty items, asking for students' answers, reasonings, and levels of confidence. Three-stages development models developed by Treagust applied in developing TT-DICE. First-stage was a literature study to analyze the main concepts and students' misconceptions in CE. The next step is the development of the TT-DICE prototype, consisting of item development, preliminary experiments, interviews, and revisions. The third stage of validity checking starts with content validation by experts and practitioners. A field test conducted to determine the reliability and quality variables are supporting TT-DICE items. These variables are Item Validity (IV), Difficulty Level (DL), Discrimination Index (DI), and effectiveness of distractors. Content validity by validators showed that the TT-DICE is valid with a very high category (96.1%). The results of item validity show that 88.9% of the items are valid. Testing of DL, DI, and the effectiveness of distractors in general, shows a good category. The Reliability of the three-tiers TTDICE using Alpha Cronbach is a high and very high category, respectively (0.83, 0.81 & 0.91). The overall validity test results also concluded if the TT-DICE is valid, reliable, and suitable to diagnose misconceptions of Chemical Equilibrium.

Desarrollo de un instrumento de diagnóstico de tres niveles sobre equilibrio químico (TT-DICE)

Resumen

Este estudio tiene como objetivo desarrollar un Instrumento de diagnóstico de tres niveles sobre equilibrio químico (TT-DICE) para diagnosticar los conceptos erróneos de los estudiantes de secundaria relacionados con este tema. TT-DICE consta de treinta elementos, que solicitan respuestas, razonamientos y niveles de confianza de los estudiantes. Modelos de desarrollo de tres etapas desarrollados por Treagust aplicados en el desarrollo de TT-DICE. La primera etapa fue un estudio de literatura para analizar los conceptos principales y los conceptos erróneos de los estudiantes en CE. El siguiente paso es el desarrollo del prototipo TT-DICE, que consiste en el desarrollo de ítems, experimentos preliminares, entrevistas y revisiones. La tercera etapa de verificación de validez comienza con la validación de contenido por parte de expertos y profesionales. Una prueba de campo realizada para determinar las variables de confiabilidad y calidad es compatible con los elementos TT-DICE. Estas variables son Validez del ítem (IV), Nivel de dificultad (DL), Índice de discriminación (DI) y efectividad de los distractores. La validez del contenido por parte de los validadores mostró que el TT-DICE es válido con una categoría muy alta (96.1%). Los resultados de la validez del artículo muestran que el 88.9% de los artículos son válidos. Las pruebas de DL, DI y la efectividad de los distractores en general muestran una buena categoría. La confiabilidad del TTDICE de tres niveles con Alpha Cronbach es una categoría alta y muy alta, respectivamente (0.83, 0.81 y 0.91). Los resultados de la prueba de validez general también concluyeron si el TT-DICE es válido, confiable y adecuado para diagnosticar conceptos erróneos de equilibrio químico.

[1] State University of Malang and State University of Makassar.
[2] State University of Malang.

Introduction

Concepts or topics in chemistry are hierarchically built from the basic into the more complex ones (Ealy, 2018; O'Connor, 2015; Seery, 2009). Therefore, a proper understanding of prerequisite concepts or topics is essential in chemistry learning, especially to comprehend the more advanced concepts or topics (Shing & Brod, 2016; Carey, 2010; Bilgin & Uzuntirkayi, 2003; Effendy, 2002). Hence, students who mastered the prerequisite concept properly tend to be easier to understand further related concepts. Improper understandings or inability to link prerequisite concepts with the new ones will cause difficulties in understanding the new concepts (Taber, 2015; Ambrose *et al.*, 2010; Taber, 2009). These difficulties tend to form perceptions consistent with his new understanding (Osborne & Wittrock, 1983).

Chemical knowledge, especially chemical phenomenon, is generated, expressed, taught, and communicated at three levels of representations, namely macroscopic, submicroscopic, and symbolic. It has been one of the most powerful and productive ideas in chemical teaching for three decades (Talanquer, 2011; Gilbert & Treagust, 2009). The macroscopic representation relates to chemical processes mostly observed with our eyes. The submicroscopic relate to macroscopic phenomena at a particulate level that is mostly abstract (Talanquer, 2010; Cook *et al.*, 2008; Johnstone, 2000). The symbolic representations are mostly the translations of macroscopic and submicroscopic in the forms of symbols, formulas, and equations. Hence, students are required to have a high level of understandings of submicroscopic and symbolic representation to comprehend the macroscopic phenomena (Stojanovska *et al.*, 2017; Talanquer, 2011; Johnstone, 2000). Arroio (2016) argues that the students face many difficulties in operating at all the representational levels. Talanquer (2011) stated that Students' inability to interrelate these three levels of representations is another cause of students' difficulties in understanding concepts in chemistry.

Difficulties in learning concepts in chemistry can result in incorrect understanding. This incorrect understanding that occurs consistently will cause misconceptions. Misconceptions are the understanding of concepts that disconfirm the views of the experts (Barke *et al.*, 2009; Nakhleh, 1992). The misunderstanding may occur due to several factors, such as prior knowledge (Durmaz, 2018; Taber, 2015), ineffective communication (Dhindsa & Treagust, 2014; Johnstone, 2010), insufficient information by teachers and limited textbook content (Erman, 2017; Devetak *et al.*, 2010; Garnet *et al.*, 1995; Peterson & Treagust, 1989), insufficient conceptual understanding (Alghazo & Alghazo, 2017; Caleon & Subramaniam, 2010), abstract and symbolic properties of chemical concepts (Yakmaci-Guzel, 2013), and student preconceptions (Barke *et al.*, 2009; Horton, 2007). The problems would be, as highlighted by Taber (2011), that students could potentially perceive new improper knowledge from their misconceptions.

Chemical Equilibrium (CE) is one of the chemistry topics taught to high school students. Several studies have successfully identified misconceptions among students on this topic. Some misconceptions reported are as follows. The forward reaction rate becomes faster at the equilibrium (Niaz, 1998a; Hackling & Garnett, 1985). The state of equilibrium occurs when the concentrations of the reactants and products are the same (Yakmaci-Gusel,

2013; Barke *et al.*, 2009; Özmen, 2008). The rise in temperature would cause the rate of the forward reaction to decreases, at the same time, increases the reverse reaction (Barke *et al.*, 2009; Bilgin & Uzuntiryaki, 2003; and Hackling & Garnet, 1985). Equilibrium is a static process (Yakmaci-Gusel, 2013 and Barke *et al.*, 2009). The rate of the forward reaction will decrease in exothermic reactions if increasing the temperature system (Sozbilir, 2010 and Banerjee, 1991). Adding a reactant to the gas equilibrium system will shift the equilibrium towards products (Karpudewan *et al.*, 2015). Catalysts would cause an increase in product concentration (Bilgin & Uzuntiryaki, 2003; Voska & Heikkinen, 2000; Gorodetsky & Gussarsky, 1986; and Hackling & Garnett, 1985). In a nutshell, misconceptions have been observable in almost all the concepts of CE.

To identify misconceptions about the CE topic, various methods, as well as instruments, have been employed in some studies. Hackling & Garnett (1985) and Cheung *et al.* (2009), for instance, chose to conduct interviews to allow them to investigate misconceptions thoroughly. The method, however, required a long time with a limited number of sample involvements (Chandrasegaran *et al.*, 2007), and the analysis would be rather difficult and complicated (Adadan & Savasci, 2012). Another method used was a multiple-choice test, as performed by Bilgin and Uzuntiryaki (2003) and Nakhleh (1992). This type of test allows larger samples, easy to analyze, and covers broad generalizations (Beichner, 1994). Unfortunately, it is unable to reveal students' reasoning supporting their answers (Peşman & Eryilmaz, 2010), which potentially allows guessing answers (Gurel *et al.*, 2015). To avoid the possibility of guessing answers, Banerjee (1991) combined the multiple-choice test with short solutions in the instrument he developed. The design was good but was not enough to reveal students' explanations nor reasonings for choosing their answers. A concept map was proposed by Novak (1990) to explore students' conceptual understanding and by Hay & Kinchin (2006) to uncover conceptual typologies in science. However, the method seems quite hard to conduct as it requires students to have a good mastery of hierarchical vocabularies to express their ideas logically (Kinchin, 2000). In addition to the methods mentioned above, Treagust (1988) developed a two-tier instrument to identify misconceptions in science subjects. In chemistry, Peterson and Treagust (1989) have developed a similar kind of test for the concept of chemical bonds, and Chandrasegaran *et al.* (2007) for the chemical reaction equations topic. Items within this kind of test consist of two levels, with multiple-choice or true-false type in the first tier and choices of causal reasons in the second tier. However, although the two-tier tests have provided ideas to clarify students' answers, they are still unable to distinguish students who experienced misconceptions and those having insufficient understandings or lack knowledge (Arslan *et al.*, 2012; Hasan *et al.*, 1999).

The limitations possessed by each method described above require that a method be designed to overcome all existing weaknesses and be more practical to reveal misconceptions. Design a three-tier instrument by adding a third-tier to two-tier test (Dindar & Geban, 2011). The third-tier was asking for students' level of confidence (LC) when answering the first and second tiers from each item. Such an action would provide certainty response index (CRI) that can help distinguish between students' suffering from misconceptions and those who are simply lack of understanding.

Yang & Sianturi (2019), Arslan *et al.* (2012), and Hasan *et al.* (1999) used the three-tier instrument to classify students' conceptual understanding into three categories, namely, (1) scientific knowledge, (2) misconceptions, and (3) insufficient opinions. The first category belonged to those having correct answers in their answer (first-tier) and

reasonings (second-tier) and was also sure about their responses (third-tier). The second category divided into three subcategories, that is, specific misconceptions (incorrect answers, wrong reasons, sure), false-negatives (correct answers, incorrect reasons, sure), and false-positive (incorrect answer, correct reasons, sure). The third category covered students with wrong answers as well as reasonings and was only do guessing when performing the test. The three-tier instrument has been shown in some fields but is still absent from the topic of Chemical Equilibrium (CE). This research is intended to develop a valid and reliable three-tier diagnostic instrument on the topic of Chemical Equilibrium abbreviated with TT-DICE.

Research method

This research is a developmental study aiming at developing a TT-DICE test used to identify students' misconceptions on the topic of CE. The development of this test adopts three stages of test development formulated by Treagust (1988). These three stages are analysis of misconception propositions, a prototype of test development, and validity checks. The study employed a mixed-method approach combining qualitative and quantitative data analyses. The qualitative approach is used to describe the process of developing TT-DICE prototypes based on misconception propositions reported in the articles and other sources. The quantitative method is used to assess content validity, item validity, and reliability of the designed instruments.

The three stages of the development of TT-DICE were as follows. The first stage, analysis of misconception propositions, was conducted through a literature study. This included formulate concept analysis, identify common misconceptions, and formulate misconception propositions on the topic of CE. The second stage, a prototype of test development, included preparation of 30 three-tier items, initial trials, interview to six student-participating in the trial, and revision of item tests. The third stage, validity check, included determination of content validation judged by three lectures on chemistry learning and three practitioners, revisions, and then field testing to check off item validity (IV), difficulty level (DL), discrimination index (DI), the effectiveness of distractor, and reliability TT-DICE.

TT-DICE was a three-tiers test consisting of 30 items. Some of the items tests were based on misconception reported by Barke *et al.* (2009), Özmen (2008), Tyson *et al.* (1999) Banerjee (1991), and Hackling & Garnet (1985). The first-tier of each item test was a multiple-choice question with four different possible answers. The second-tier was four possible reasons related to choice in the first tier. The third tier related to students' level of confidence (LC). The students' choosing LC about answers and reasons. The level of confidence consisted of three categories, namely, sure, not sure, and guessing. The correct answer for the first and second-tier was given a score of one. The incorrect answer was given a score of zero. Guessing answer of the third tier was given a score of zero, not sure answer was given a score of one, and the sure answer was given a score of two.

The LC in the third tier of TT-DICE was intended to distinguish between students giving an incorrect answer because of insufficient understanding with students having misconceptions. Students were categorized as inadequate understandings if the choices in the first and second tiers were incorrect, or one of the answers of both levels was incorrect with an LC were "guessing" or "not sure." Like students who lack understanding, students experienced misconceptions chose the "sure" answer for LC in the third-tier. Students were

categorized to understand the concept if the answer and reason were correct and chose the "sure" for LC in the third-tier. In contrast, students were classified to understand but lack confidence if the answers and reasons were correct and chose the "not sure or guess" for LC in the third-tier.

This study involved 30 first-year of a bachelor chemistry education program at Makassar State University in the initial trial of the TT-DICE prototype. Following improvement, the content validity of TT-DICE has judged six validators. Field testing was conducted involving 111 middle school students studying CE in their chemistry class.

In the initial trials, there were fewer samples involved. The reason for the initial trial was to explore the coherence of the answers and the reasons for exposing the misunderstanding. Tracing the suitability between the answers and the reasons was done by interview. Field testing was designed to test the quality of TT-DICE items. Variables were Discrimination Index (DI), Difficulty Level (DL), Distractor Effectiveness (DE), Item Validity (IV), and Reliability. In terms of validity, assessment criteria were used, as shown in Table 1 for content validity criteria, and Table 2 for field testing for DI, DL, IV, and reliability criteria.

Table 1 The category of content validity

Validity (%)	Category
81–100	Very high
61–80	High
41–60	Moderate
21–40	Low
0–20	Very low

Table 2 The criteria used to interpret the item analysis aspects

Difficulty Level (DL)		Discrimination Index (DI)		Reliability		Item validity	
Value	criteria	value	Criteria	Value	Criteria	$R_{table} = 0{,}157$	Criteria
0.00–0.30	Difficult	0.03 - <0.10	Poor	> 0,90	Excellent	$r_{count} \geq r_{table}$ (0,157)	Valid
		0.10 - <0.30	Moderate	0.81–0.90	Very good		
0.31 –0.70	Moderate	0.30 - <0.75	Good	0.61–0.80	Good	$r_{count} < r_{table}$	Invalid
				0.40–0.60	Moderate		
0.71 – 1.00	Easy	0.75 -1.00	Excellent	< 0.40	Poor		

Results and discussion

First Stage: literature study to analyse the main concept and misconception

Results of a literature study on CE misconceptions reported by previous research were given in Table 3.

Table 3 Main concepts and misconceptions propositions of CE

Main concepts	Misconception identified
Equilibrium state	At equilibrium concentrations of reactants and products are equal (Hackling & Garnett, 1985)
Dynamic equilibrium	Equilibrium is a static process (Yakmaci-Guzel, 2013; Barke *et al.*, 2009) The rate of the forward reactions increases by time, starting when the reactants are mixed until the equilibrium is established (Hackling & Garnet, 1985) At equilibrium, the sum of the concentrations of the reactants is equal to the products (Barke *et al.*, 2009; Özmen, 2008)
Effect of temperature on the equilibrium constant	The equilibrium constant (K_{eq}) increases as the temperature of an exothermic reaction increases (Özmen, 2008) As the temperature in an exothermic reaction decreases, the rate of the forward reaction will increase (Banerjee, 1991; Hackling & Garnett, 1985)
Effect of catalyst on the equilibrium system.	The rates of the forward and reverse reactions could be affected differently when the catalyst added (Özmen, 2008) Catalysts can affect the rates of the forward and reverse reactions separately (Hackling & Garnett, 1985) Catalysts cause an increase in product concentration (Bilgin & Uzuntiryaki, 2003; Voska & Heikkinen, 2000; Gorodetsky & Gussarsky, 1986; and Hackling & Garnett, 1985)
The equilibrium constant	The larger the value of K, the faster the forward reaction (Bilgin & Uzuntiryaki, 2003; Hackling & Garnet, 1985)
The effect of pressure on the gases equilibrium system	When gas volume decreased, equilibrium is reestablished, the equilibrium constant is higher than under its initial conditions (Hackling & Garnet, 1985) As the volume decreased, the rate of the reverse reaction also decreased (Hackling & Garnet, 1985)
The effect of concentration on the equilibrium system	When a substance added to an equilibrium mixture, the equilibrium will shift to the side of addition (Özmen, 2008) Adding a reactant to the gas equilibrium system will change the equilibrium towards products (Karpudewan *et al.*, 2015)

Based on the literature study, there were fifteen misconceptions related to nine main concepts in the CE material. These misconceptions were used as a basis for preparing TT-DICE.

Second Stage: TT-DICE prototype development

There were 30 items of TT-DICE prototype prepared bases on misconceptions given in Table 3. TT-DICE was a three-tier test. First-tier was a multiple-choice test with four possible answers. Second-tier was a reason for the choice of answer on the first tier. There were four possible reasons given in the second tier. Third-tier was about students' level of confidence in selecting answers in tier one and two. There were three choices of the level of confidence provided, namely guess, not sure, and sure.

Of the total 30 items (see Appendix 1b), 15 items were prepared based on misconceptions given in Table 3, and 15 items were prepared based on the learning outcome of CE topic contained in the Indonesian senior high school chemistry curriculum. Examples of items prepared based on Table 3 were given in Table 4, whereas examples of items prepared based on Indonesian senior high school chemistry curriculum were given in Figure 1.

An initial trial of a prototype of items test was carried out on 30 first-year chemistry education program students.
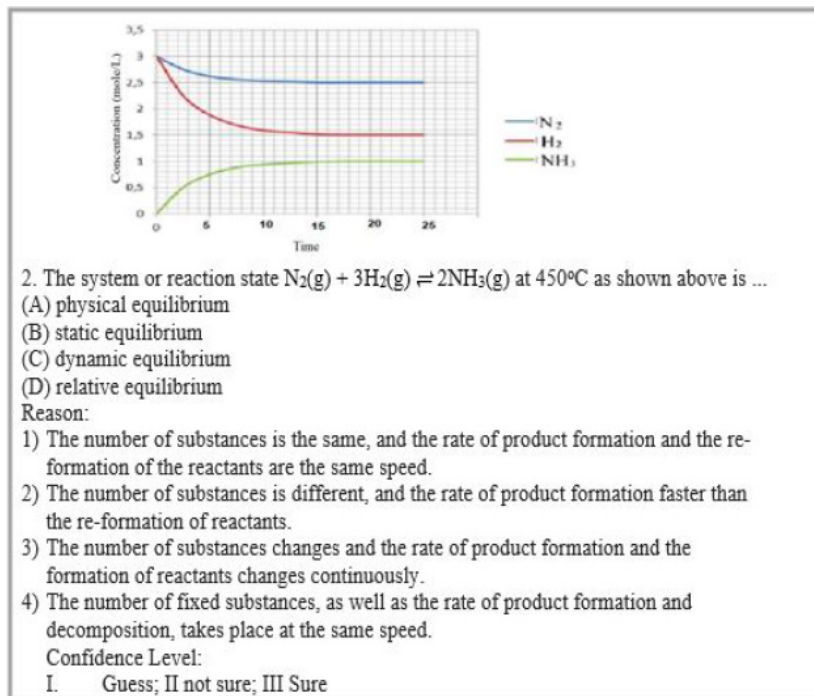
2. The system or reaction state $N_2(g) + 3H_2(g) \rightleftharpoons 2NH_3(g)$ at 450ºC as shown above is ...
(A) physical equilibrium
(B) static equilibrium
(C) dynamic equilibrium
(D) relative equilibrium
Reason:
1) The number of substances is the same, and the rate of product formation and the re-formation of the reactants are the same speed.
2) The number of substances is different, and the rate of product formation faster than the re-formation of reactants.
3) The number of substances changes and the rate of product formation and the formation of reactants changes continuously.
4) The number of fixed substances, as well as the rate of product formation and decomposition, takes place at the same speed.
Confidence Level:
I.   Guess; II not sure; III Sure

Figure 1. Example of item developed by the researcher

Table 4 Example of items adapted from previous studies

No	Questions	Reference
23	At a certain temperature, sulfur dioxide, and oxygen gas react to form sulfur trioxide, and equilibrium occurs according to the reaction: $2SO_2(g) + O_2(g) \rightleftharpoons 2SO_3(g) \ \Delta H = -197.78 \text{ kJ}.$ If a catalyst added to the equilibrium system, then the ratio of the forward reaction rate to the reverse reaction rate will be (A) greater than 1 (> 1) (B) less than 1 (<1) (C) equal to 1 (= 1) (D) equal to 0 (= 0) Reason: 1) Catalysts can increase collisions between reactant molecules and produce more products. 2) Catalysts reduce the activation energy to form the product and will again react at the same rate. 3) Catalysts increase activation energy so that the reaction rate progresses faster than the reverse reaction. 4) Catalysts do not affect activation energy, so fewer products formed. Confidence Level: I. Guess; II. Not sure; III. Sure	(Özmen, 2008) CERP (Hackling & Garnet, 1985). IJSE

Results of the Initial Trial and Interview

The objective of the initial trial is used to find out the legibility and usability of TT-DICE in identifying misconceptions about CE material. The trial revealed some students' misconceptions, especially ones related to an equilibrium state, dynamic equilibrium, and shift in equilibrium. Some of the misconceptions related to equilibrium state identified are *(1) the rate of an increasing number of products is lower than that of the reactants, (2)*

the forward reaction rate is faster than the reverse reaction, and (3) the concentration of reactants and products are the same. In the case of dynamic equilibrium, students argued that the number of reactants and product changes because the rate of the forward and reverse reaction also changed. Here, students might see the word "dynamic" as "changing," and thus, they regarded dynamic equilibrium as the forward reaction and the reverses that change accordingly.

Some of the misconceptions related to shifting of equilibrium identified are *(1) catalyst increases activation energy so that the rate of the forward reaction is faster than that of the reverse one. (2) The number of subscript of an element in the reactant and product in the gas equilibrium system affects the shift in equilibrium due to changes in volume. (3) in an exothermic gas equilibrium system, an increase in temperature will shift the equilibrium to the product. (4) in heterogeneous equilibrium, the addition of solid reactants will shift the equilibrium to the product. (5) The heterogeneous equilibrium constant is the result of the concentration of the product with the reactants raised by each coefficient. (6) addition of inert gas to the equilibrium with the same number of moles of reactants and products will shift the equilibrium.* The findings of misconceptions in the initial trial and the field test are given in Appendix 2.

Some statements from student interviews can take into consideration in revising options on answers and reasons. One example of the item 26 results from an interview with (SY) student who argues that in the $H_2(g) + Br_2(g) \rightleftharpoons 2HBr(g)$ equilibrium, the decreasing volume of gas shift equilibrium to the right. The reason is that the decrease in the volume of the system will increase the concentration, so that it will shift to the small number of moles, namely the right side. According to SY, the number of moles on the right side is two moles, and on the left side, there are four moles. This student seems to be inconsistent in distinguishing between subscripts with coefficients that determine the number of moles.

The results of the interview with the "AH" students in item two (Figure 1) suggest the correct answer for the reaction $N_2(g) + 3H_2(g) \rightleftharpoons 2NH_3(g)$ as a dynamic equilibrium with the answer pattern C3-III. The student's answer is correct, the reason is wrong and certain. AH, interview results say that the word dynamic means change. So, the amount of substances in the system changes the rate of the forward, and the rate of reverse reaction also changes. The correct concept is that the dynamic equilibrium concentration of substances in the system is constant, where the rate of forward reaction is equal to the rate of reverse reaction (Effendy, 2007).

Results of the initial trial indicate that 30% of questions were not valid in their first-tier. These items are Q4, Q6, Q7, Q8, Q11, Q12, Q13, Q21, and Q26. In the second-tier, there 46% items are not valid, namely Q3, Q4, Q7, Q9, Q11, Q13, Q14, Q17, Q19, Q21, Q22, Q23, Q27, and Q29. Items are not valid in the first and second-tier revised by simplifying the language in answers and reason provided. The small number of samples has an effect on the value of the r xy table. The greater number of samples, the value of r will be smaller (Arikunto, 1998). This weakness is then refined in field testing by increasing the number of samples.

The Third Stage: Validity Checks

This stage discusses the results of the testing of TT-DICE test. Discussion includes content validity, DL, DI, the effectiveness of distractor, IV, and reliability. Discussion is also about the feasibility and usefulness of TT-DICE.

Results and Analysis of Content Validation

Content validity was judged by three experts and three practitioners to ensure the quality of the prepared items of TT-DCE. This type of validity addresses how well the things developed to operationalize a construct provide an adequate and representative sample of all the items that might measure the construct of interest (Kimberlin & Winsterstein, 2008). Content validity usually depends on the judgment of experts in the field. The content validity examined (1) the correctness of the items, (2) the suitability of aspects measured by the questions, and (3) elements of language used in the questions. The assessment criteria are presented in Table 1. The result, given in Table 5, showed content validity of TT-DICE was very high with a score of 96.1%. The consistency of evaluation among validators was 88.9%, including the best category. Percentage agreements determined the flexibility of assessment between validators. It is often called the reliability coefficient (R). Borich (1994) argues if R ≥ 75% could be classified as a good percentage of agreement by validators. In other words, the TT-DICE believed based on decisions between validators.

Both the experts and practitioners claimed that the TT-DICE has successfully met the criteria, such as: (1) the items have strictly followed indicators of concepts or misconception propositions; (2) the language is clear and easy to understand; (3) the thoughts measured are conceptually and logically correct. Remarks from two validators asked for a minor revision in the language of four items, such as item number 15, 16, 22, and 28. Furthermore, we revised several questions for improvement ambiguity in the content of questions that makes each question sufficiently clear. To ensure that issues, answers, and reasons are representative and not outside the scope of the CE materials in the Curriculum for eleventh-grade students.

Table 5. Result of content validity of TT-DICE

Instrument	Validator Assessment (%)						Average (%)	Validity
	1	2	3	4	5	6		
TT-DICE	96.7	96.7	96.7	95.0	95.0	96.7	96.1	Very high

The result of the field test for TT-DICE interpreted according to Table 2. The variables tested were the Difficulty Level (DL) of the items (see Table 6). The Discrimination Index (DI) of the questions expressed in Table 6. The effectiveness of the distractor showed in Table 7 and the item validity of the TT-DICE at each level in Table 8. The reliability of the instruments of each level showed in Table 9.

Result and Analysis of the Difficulties Level

Data in Table 6 showed that the TT-DICE items are proportionally distributed in terms of their difficulties in each tier. The first-tier contained eight natural items (Q1, Q2, Q3, Q7, Q8, Q19, Q28, and 29), eight delicate items (Q13, Q14, Q15, Q21, Q24, Q25, Q26, and Q27), the other 14 moderate items, with the average level of difficulty rate is about 0.49. Meanwhile, for the second-tier, there are four natural items (Q3, Q7, Q25, and Q28), nine delicate questions (Q13, Q14, Q15, Q17, Q20, Q21, Q22, Q24, and Q28), and 17 moderate items, with the average level of difficulty, reached 0.46. At last, in the third-tier, 16 items are relatively easy, and 14 pieces are classified as moderate, and the average level of difficulty is around 0.67, which is in the easy category. With these results, the proportionality condition in terms of DL has been fulfilled.

As observed, the second tier items seem to have a higher level of difficulty compared to that of the first tier items. This might be due to the disclosure of reasons in the second tier that requires the ability to explain causal relationships rather than in the first tier as limited as revealing detailed answers, as described by Caleon & Subramaniam (2010). That items in the first-tier evaluate descriptive knowledge while in the second-tier evaluate explanatory knowledge. While for the third level, LC options in the form of confidence in the answers and reasons do not contain the understanding of the concept, so the determination of DL does not significantly affect the quality of the item.

CRI, as a feature of multi-tiered instruments termed LC in the TTDICE, can clearly categorize the level of student understanding based on the pattern of answers. For example, the model of solutions students have a scientific knowledge for item 2 in Figure 1 is C4-III. The student understands with high confidence that the reaction $N_2(g) + 3H_2(g) \rightleftharpoons 2NH_3(g)$ is a dynamic equilibrium because the amount of the substances is constant and the rate of product formation and decomposition, takes place at the same speed. Students have experience misconceptions had a C3-III answer pattern. The student believes that equilibrium is called dynamic because the amount of substances varies, and the rate of product formation and reactant re-formation also changes. Students who lack knowledge have answer patterns B1-II. Students uncertainly understand that the equilibrium is static because the amounts of the same substances and the rate of product formation and decomposition take place at the same speed.

Result and Analysis of the Discrimination Indexes (DI)

Tuckman & Harper (2012) and Arikunto (1998) explained that the discrimination index (DI) of an item represents how well the thing can distinguish between the upper and lower group students who gave the right answer for each question. A high DI value indicates the better quality of the item in identifying between the top and the low achievers. The DI an item can be measured by computing its discrimination coefficient, which is the correlation between examinees' overall test scores and the scores that they have obtained on the question under consideration DiBattista & Kurzawa (2011). Tuckman & Harper (2012) suggested the value of DI above 0.20 recommended as a useful item. As shown in Table 6, all the questions in the TT-DICE for three-tier have a relatively good DI value ranging from the moderate category into the perfect ones, and none of them having a bad or negative DI value. Besides, there are 43.3% of the items have either good or very good in all the three tiers, indicating their consistencies in distinguishing the high-and the low achievers in their answers, reasons, and levels of confidence. Meanwhile, the other items also have considerably right consistency in either sufficient or good categories.

The parameters of the DL and DI confirm that the answers and reasons for the TTDICE questions can be used to identify students' understanding of CE. TT-DICE function is twofold, first to analyze misconception understandings based on students' LC choices as there are categories. Secondly, determine students' understanding of concepts or cognitive learning outcomes carried out without consideration of the LC at the third-tier. Especially for LC at the third -tier, the value of DI, as well as DL in the previous presentation is less meaningful because the function of choice is to categorize student understanding. The DI of item depends heavily on the quality of its distractors (DiBattista & Kurzawa, 2011).

Tabel 6 Difficulty level and discrimination index of TT-DICE

Item	Difficulty Level			Discrimination Index			Item	Difficulty Level			Discrimination Index		
	Tier 1	Tier 2	Tier 3	Tier 1	Tier 2	Tier 3		Tier 1	Tier 2	Tier 3	Tier 1	Tier 2	Tier 3
1	0.71	0.54	0.91	0.38	0.54	0.45	16	0.63	0.43	0.71	0.34	0.75	0.73
2	0.75	0.39	0.91	0.53	0.63	0.45	17	0.39	0.25	0.73	0.28	0.18	0.60
3	0.91	0.79	0.89	0.38	0.32	0.44	18	0.46	0.56	0.41	0.52	0.55	0.89
4	0.37	0.41	0.79	0.45	0.18	0.47	19	0.78	0.51	0.63	0.34	0.36	0.70
5	0.64	0.70	0.72	0.58	0.68	0.14	20	0.51	0.29	0.79	0.54	0.80	0.25
6	0.65	0.61	0.91	0.31	0.11	0.45	21	0.19	0.15	0.39	0.18	0.15	0.68
7	0.84	0.75	0.89	0.22	0.23	0.38	22	0.52	0.30	0.50	0.43	0.11	0.58
8	0.79	0.47	0.64	0.23	0.51	0.47	23	0.37	0.45	0.58	0.46	0.44	0.71
9	0.63	0.70	0.80	0.36	0.11	0.61	24	0.18	0.15	0.44	0.75	0.62	0.29
10	0.57	0.70	0.88	0.53	0.68	0.44	25	0.21	0.79	0.41	0.58	0.35	0.64
11	0.40	0.19	0.81	0.30	0.23	0.54	26	0.18	0.33	0.84	0.49	0.50	0.50
12	0.59	0.45	0.66	0.13	0.37	0.73	27	0.15	0.60	0.46	0.69	0.38	0.73
13	0.26	0.10	0.48	0.29	0.15	0.69	28	0.77	0.24	0.47	0.61	0.49	0.48
14	0.27	0.23	0.87	0.24	0.24	0.10	29	0.78	0.80	0.63	0.46	0.11	0.57
15	0.14	0.43	0.47	0.66	0.75	0.51	30	0.19	0.55	0.54	0.33	0.10	0.87
	-	-	-	-	-	-	avarage	0.49	0.46	0.67	0.42	0,39	0.54

Note:

	DL easy		DI very good
	DL moderate		DI good
	DL difficult		DI moderate

Result and Analysis of Effectiveness of the Distractors

Each item on TT-DICE consists of questions, answers, reasons, and the level of confidence as a whole. However, what was analyzed for this distractor was only the pattern of students' answers and reasons. The level of confidence is not analyzed because it is only needed for categorizing student conceptions. The student's task is to choose an option that is the best answer to the questions asked in the one tier, or the best reason for the answer in tier two. The best answer is called the keyed option, and the remaining choice is called distractors.

Distractors in a multiple-choice question are tricky answers in addition to the correct answer in the options. A good distractor will make at least a 5% testee choosing it. Otherwise, it is considered a bad distractor DiBattista & Kurzawa, 2011. In the TT-DICE, we provided three distractors for each item in the first and second tier, and thus, there are 90 distractors overall for each tier. The try out revealed if 14.4% distractors in first-tier are classified bad, while the rest are considered good. Meanwhile, in the second-tier, there are 15.6% bad distractor reasons, and 84.4% are classified good (see Table 7). In TT-DICE, it is more important to uncover potential misconceptions with a consistent pattern between the answers and the reasons chosen. Distractors that are not selected by Testee will be reconsidered. When deciding whether a question should be revised or replaced the values of all parameters should be considered. In some circumstances, even a question with a bed distractor can be retained because the primary purpose of this instrument is to identify students' understanding instead of differentiating between high low achieving students (Suruchi & Rana, 2014).

The effectiveness of the distractor does not need to be tested in LC (Tier-3) because of the choice of options to categorize the level of understanding of students whether understanding concepts, misconceptions, and insufficient knowledge of concepts. The LC used on TT-DICE is more straightforward, namely three choices (guess, not sure, and sure) with a score range of 0-2. This range of grades is not ordinal; for example, students choosing "sure" with a score of 2 may be included in the category of conceptual understanding or misconception. This depends on the choice of answers and reasons. If both are true and sure, then the concept is understood to be categorized, but if at two levels, the responses and ideas are wrong and confident, then misconception is classified. LC used in TT-DICE is more practical to use than has been used by previous researchers such as Hasan *et al.* (1999) and Caleon & Subramaniam (2010) using the term Certainty Response Index (CRI) on a scale of 0-5.

Table 7. Effectiveness of the Distractor (%) for Each Item TT-DICE

Item		1		2		3		4		5		6		7		8	
Option		T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
T1	T2	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
A	1	9	32,5	1.8	7.2	3.6	9.0	12.6	27.0	62.2*	6.3	0	6.3	63*	5.4	64*	52.2*
B	2	71*	54*	16	57	81.0*	12.6	32.4	9.1	22.5	23.4	9	75*	26	12*	27	11.7
C	3	5	4,5	79*	6.3	9.0	73*	18.1	24.3	15.3	70.3*	66*	6.3	7.2	27	3.6	27.1
D	4	15	9	4.5	30*	5.4	5.4	36.9*	34.2*	0	0	25	14	3.6	9	5.4	9

Item		9		10		11		12		13		14		15		16	
Option		T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
T1	T2	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	
A	1	39.6	1.8	66*	6.3	5.41	9	0	5.4	14	6.3	64	31.5	7.2	9.0	51.4*	43.2*
B	2	53.2*	82*	14	35	41.4	11.7*	28	24	19	35.1	27*	29.7	55	0	33.3	27.0
C	3	7.21	16.2	13	45*	0	15.3	59*	45*	41	40.5	3.6	38.8*	13	43.2*	6.3	20.7
D	4	0	0	7.2	18	53.2*	64	13	25	26*	18.1*	5.4	0	11*	33.3	9.0	9.1

Item		17		18		19		20		21		22		23		24	
Option		T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
T1	T2	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
A	1	5	58.6	46*	6.3	78*	0	56	8.1	19*	59.5	28.8	29.7*	46.8	33.3	34	30.5

B	2	40*	27	22	55.9*	17	51.4*	32*	28.8*	49	9	52.3*	36.9	6.3	45*	4.5	6.3
C	3	51	14.4*	27	31.5	5	21.6	6	4.5	27	15.3*	14.4	33.3	37.8*	12.6	24*	16.2*
D	4	4	0	5	6.3	0	27.0	5	58.6	5	0	4.5	0	9.0	9	27	9.0

Item		25		26		27		28		29		30	
Option		T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
T1 T2		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
A	1	54.1	0	32.4	55	13	6.3	7.2	28.8	14.3	9.9	18.9*	15.3
B	2	25.2*	0	43.2	6.3	36	60.4*	69.4*	46.8	78.3*	5.4	68.5	25.2
C	3	6.31	79.3*	19.8*	33.3*	36	0	14.4	24.3*	5.4	80*	7.2	54.1*
D	4	6	6.3	4.5	5.4	15*	33.3	9.0	0	0	4.5	5.4	5.4

*The correct answer and reason

Result and Analysis of Item Validity

According to Kimberlin & Winsterstein (2008), validity refers to whether the information obtained from a test represents the actual understanding of the examinees. The item validity is shown by the value of the Pearson correlation index (r count). To determine whether an item or question is categorized as valid or invalid, the value of the r count calculation of each item is compared with the value of the r table. The higher the r count, indicating the greater the validity, which means that students' answers to the question represent their actual understanding. Data on Table 8 showed that the items in the three tiers contained 88.9% valid items, and 11.1% were invalid. There are 66.7% of 30 items in the TT-DICE are consistently valid in their three tiers. Meanwhile, the other 33.3% items are still in any combinations invalid in one of their tiers. This overall result indicates if the items in the TT-DICE have considerably good construction validity and have precisely measured the expected aspects of the content. Further, look at the two invalid items in the first-tier (Q12 and Q21) revealed if the two items also have a small DI (discrimination index), that is, 0.13 and 0.18 consecutively which might show if the two items should be reconsidered when was about to distinguish between the high and lower achievers. Item validity is closely related to DI because if the item can distinguish between high and low achievement students, it means that the item has been trusted to measure students' conceptions.

Overall, the second-tier has the highest percentage of invalid items. This might be related to the level of difficulties and the abilities the students should have in performing the tasks, i.e., the ability to explain the answer they have chosen. Such a phenomenon is strengthened by the value of the DI of the invalid items, which are in the sufficient category (0.1- 0.3). Hence, the probability of the guessing answers in this tier is likely higher (Gurel *et al.*, 2015).

Table 8. The Validity of Each Item at 95% Significance Level

Items		1	2	3	4	5	6	7	8	9	10
First-tier	r_{count}	0.38	0.53	0.38	0.45	0.58	0.31	0.22	0.23	0.36	0.53
	category	valid	valid	valid	valid	valid	valid	valid	valid	valid	valid
Second-tier	r_{count}	0.55	0.63	0.32	0.18	0.68	0.11	0.23	0.51	0.10	0.68
	category	valid	valid	valid	invalid	valid	invalid	Valid	valid	invalid	valid
Third-tier	r_{count}	0.45	0.45	0.44	0.47	0.14	0.45	0.38	0.47	0.61	0.44
	category	valid	valid	valid	valid	invalid	valid	valid	valid	valid	valid

Items		11	12	13	14	15	16	17	18	19	20
First- tier	r_{count}	0.30	0.13	0.29	0.24	0.66	0.34	0.28	0.52	0.34	0.54
	category	valid	invalid	valid	valid	valid	valid	valid	valid	valid	valid
Second-tier	r_{count}	0.23	0.37	0.15	0.24	0.75	0.75	0.18	0.55	0.36	0.80
	category	Valid	valid	valid	valid	valid	valid	Invalid	valid	Valid	valid
Third-tier	r_{count}	0.54	0.73	0.69	0.10	0.51	0.72	0.60	0.89	0.70	0.25
	category	Valid	valid	valid	valid	valid	valid	valid	valid	Valid	valid

Items		21	22	23	24	25	26	27	28	29	30
First-tier	r_{count}	0.18	0.43	0.47	0.77	0.58	0.49	0.69	0.61	0.46	0.33
	category	Invalid	valid	valid	valid	valid	valid	valid	valid	Valid	valid
Second-tier	r_{count}	0.15	0.11	0.44	0.61	0.35	0.50	0.38	0.49	0.11	0.10
	category	Valid	invalid	valid	valid	valid	valid	valid	valid	Invalid	Invalid
Third-tier	r_{count}	0.68	0.58	0.71	0.29	0.64	0.50	0.73	0.48	0.57	0.87
	category	Valid	valid	valid	valid	valid	valid	valid	valid	Valid	valid

The item validity for third-tier shows that 29 items (96.7%) are classified as valid, and only 3.3% is invalid. This shows that three options in LC are appropriate for measuring students' confidence in each item that measures their understanding of the CE concept. Invalid items at level three are item five with validity values of 0.14 (sufficient). Although invalid, the item is still positive so that it can still be used by revising the language of the questions at the first level and the reasons at the second level, because it might be caused by a malfunction distractor on the two-level. Based on the distractor analysis on item "5" (see Table 7), it appears that the option "D" in tier one and the choice of reasons "4" no testee chooses ("0%"). This item is classified as having a bad distractor, and affect the value of item validity (DiBattista & Kurzawa, 2011).

Another factor that might affect item validity is the number of testee that guess. This fact, based on the analysis results for item "five," found that students found classified as understanding concepts were 62.2%, misconceptions 16.2%, and less understanding concepts by 21.6%. The results of the students' level of confidence score for item "5" found that there were 65.8% of students chose confidence (score 2). The misconception category is that students are wrong and confident. That means as many as 16.2% of students with misconceptions must get a score of 2. Thus, students who understand the concept are only 49.6% who choose a score of 2, meaning that there are 12.6% of students whose choices are correct at the first and second tier, but have a guess or not sure at the third-level. This indicates that the student understands the concept but lacks confidence (Arslan *et al.*, 2012).

Table 9 Result of Reliability Test

TT-DICE Aspect	Alpha Cronbach Coefficient	Category
Answer	0.83	Very good
Reason	0.81	Very good
Level of confidence	0.91	Excellent

Result and Analysis of Reliability of TT-DICE

The method most used to estimate the internal consistency or reliability of the instrument is Cronbach's alpha. Cronbach's alpha is a function intercorrelation of an item and the number of questions in the scale (Kimberlin & Winsterstein, 2008). The internal consistency coefficient gives an estimation of measurement reliability. Based on that assumption, the item that measures the same construct must be correlated. The statistical reliability test was performed using the Cronbach Alpha coefficient in the three tiers for the TT-DICE and showed excellent results (see Table 9). These results indicate that the test instruments have internal consistency or high regularity (Creswell, 2012: 162). Tuckman & Harper (2012) suggested the reliability score be equal to or greater than 0.75 to satisfy the internal consistency standard in terms of measuring learning achievements and conceptual understanding.

The reliability of TT-DICE is related to the value of the discriminatory index (DI) (Ebel, 1967). This fact is supported by the reliability coefficient for answers, reasons, and the LC in accordance with the average value of DI, respectively 0.42, 0.39, and 0.54. DI, with a value of ≤ 0.2 as a bad item based on criteria by Tuckman & Harper (2012), namely, the first level consists of two items, the second level of nine items, and the third level of two items. Another contribution that makes the reliability coefficient of the three levels of TT-DICE high value is the absence of things that have a DI value is negative. Validity and reliability are the most critical parameters in determining instrument quality (Kimberlin & Winterstein, 2008). These two parameters are related to the DI and distractor parameters. Based on the results and analysis was done, several items revised, and the TTDICE final product is declared valid, reliable, and useful for future purposes.

Conclusions

The three-tier test is strongly recommended in diagnosing misconceptions among students. This type of test facilitates teachers distinguish between students suffering from misunderstanding and those who are simply lack of knowledge. For this occasion, the third tier measuring the students' confidence level or CRI takes place. Here, the misconceptions were indicated as having either incorrect answers or inappropriate reasonings, but problem solver (student) was sure (confident) about their response. The TT-DICE development has been declared valid, reliable, and applicable in investigating students' misconceptions in the concepts of Chemical Equilibrium. The misunderstanding identified in this trial includes misconceptions of equilibrium conditions, dynamic equilibrium, heterogeneous equilibrium, effects of changes of temperature, pressure, and concentration, the addition of inert gas, and catalyst for balance shifting systems. Those findings mostly confirmed the misconception propositions reported in our referred studies, with two additional items. The first, under the dynamic equilibrium, the concentration of substances changes, and the rate of the forward and backward reaction also changes. The second, adding inert gas does not affect a shift in the chemical equilibrium system. Also, the use of TT-DICE would highly be expected, especially in high schools, to develop the students' complete understanding of concepts. Continual diagnosis would be beneficial not only for students but also for teachers to help them reflect their overall lessons and to be more aware of their students' incomplete understanding issues. Further studies on the usefulness of the TT-DICE in measuring students' learning outcomes and scientific thinking skills would be suggested to complete the findings of this study.

References

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.

Alghazo, Y. M. & Alghazo, R. (2017). Exploring common misconceptions and errors about fractions among college students in Saudi Arabia. *International Education Studies,* 10(4), 133–140.

Ambrose, S. A., Bridges, M. A., Dipietro, M., Lovet, M.C., & Norman, M.K. (2010). *How does learning works*? USA: Jossey Bass Willey Imprint.

Arikunto, S. (1998). *Research procedure: A practical approach*. Jakarta: Rineka Cipta.

Arslan, H.O., Cygdemoglu, 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.

Banerjee, A. C. (1991). Misconceptions of students and teachers in Chemical Equilibrium. *International Journal of Science Education*, 13(4): 487-494

Barke, H. D., Hasari, A., & Yitbarek, D. (2009). *Misconceptions in Chemistry.* Book Chapter (145-170). Verlag Berlin Heidelberg: Springer

Beichner, R. J. (1994). Testing student interpretation of kinematics graphs. *American Journal of Physics*, 62(8), 750-762

Bilgin, I & Uzuntiryaki, E (2003). Students' misconception on the concept of Chemical Equilibrium. *Journal of Education and Science*, 28(127), 10-17.

Borich, G. D. (1994). *Observation Skills for Effective Teaching.* 2th Edition. New York: Macmillan Publishing Company.

Carey, S. (2000). Science education as conceptual change. *Journal of Applied Development Psychology,* 21(1), 13-19.

Caleon, I., and Subramaniam, R. (2010). 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.

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 Educational Research Practice*, 8(3), 293–307

Cheung, D., Ma, Hong-Jia, Yang, J. (2009). Teachers' misconceptions about the effects of the addition of more reactants or products on Chemical Equilibrium. *International Journal of Science and Mathematics Education*, 7: 1111-1133

Cook, M, Wiebe, E. N & Carter, G. (2008). The influence of prior knowledge on viewing and interpreting graphics with macroscopic molecular representations. *Science Education,* 92, 848-867

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Boston, Pearson, p.347.

Devetak, I., Vogrinc, J., & Glazar, S.A. (2010). States of matter explanations in Slovenian textbooks for students Aged 6 to 14. *International Journal of Environment & Science Education*, 5(2): 217-235.

Dhindsa, H. S.,& Treagust, D. F. (2014). Prospective pedagogy for teaching chemical bonding for smart and sustainable learning. *Chemistry Education Research and Practice,* 15, 435–446.

DiBattista, D., and Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests, CJSoTL, 2 (2), 4. http://dx.doi.org /10.5206/cjsotl

Dindar, A. C. & Geban, O. (2011). Development of a three-tier test to assess high school students' understanding of acids and bases. *Procedia Social and Behavioral Science,* 15, 600–604.

Durmaz, M. (2018). Determination of prospective chemistry teachers' cognitive structures and misconceptions about stereochemistry. *Journal of Education and Training Studies*, 6(9), 13. https://doi.org/10.11114/jets.v6i9.3353

Ealy, J. (2018). Analysis of students' missed organic chemistry quiz questions that stress the importance of prior general chemistry knowledge. *Science Education,* 8(42), 1-13.

Effendy. (2002). Upaya untuk mengatasi kesalahan konsep dalam pengajaran kimia dengan menggunakan strategi konflik kognitif. [Efforts to overcome misconceptions in chemistry learning using cognitive cognitive strategy]. *Media Komunikasi Kimia,* 6(2), 1-22.

Effendy. (2007). *A-Level Chemistry for Senior High School Students 2 A*. Malang: Bayumedia Publishing.

Erman, E. (2017). Factors contributing to student's misconceptions in learning covalent bonds. *Journal of Research in Science Teaching*, 54(4), 520-537

Garnett, P. J., Garnett, P. J., & Hackling, M. W. (1995). Student alternative conception in chemistry: A review of research and implications for teaching and learning. *Studies Science Education.,* 25, 69-95.

Gilbert, J. K., & Treagust, D. (2009). Introduction: Macro, sub-micro, and symbolic representations and the relationship between them: Key models in chemical education. In J. K. Gilbert & D. Treagust (Eds.), Multiple representations in chemical education (pp. 1–8). the Netherlands: Springer.

Gorodetsky, M. & Gussarsky, E (1986). Misconceptualization of the chemical equilibrium concept as revealed by different evaluation methods. *European Journal of Science Education,* 8(4): 427-441.

Gurel, D. K, Erylmaz, A. & Mcdermott, L. C. (2015). A review and comparison of a diagnostic instrument to identify student misconceptions in science. *Eurasia Journal Mathematic. Science, and Teaching Education*, 11(5), 1-20.

Hasan, S., Bagayoko, D., & Killey, E. L. (1999). Misconception and the certainty of the response index. *Journal Physics Education*, 34(5): 294-299.

Hackling, M. W., & Garnett, P. J. (1985). Misconceptions of chemical equilibrium. *European Journal of Science Education*, 7(2), 205-214.

Hay, D. B & Kinchin, I. M. (2006). Using concept maps to reveal conceptual typologies. *Emerald Insight*, 48 (2/3): 127-142.

Horton, C. (2007). Students' alternative conception in chemistry. *California Journal of Science Education*, 7(2): 18-38

Johnstone, A. H. (2000). The teaching of chemistry logical or psychological. *Chemistry Education Research and Practice*, 7(2), 49-63.

Karpudewan, M., Treagust, D.F., Mocerino, M., Won, M. & Chandrasegaran, A.L. (2015). Investigating high school students' understanding of chemical equilibrium concepts. *International Journal of Environmental & Science Education*, 10(6), 845-863.

Kinchin, I. M. (2000). Using concept maps to reveal understanding: A two-tier analysis. *Sch. Sci. Rev.,* 81(296), 41–46.

Kimberlin, C. L., and Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research *Am. J. Health Syst. Pharm*., 65 (23), 2276–2284.

Nakhleh, M. B. (1992). Why don't some students learn chemistry (chemical misconception)? *Journal of Chemical Education*, 59(3), 191-194.

Niaz, M. (1998a). A Lakatosian conceptual change teaching strategy based on student ability to built models with varying degrees of conceptual understanding of Chemical Equilibrium. *Journal of Science Education, 7,* 107-127.

Novak, J. D. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science and Technology, 27* (10), 937–949

O'Connor, C. (2015). A practice-led approach to aligning learning theories with learning and teaching strategies in third level chemistry education. *Irish Journal of Academic Practice, 4*(1), 7. https://doi.org/10.21427/D7B42F

Osborne, R. J., & Wittrock, M. C. (1983). Learning science: A generative process. *Science Education, 67*(4), 489–508.

Özmen, H. (2008). Determination of students' alternative conceptions about Chemical Equilibrium: A review of research and the case of Turkey. *Chemistry Education Research and Practice, 9*(3), 225–233. https://doi.org/10.1039/B812411F

Peşman, H. & Eryilmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *Journal of Educational Research, 103*(3), 208–222. https://doi.org/10.1080/00220670903383002

Peterson, R.F. & Treagust, D.F. (1989). Development and application if diagnostic instrument to evaluate grade-11 and 12 students' concept of covalent bonding and structure following a course of instruction. Journal of Research in Science Education, 26, 301–314.

Piquette, J. S. & Heikkinen, H. W. (2005). Strategies reported used by instructors to address alternate student conceptions in Chemical Equilibrium. *Journal of Research in Science Teaching, 42* (10), 1112-1134

Santos, V. C & Arroio, A. (2016). The representational levels: Influences and contributions to research in chemical education, *Journal of Turkish Science Education,* 13(1): 3-18.

Seery, M. K. (2009). The role of prior knowledge and student aptitude in undergraduate performance in chemistry: A correlation-prediction study, *Chemistry Education Research, and Practice.,* 10 (3), 227–232.

Sendur, G; Toprak, M; Pekmez, E.S. (2011). How can secondary school perceive Chemical Equilibrium? *New World Sciences Academy, 6*(2), 1512-1531.

Shing, Y. L. & Brod, G. (2016). Effects of prior knowledge on memory: Implications for education. *International Mind, Brain, and Educational Society,* 1-9. *https://doi.org/10.1111/mbe.12110.*

Sozbilir, M., Pinarbasi, T., Canpolat. (2010). Prospective chemistry teachers ' conceptions of chemical, *Eurasia Journal of Mathematics, Science, and Technology Education*, 6(2): 111-120.

Stojanovska, M., M. Petruševski, V., & Šoptrajanov, B. (2017). Study of the use of the three levels of thinking and representation contributions. *Section of Natural, Mathematical and Biotechnical Sciences*, *35*(1), 37–46.

Suruchi, S., and Rana, S.S. (2014). Test item analysis and the relationship between difficulty level and discrimination index of test items in an achievement test in biology, *Paripex India Journal of Research*, 3 (6), 56–58.

Taber, K. S. (2009). Challenging misconceptions in the chemistry classroom: Resources to support teachers. *Educacion Química*, 4, 13–20. DOI: 102436/20.2003.02.27

Taber, K. S. (2011). Models, molecules, and misconceptions: A commentary on "Secondary School students' misconceptions of covalent bonding." *Journal of Turkish Science Education, 8*, 3–18.

Taber, K. S. (2015). Prior Knowledge. Book Chapter *Encyclopedia of Science.* 785-786. DOI: 10.1007/978-94-007-2150-0_483.

Talanquar, V. (2010). Exploring dominant types of explanations built by general chemistry students. *International Journal of Science Education,* 32(18), 2393-2412.

Talanquer, V. (2011). Macro, submicron, and symbolic: The many faces of the chemistry "triplet." *International Journal of Science Education, 33*(2), 179-195.

Treagust, D.F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159–169.

Tuckman, B.W. & Harper, B.E. (2012). *Conducting educational research* (6th ed) New York: Rowman & Littlefield Publishers, INC.

Tyson, L., Treagust, D.F., & Bucat, R.B. (1999). The complexity and teaching and learning Chemical Equilibrium. *Journal of Chemical Education, 79*(4), 554-558.

Voska, K. W., & Heikkinen, 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. https://doi.org/10.1002/(SICI)1098-2736(200002)37

Yakmaci-Guzel, B. (2013). Preservice chemistry teachers in action: An evaluation of attempts for changing high school students' chemistry misconceptions into more scientific conceptions. *Chemistry Education Research and Practice, 14*(1), 95–104. https://doi.org/10.1039/c2rp20109g

Yang, D.C. & Sianturi, I.A.J. (2019). Assessing students' conceptual understanding using an online three-tier diagnostic test. *Journal of Computer Assisted Learning.*1-12. DOI: 10.1111/jcal.12368