International Education Studies; Vol. 13, No. 9; 2020
ISSN 1913-9020 E-ISSN 1913-9039
Published by Canadian Center of Science and Education

Investigation of Three-Tier Diagnostic and Multiple Choice Tests on Chemistry Concepts with Response Change Behaviour

Suat Türkoguz[1]

[1] Buca Faculty of Education, Dokuz Eylül University, İzmir, Turkey

Correspondence: Suat Türkoguz, Buca Faculty of Education, Dokuz Eylül University, İzmir, Turkey.

Received: January 24, 2020 Accepted: June 9, 2020 Online Published: August 24, 2020

doi:10.5539/ies.v13n9p10 URL: https://doi.org/10.5539/ies.v13n9p10

Abstract

This study aims to investigate the test scores of the three-tier diagnostic chemistry test (TD̈C̈T) and multiple choice chemistry test (MCCT) by response change behaviour (RCB). The study is a descriptive research study aiming to investigate the item response efforts of TD̈C̈T and MCCT in a computerized testing environment (Quizzer test program, QTP). In both TD̈C̈T and MCCT, QTP maintains a continuous record for each tier of the test. Participants in the study are students in the Science Education Department at the state university in the Aegean region of Turkey (n=115). The study was conducted in two groups: there were 58 students in Group 1 and 57 students in Group 2. In Group 1, a TD̈C̈T was used; in Group 2, an MCCT test was applied. Tests were distributed by random sampling between Group 1 and Group 2. The data were collected by adding a confirmation tier to the TD̈C̈T involving 44 items. The TD̈C̈T was applied to 115 pre-service teachers; the reliability coefficient of the test was found to be 0.72. SPSS and MS Excel programs were used to analyse the data. Data were analysed using descriptive statistical methods. Considering the results obtained from the study, the rate of completing the test with RCB of test items for both tests is approximately 7–12 per cent. Another important consequence is that RCB does not provide an advantage or disadvantage in terms of scoring.

Keywords: misconception, tier diagnostic tests, response change behaviour

1. Introduction

The first studies on misconceptions in scientific concepts were in the early 1970s, and by the middle of the 1980s, they were the focus of several researchers (Driver, 1981; Linke & Venz, 1979; Osborne & Cosgrove, 1983; Tamir, 1971). Up to 2015, there were many studies on misconceptions in science education. Open-ended questions, multiple-choice tests (MCTs) and interviews were used in 91 per cent of these studies. A very small proportion (9%) benefited from tier diagnostic tests ($\widehat{\text{TDT}}$ s) (Gürel, Eryılmaz, & McDermott, 2015). Nowadays, misconceptions are determined by $\widehat{\text{TDT}}$s, and these tests are supported by interviews. Two-tier diagnostic tests (TD̈T), three-tier diagnostic tests (TD̈T) and four-tier diagnostic tests (TD̈T) are being developed (Odom & Barrow, 1995; Peşman & Eryılmaz, 2010).

1.1 What Are the Features of the Tests Used to Determine Misconceptions?

When the measurement tools used to determine misconceptions are examined in terms of validity and reliability, it is seen that some have advantages and disadvantages.

MCTs are often preferred because they have a large number of test items, can be applied to large samples, and are easily prepared and evaluated. MCTs can measure students' behaviour at the level of knowledge and conceptual learning, but they cannot measure students' inquiry and reasoning ability. These features of MCTs make it very difficult to distinguish between students with conceptual knowledge and students with misconceptions. Therefore, an MCT is insufficient to detect misconceptions in students (Gönen, S. Kocakaya, & F. Kocakaya, 2011). Therefore, considering the disadvantages of MCT, the importance of $\widehat{\text{TDT}}$s has increased.

Tests consisting of open-ended questions often contain two to four test items and therefore have less validity. Since tests consisting of open-ended questions are evaluated qualitatively by content analysis, they have reliability problems. The analysis process takes a long time and fails to give valid results for students who have low ability to express themselves in writing (Abraham, Grzybowski, Renner, & Marek, 1992; Özdemir & Kocakülah, 2016). Although these measurement tools do not conform to the basic approaches to classical test theory and item response theory, they create mathematical modelling and analysis problems.

TDT or TDT may be more standardized than other tests. In addition, since the knowledge of students is questioned in the second tier (II) of TDTs, the chance factor can be reduced. In the context of measurement theories, the error rate of measurement tools reducing the chance factor decreases, and thus reliability increases (Çakır & Aldemir, 2011; Özbayrak & Kartal, 2012). However, there are some problems with the writing of distractors in $\widehat{TDT}s_{(II)}$. In addition, the use of a small number of test items and the scoring of the stages with different models in TDTs raise validity and reliability problems (Xiao, Han, Koenig, Xiong, & Bao, 2018). $\widehat{TDT}s_{(II)}$ can be designed as MCTs or open-ended tests. It may be a disadvantage that $\widehat{TDT}s_{(II)}$ have open-ended items. For example, in a diagnostic test developed based on the PISA (Programme for International Student Assessment) exam, it was stated that students could easily do the first tier with its multiple choice items, but had difficulty writing reasons for their responses in the second tier consisting of open-ended question items (Sadıç & Çam, 2015). Therefore, it may be a disadvantage if TDTs are prepared using open-ended test items. The validity and reliability of the tests can be solved if TDTs are prepared with item choices using standard misconceptions.

As the second tier is integrated into the first tier of TDTs, the response time and performance of the test are affected. Therefore, these tests can be applied with a limited number of items. Therefore, $\widehat{TDT}$s can be criticized in terms of content validity. Research has found that while students were successful in numerical problem content tests, they failed in $\widehat{TDT}$s where relationships between concepts were explored, and in $\widehat{TDT}$s, items are often used to address the relationship between concepts (Bernhard, 2000; Crouch & Mazur, 2001). There are validity and reliability problems due to scoring problems of open-ended questions, drawing-based qualitative measurement tools and MCTs. Therefore, these measurement tools are inadequate to measure students' abilities, skills and knowledge. For these reasons, the need for studies on adaptive $\widehat{TDT}$s is increasing day by day.

1.2 How are $\widehat{TDT}$s Prepared and Scored?

Although $\widehat{TDT}$s are similar to MCTs in terms of their practice, there are differences in the development process of the tests. Firstly, in the preparation of TDT or TDT, concept map drawing comes to the forefront. It is important to show the relationships between concepts correctly while preparing a $\widehat{TDT}$, because $\widehat{TDT}$s pay attention to the relationships between these concepts. While preparing a $\widehat{TDT}$, learning objectives in the curriculum are determined on the concept map first; then misconceptions from literature in pursuant of the determined learning objectives are listed; finally, the test items are created (Treagust, 1988). The first tier (I) of $\widehat{TDT}$ is similar to an MCT item that measures concepts about the subject; $\widehat{TDT}_{(II)}$ is composed of choices that are thought to be related to $\widehat{TDT}_{(I)}$. A $TDT_{(II)}$ prepared in this way may consist of items which express the reasoning for the response in connection with $TDT_{(I)}$ (Mutlu & Şeşen, 2015; Taber, 1999; Uyulgan, Akkuzu, & Alpat, 2014). In some studies, when developing a $\widehat{TDT}$, different methods may be followed in the process of test development than that proposed by Treagust (1988). Some researchers prefer to quote the validated and reliable test items used in theses, scientific articles, and national and international exams in the first tier of the tests when developing $\widehat{TDT}$s (Sadıç & Çam, 2015). After $TDT_{(II)}$, an 'Are you confident of your response?' is added, and TDT is converted into TDT. The students are asked to confirm this question as 'Yes/No' and their consistent attitude towards the concept is determined: in fact, it is the student's consistent attitude that proves the existence of misconceptions in the test item. One of the choices in TDTs should be correct in terms of scientific proposition, and the propositions in the other choices should contain misconceptions. Similar to TDTs and TDTs, a TDT can be prepared. In TDTs, after $TDTs_{(I)}$ and $TDTs_{(III)}$, the question proposition confirming the students' confidence in their answers is added an extra two times. In summary, $TDTs_{(II)}$ and $TDTs_{(IV)}$ are the confirmation stage.

It is important to identify misconceptions as well as to measure current knowledge and the curriculum objectives achieved at the beginning or end of the learning process (Bektaş & Kudubeş, 2014). The scoring of $\widehat{TDT}$s is done differently from other measurement and evaluation tools. In the scoring of a $\widehat{TDT}$ in the light of classical test theory, when the participant correctly answers both tiers of the $\widehat{TDT}$, the score of the test item becomes 1; in other cases, the score of the test item is 0. A binary rating such 0/1 is considered to be more reliable computation. However, there is still controversy about the reliability and scoring of $\widehat{TDT}$s and MCTs (Bademci, 2006; Taber, 2017). Due to reliability and scoring difficulties, each tier of a $\widehat{TDT}$ can be evaluated under separate parameters with logistic or Rasch models (Xiao, Han, Koenig, Xiong, & Bao, 2018). In $\widehat{TDT}$s, it is important to find the reason for the students' answers alongside the score and to determine any misconceptions about the subject. Scoring the tiers in a $\widehat{TDT}$ is advantageous both in determining the lack of knowledge and in arriving at test scores.

Hestenes and Halloun (1995) proposed the definition of the false positive (FP) and the false negative (FN) as evidence of external validity in TDTs. By Hestenes and Halloun (1995), FP is defined as the correct response to the test item with a confident attitude based on an incorrect reason, while FN is defined as an incorrect response to the test item with a confident attitude based on the correct reason. The researchers also noted that minimizing the probability of FPs and FNs could provide higher validity in $\widehat{TDT}$s. FN for external validity in $\widehat{TDT}$s should be less

than 10 per cent (Gürçay & Gülbaş, 2015; Şen & Yılmaz, 2017). However, it is very difficult to reduce FP in $\widehat{TDT}$s. Due to the nature of the test, students can choose the right alternative in the content layer even if they have misconceptions (Peşman & Eryılmaz 2010).

There are some problems with the scoring of $T\ddot{D}Ts_{(I)}$ and $T\ddot{D}Ts_{(II)}$ separately or jointly. These are the chance factor, preference interactions between tiers, and uncertainties in calculating the reliability coefficient. In the model proposed by Hestenes and Halloun (1995), if the students know that the scoring of the test will compute $\widehat{TDT}s_{(I)}$ and $\widehat{TDT}s_{(II)}$ together, their choices can contribution positively to FP (the first tier is the correct, the second tier is incorrect). However, more in-depth scientific understanding and reasoning processes may not be determined by their scoring model. In this case, students can avoid guessing to the test items. If students know that $\widehat{TDT}s_{(I)}$ and $\widehat{TDT}s_{(II)}$ will be scored separately in $\widehat{TDT}$s, their choices may have a negative effect on FP. Students can make separate estimates for both tiers and increase the chance factor in the $\widehat{TDT}$ tiers. In this type of scoring, students can establish a systematic relationship between the stages and thus predict (Xiao, Han, Koenig, Xiong, & Bao, 2018). In compared with their model, this study may be important in terms of seeing the positives and negatives of the three-tier diagnostic chemistry test ($TD\ddot{C}T$).

1.3 Response Change Behaviour (RCB) in Tests

In the tests, the effects of students' RCB on the test scores can be examined. There is a common belief that the first response to the test items is correct in the test and the second choice is incorrect if the response is changed later. Although test participants have worries about changing responses, they persistently change their responses (Cox-Davenport, Haynes, & Lawson, 2014). Therefore, the effect of RCB on test scores has always been a matter of interest. For this purpose, permanent markings and deleted markings were initially examined in paper and pencil tests. Nowadays, RCBs and response time can be followed by the software system in computerized tests. This recording feature of computers may be an important source of data for researchers, and maybe evaluated as a parameter in ability estimation.

When the deleted markings or choices in the open-ended, true-false and MCTs made with paper and pencil items were examined, it was found that there was a general increase in the test scores of test participants (Al-Hamly & Coombe, 2003; Baştürk, 2011; Beck, 1978; Cox-Davenport, Haynes, & Lawson, 2014; Kim, 2019; Lynch & Smith, 1972). Only Noorbala and Mohammadi (2011) explained that RCB had a negative effect on test scores in a study conducted with medical students. It was found that test participants with a higher test score or more talent showed less frequent RCB than other weaker candidates (Beck, 1978; McMorris, Schwarz, Richichi, Fisher, Buczek, Chevalier, Meland, 1991). It was found that repeating RCB of a test item does not contribute to the test score (Lynch & Smith, 1972). It was observed that the test type had no effect on RCB (McMorris, Schwarz, Richichi, Fisher, Buczek, Chevalier, Meland, 1991). There were no significant differences between the sexes in studies of RCB (Baştürk, 2011). In the comparison of RCB with variables such as test item difficulty, the frequency of RCB was parallel to item difficulty. It was seen that students showed more RCB with difficult items (Baştürk, 2011; Beck, 1978; Lynch & Smith, 1972). Some studies have emphasized that students should be encouraged to change their response behaviour (Al-Hamly & Coombe, 2003; Casteel, 1991; McMorris et al., 1991). In addition, participants who change their response behaviour during the test spend more time and exhaust their minds. Therefore, the response performance for test items may be affected. Therefore, RCB can be used in two-, three- and four-parameter logistic modelling for ability estimation in tests (Kim, 2019; Yen, Ho, Liao, & Chen, 2012). In addition, when the probability value for RCB is used in three-parameter logistic models, students engaging in cheating can be identified. If the RCB is examined in computerized test environments, it can be determined how many times the RCB is repeated and how long it takes to decide. Therefore, RCB should be considered in computerized tests (Van Der Linden & Jeon, 2012). RCB should be considered when developing test items (Lynch & Smith, 1972). RCB can be utilized in test development processes.

1.4 Applications of Computer-Based Tests from Past to Present

Computers have been included in administration offices of schools as an auxiliary tool outside the learning process since the 1980s, and in the classroom as a teaching tool since the 1990s. Since the 2000s, their use in measurement and evaluation has become widespread and is today very active (Linden & Glas, 2002). When the literature on the use of computers for measurement and evaluation is examined, it is seen that they are applied in different ways (Aybek, 2012). In computer-based tests, the projection of test items to the screen can take different forms, such as similar to paper and pencil tests, one by one sequential order, one by one blended order, or by students' preference. In addition, multimedia and visuals can be used for the presentation of test items on the screen. Data in computer-based tests can be collected online with external data loggers, a computer connected to the network centre, or online through an internet web server. The data obtained from computer-based tests can be processed on

the basis of classical test theory or on the paper and pencil test, and can be processed on the basis of item response theory by computer-adapted algorithmic methods. Computer-adapted algorithmic methods can be used for different variables such as test response time, personal attention and motivation data, in addition to test scores (Tabakçıoğlu, Çizmeci, & Ayberkin, 2016; Weiss & Kingsbury, 1984). Computers can be used in developing $\widetilde{\text{TDT}}s$ and determining misconceptions. Lin (2016) used computer-based TDTs to identify misconceptions about electrical circuits. Maier, Wolf and Randler (2016) showed that misconceptions can be better determined by the automatic feedback given to students during the application of the multi-tiered diagnostic test in the computer software environment. Yang, Hwang, Yang, and Hwang (2015) determined that students' skills were observed better with computer-based $\widetilde{\text{TDT}}s$ to measure their computer programming skills. In this study, it was aimed to determine the positive and negative aspects of computer-based $\widetilde{\text{TDT}}s$.

1.5 Quizzer Test Program (QTP)

Since there is no licensed computer-supported test program suitable for the purpose of this study, the Quizzer test program (QTP) has been developed by the expert software programmer in computer teaching technologies at the university where the research is conducted. During the academic semester, pilot trials of QTP were carried out and missing aspects were corrected. QTP aimed to perform the tests easily and effectively in the experimental and control groups. Examples of QTP are shown in Figure 1 and 2 in data collection section.

1.6 Problem Status, Importance and Purpose of the Study

In this study, it is understood from the literature review that there are very few studies on response behaviour in computer-supported exams. In addition, computer-supported exams were not encountered in relation to $\widetilde{\text{TDT}}$s, especially $\widetilde{\text{TDCT}}$s. Yang and Sianturi (2018) used a computerized online test in a three-tier mathematical diagnostic test, but this test was not related to chemistry. Chiang and Chiu (2015) performed a computer-supported chemistry test, but this computer-supported test was not a $\widetilde{\text{TDCT}}$. The purpose of their test was to reveal mental models in chemistry. There are not many studies investigating RCB with computer-supported $\widetilde{\text{TDCT}}$. An indicator of misconceptions is that students insist on their response. Students may exhibit RCB for some test items in $\widetilde{\text{TDT}}$s. Because RCB shows that the student is not completely confident about the concept of the test item, if the student responds to the test item with a single response behaviour, it can be understood that he/she knows the response and is confident of the response. It can be said that if the decision in the single response behaviour of the student is wrong, the student has misconceptions, because the misconception arises by insisting that the correct response is wrong. Therefore, if the majority of students insist on their decision in single response behaviour, this may indicate misconceptions. There are no data showing the positive or negative contribution of RCB to tiers in $\widetilde{\text{TDT}}$s. However, the advantages and disadvantages are not known in relation to conventional tests. In this context, the aim is to compare QTP-supported TDC̈T and QTP-supported MCCT considering students' response behaviours, but also to investigate RCB in both tests and to determine the effect of $\widetilde{\text{TDCT}}$ between tiers.

1.7 Research Questions

The problem explored in the study is: What are the differences between the three-tier chemistry diagnostic test (TDC̈T) and the multiple-choice chemistry test (MCCT) for RCB? In this context, the following sub-problems were utilized in the solution of the problem.

1) Do RCB percentages differ significantly between TDC̈T and MCCT?
2) Is there a significant difference between the correct response percentages of TDC̈T and MCCT according to single response behaviour?
3) What are rates of false positive (FP) and false negative (FN) for TDC̈T according to single response behaviour?
4) What are the trend of correct (TC) and the trend of incorrect (TIC) responses for TDC̈T and MCCT according to RCB?

2. Method

This study is a descriptive study aiming to examine the test scores of students who participated in computer-supported TDC̈T and computer-supported MCCT considering RCB. In this study, a comparative research design was used within the scope of a non-experimental research design. In comparative design, the difference between two or more events or cases is investigated (Fraenkel & Wallen, 2000; McMillan & Schumacher, 2010). Therefore, the experimental and control groups were formed to determine different features of TDC̈T and MCCT for this study, but no experimental treatment process affecting the groups was performed. The data were collected within the context of the Chemistry II course in the Science Teacher Training programme in a

state university in Turkey.

2.1 Participants in the Study

The participants were pre-service science teachers at a state university in Turkey (n=115) in the 2017–2018 academic year. Experimental and control groups were selected randomly from the classes in which the students were officially registered. For this reason, the study was conducted with two groups determined by random sampling. In Group 1 (n=57), TD̈C̈T was performed, while MCCT was applied in Group 2 (n=58). Participants comprised 89 female teachers and 26 male pre-service teachers, distributed as 46 female and 11 male pre-service teachers in the experimental group and 43 female and 15 male pre-service teachers in the control group.

2.2 Data Collection

The data were collected by adding a confirmation tier to TD̈T involving 44 items developed by Mutlu and Şeşen (2015). The tests consisted of chemistry concepts such as acids-base, electrochemistry, thermodynamics, chemical kinetics and equilibrium. Their test was developed with 151 pre-service teachers. The test reliability was found to be 0.84. In this study, a third tier was added to their test and TD̈C̈T was converted to TD̈C̈T. TD̈C̈T was then applied to 115 pre-service teachers. TD̈C̈T$_{(I)}$ and TD̈C̈T$_{(II)}$ were coded by the graded scoring of Milenković, Hrin, Segedinac, and Horvat (2016), and TD̈C̈T's reliability was calculated as 0.72. KR_{20} was found to be 0.51 when scored as 1 in the correct response in both tiers (first and second) of TD̈C̈T and 0 in other response situations. When the KR_{20} coefficients of TD̈C̈T$_{(I)}$ and TD̈C̈T$_{(II)}$ were calculated separately, the results were 0.18 for TD̈C̈T$_{(I)}$ and 0.51 for TD̈C̈T$_{(II)}$. During the test process, response performances in the items following the item marked by RCB may be affected by participants' RCB (Kim, 2019; Yen, Ho, Liao, & Chen, 2012). In other words, scores of test items may be affected by RCB during the test process. TD̈C̈T$_{(I)}$ included the question forms of items and their distractors; TD̈C̈T$_{(II)}$ included the misconception choices which made a causative inquiry process related to TD̈C̈T$_{(I)}$; and the final tier included the stage in which the responses were confirmed. MCCT involved the question forms of items, their distractors and the confirmatory response in last test item. The test structures for TD̈C̈T$_{(I)}$ and MCCT were the same. Therefore, there may be differences in reliability coefficients. These test items, taken from Mutlu and Şeşen (2015) and transferred to QTP, were applied as TD̈C̈T for the experimental group and MCCT for the control group. These tests were performed individually in the computer laboratory. Figure 1 shows the process of time recording in QTP for TD̈C̈T. Figure 2 shows sample screenshots from QTP.

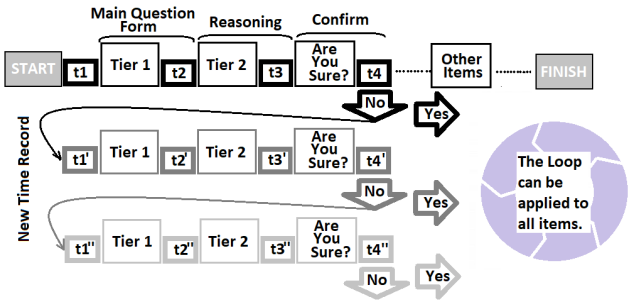

Figure 1. Application image cycle in QTP for TD̈C̈T (Group 1)

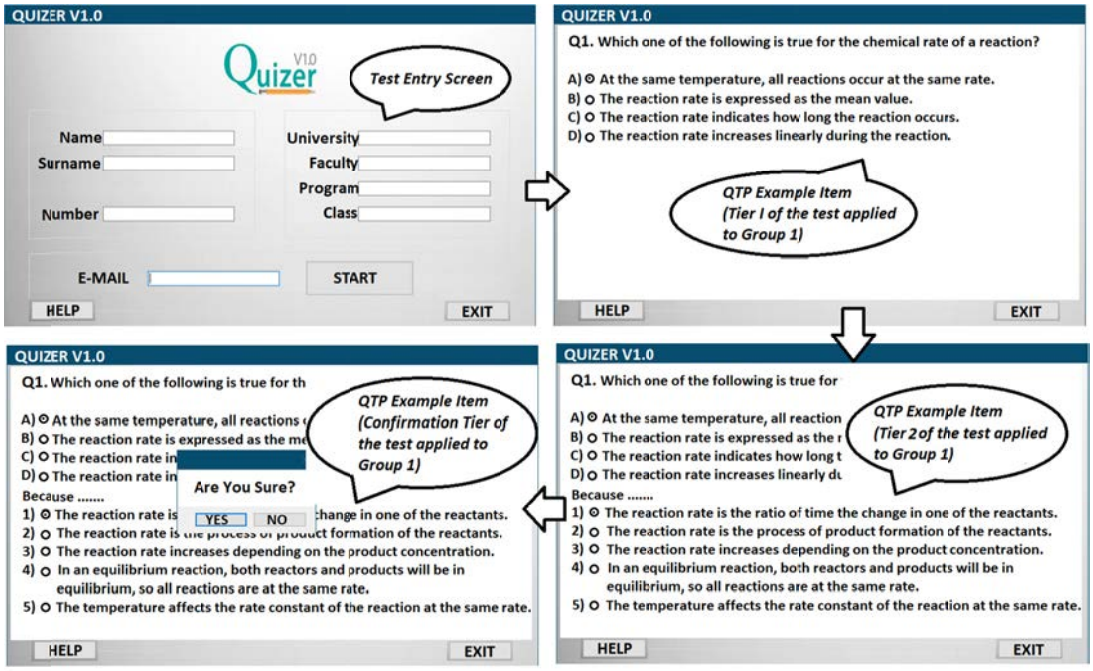

Figure 2. The process of time recording in QTP for TDÖCT (Group 1)

2.3 Data Analysis

2.3.1 Calculating Percentage of RCB

The percentage of RCB is calculated by the equation $RCB = \frac{\sum_{i=1}^{N} \frac{\sum_{j=1}^{K} RC_{ij}}{K}}{N} x100$. In the equation, RCB represents the percentage of items that show RCB from the test items (K) for the person (N). Response behaviour (RC_{ij}) represents the preference given by the person i to test item j (changing response). Person i can receive a maximum score of $RC_{ij} = 1$ from item j of the test. The formula was developed from Wise and Kong (2005).

2.3.2 Calculating Correct Response Percentage According to Single Response Behaviour

The correct response percentage according to single response behaviour (RCB_+) was calculated by $RCB_+ = \frac{\sum_{i=1}^{N} \frac{\sum_{j=1}^{K} RC_{ij}}{K}}{N} x100$. In the equation, RCB_+ represents the percentage of items that show correct responses in the single response behaviour from the test items (K) for the person (N). Response behaviour (RC_{ij}) represents the preference given by the person i to test item j (correct response according to single response behaviour). Person i can receive a maximum score of $RC_{ij} = 1$ from item j of the test. Here, RC_{ij} represents the correct response to the test item according to the single response behaviour. The formula was developed from Wise & Kong (2005).

After RCB and RCB_+ percentages were calculated, comparisons were made between $TDÖCT_{(I \text{ and } II)}$ and MCCT using t-test analysis for independent and dependent variables. In this study, SPSS and MS Excel programs were used for data analysis.

2.3.3 Calculating Rates of FP and FN for TDÖCT According to Single Response Behaviour

Only rates of FP and FN for TDÖCT can be calculated according to single response behavior. It is not possible to calculate the rates of FP and FN for MCCT. The response format of TDÖCT for single response behaviour is shown in Table 1, by Milenković, Hrin, Segedinac, and Horvat (2016).

Table 1. TDC̈T classification of response types (Milenković, Hrin, Segedinac, & Horvat, 2016)

$TDC̈T_{(I)}$	$TDC̈T_{(II)}$	TDC̈T(confirm)	Classification	Proportional Classification
Correct	Correct	Confident	Scientific knowledge	True positive (TP)
		Unconfident	Lucky prediction	
	Incorrect	Confident	False positive	False positive (FP)
		Unconfident	Lack of knowledge	
Incorrect	Correct	Confident	False negative	False negative (FN)
		Unconfident	Lack of knowledge	
	Incorrect	Confident	Misconception	True negative (TN)
		Unconfident	Lack of knowledge	

The FP and FN rates can be calculated for TD̈T using Table 1:

The FN rate is the power to distinguish FNs from TPs and is calculated by FN/(FN+TP).

The FP rate is the power to distinguish FPs from TNs and is calculated by FP/(FP+FN+TN).

2.3.4 Calculating TC and TIC Rates on for TDC̈T and MCCT According to RCB

Table 1 can be used in analysing response change behaviours for both TDC̈T and MCCT. In this case, when Table 1 for TDC̈T is adapted to RCB for both test types, 'first response in RCB' is used instead of $TDC̈T_{(I)}$, and 'second response in RCB' instead of $TDC̈T_{(II)}$. The confirmation tier is the last tier. In RCB, students used the right to reply a second time if they were not confident about their responses. In this case, it is accepted that the students are confident of their own answers since they have switched to the next question item with the 'I am confident' preference at the tier of confirming their responses in the second response process. Therefore, in Table 2, there is no 'unconfident' choice and the table is reduced. The final version of Table 2 is given below.

Table 2. Classification of TDC̈T and MCCT regarding RCB and response types

Responses in RCB		Confirm tier	Classification	Proportional Classification
First	Second			
Correct	Correct	Confident	Scientific knowledge	Stable correct (SC)
	Incorrect	Confident	Negative partial knowledge	Trend of correct (TC)
Incorrect	Correct	Confident	Positive partial knowledge	Trend of incorrect (TIC)
	Incorrect	Confident	Misconception	Stable incorrect (SI)

TC and TIC rates can be calculated for TDC̈T and MCCT using Table 2.

The TC rate is the power to distinguish TC responses from SC responses and is calculated by TC/(TC+SC).

The TIC rate is the power to distinguish TIC responses from SI responses and is calculated by TIC/ (TIC+SI).

3. Results

Findings are listed in order of the sub-problems: the percentage of RCB; the percentage of correct responses according to single response behaviour; the rates of FP and FN for TDC̈T according to single response behaviour; TC and TIC rates related to RCB.

3.1 Findings on Percentages of RCB

Table 3 shows the findings for the first sub-problem, explaining the percentage of RCB.

Table 3. Descriptive values of percentages of RCB

Group	n	Mean	SD	t	df	p
TDC̈T	57	12.73	7.83	3.184	113	.002
MCCT	58	8.09	7.80			

In the findings of Table 3, the percentage of RCB was 12.73 for TDC̈T and 8.09 for MCCT. The percentages of single response behaviour without RCB were 5.3 (3/57=0.053) for TDC̈T and 10.3 (6/58=0.103) for MCCT. Since the test used in the experimental group was TDC̈T, the connections between $TDC̈T_{(I)}$ and $TDC̈T_{(II)}$ led the students

to RCB. The majority of the students in both tests chose to respond to all of the items in the test without showing any RCB.

3.2 Findings for Correct Response According to Single Response Behaviour

Table 4 shows the findings of the second sub-problem, explaining correct response according to single response behavior.

Table 4. Descriptive values of RCB_{+} percentages for correct response according to single response behaviour

Group	n	Mean	SD	t	df	p
$TD\ddot{C}T_{(I)}$	57	34.61	7.44	-.545	113	.587
MCCT	58	35.31	6.23			
$TD\ddot{C}T_{(II)}$	57	30.74	8.01	-3.413	113	.001
MCCT	58	35.31	6.23			
$TD\ddot{C}T_{(I)}$	57	34.61	7.44	3.912	56	.000
$TD\ddot{C}T_{(II)}$	57	30.74	8.02			

In the descriptive values of Table 4, the average percentage of correct responses according to single response behaviour was 34.61 for $TD\ddot{C}T_{(I)}$, 30.74 for $TD\ddot{C}T_{(II)}$ and 35.31 for MCCT. There was no significant difference in descriptive values between $TD\ddot{C}T_{(I)}$ and MCCT. There is a significant difference between $TD\ddot{C}T_{(I)}$ and $TD\ddot{C}T_{(II)}$. Similarly, there is a significant difference between MCCT and $TD\ddot{C}T_{(II)}$. In $TD\ddot{C}T_{(II)}$, there was a slight decrease in scores. In the single response behaviour, the average percentage of correct responses from students ranged between 30 and 35 per cent. $TD\ddot{C}T_{(II)}$ shows an average correct response percentage of 30.74, which may indicate that students have misconceptions at this rate.

3.3 Findings of FP and FN for $TD\ddot{C}T$ According to Single Response Behaviour

Table 5 shows findings for the third sub-problem, explaining the rates of FP and FN according to single response behavior.

Table 5. Rates of FP and FN for all tiers of $TD\ddot{C}T$ according to single response behaviour

Rates	$TD\ddot{C}T$
FN	0.46
FP	0.15

In Table 5, the rates of FP and FN for $TD\ddot{C}T$ are 0.15 and 0.46 respectively. In this case, when the number of participants who make $TD\ddot{C}T_{(I)}$ incorrect and $TD\ddot{C}T_{(II)}$ correct is divided by number of participants who make both tiers of the test correct, the joint conditional rate is 0.46. In the same way, when the number of participants who make $TD\ddot{C}T_{(I)}$ correct and $TD\ddot{C}T_{(II)}$ incorrect is divided by number of participants who make both tiers of the test incorrect, the joint conditional rate is 0.15.

3.4 Findings of TC and TIC Rates for $TD\ddot{C}T$ and MCCT According to RCB

Table 6 shows the findings of the fourth sub-problem, explaining TC and TIC rates for $TD\ddot{C}T$ and MCCT according to RCB.

Table 6. TC and TIC rates of $TDCT_{(I \text{ and } II)}$ and MCCT according to RCB

Rates	$TD\ddot{C}T_{(I)}$	$TD\ddot{C}T_{(II)}$	MCCT
TC	0.86	0.38	0.98
TIC	0.37	0.11	0.40

Table 6 shows that the MCCT trend rates are highest in both TC and TIC considering trends in both tests. It is determined that the $TD\ddot{C}T_{(II)}$ trend rates are the lowest in both TC and the TIC considering trends in both tests. In both groups, it is seen that students change from an incorrect choice to a correct choice. This trend is greater in MCCT, while $TD\ddot{C}T_{(I)}$ is slightly lower than MCCT. This trend rate is very low in $TD\ddot{C}T_{(II)}$. The relationship between the tiers of $TD\ddot{C}T$ presents a problem for students to distinguish the correct from the incorrect choice.

Students' tendency to change from the correct choice to the incorrect choice is less than their tendency to change from the incorrect choice to correct choice. TD̈C̈T$_{(II)}$ has a low trend from correct to incorrect choice. MCCT directs students from the correct choice to the incorrect choice. All tiers of test items give clues to students.

4. Discussion

In the findings on the percentages of RCB in this study, it was seen that students participating in both the TD̈C̈T and the MCCT insisted on responding to some test items with single response behaviour, but not all of the items in the test. In other words, it is seen that they do not prefer RCB. Nevertheless, the majority of participants in both tests (TD̈C̈T/MCCT) felt the need to complete the test with RCB. This requirement appears to be higher in TD̈C̈T. It was seen that the students participating in TD̈C̈T had less insistence on continuing the test with single response behaviour than those doing the MCCT and relied more on RCB. It can be said that TD̈C̈T is more advantageous in terms of measurement and evaluation than MCCT. It is seen that the students participated in reasoning due to the tendency of their choices in these TD̈C̈T. However, this rate is not very high when comparing the values of TD̈C̈T and MCCT. The part of MCCT criticized by TD̈C̈T is that MCCT does not constitute a reasoning process. Another important finding is that students replied to most items in the test with single response behaviour, even though the majority of the students completed the test using RCB. This finding may be evidence that students do not want to use RCB. In this case, it is not advantageous to carry out tests with RCB.

In the single response behaviour of the study, the average percentage of TD̈C̈T$_{(I)}$ and the average percentage of MCCT were equivalent in terms of the percentage of correct responses. However, the average percentage in TD̈C̈T$_{(II)}$ was lower than the average percentage in both MCCT and TD̈C̈T$_{(I)}$. This finding differs from the result in Adodo's (2013) comparison of an MCT with a TD̈T. Adodo (2013) compared the adequacy of the multiple tier test with the TD̈T in a pre-test and post-test control group experiment. In his study, a slightly higher score was observed for TD̈T. This finding is similar to the studies by Li and Yang (2010), Yang, Li and Lin (2008), and Yang and Lin (2015). In these studies, it was found that the rate of correct responses was higher in TD̈C̈T$_{(I)}$ than in TD̈C̈T$_{(II)}$. In fact, in Yang and Lin's (2015) study, the percentage of correct responses in $\widehat{\text{TDT}}_{(I)}$ was around 50 per cent, while $\widehat{\text{TDT}}_{(II)}$ was around 25 per cent. There are studies showing that $\widehat{\text{TDT}}_{(I)}$ facilitates a score higher than $\widehat{\text{TDT}}_{(II)}$ (Arslan, Çiğdemoğlu, & Moseley, 2012; Peşman & Eryılmaz, 2010; Şen & Yılmaz, 2017). In this study, a significant difference was found between TD̈C̈T$_{(I)}$ and TD̈C̈T$_{(II)}$. However, the difference is not great compared to studies in the related literature. Yang and Lin (2015) considered that $\widehat{\text{TDT}}_{(I)}$ and $\widehat{\text{TDT}}_{(II)}$ were evaluated by the students as two separate problems. In this finding, the reason why TD̈C̈T$_{(II)}$ scores low can be explained in two ways: firstly, TD̈C̈T$_{(I)}$ may affect TD̈C̈T$_{(II)}$, or vice versa; secondly, it may be that TD̈C̈T$_{(II)}$ is somewhat more difficult. Because it is thought that more cognitive processes were performed in $\widehat{\text{TDT}}_{(II)}$ (Yang and Lin, 2015). In $\widehat{\text{TDT}}_{(II)}$, only one of the four choices included scientific knowledge and the other three contained misconceptions. When tests are examined in terms of the chance factor, TD̈C̈T$_{(II)}$ is disadvantageous in terms of scoring compared to both TD̈C̈T$_{(I)}$ and MCCT. This disadvantage should be considered when scoring TD̈C̈T (Xiao, Han, Koenig, Xiong, & Bao, 2018). Furthermore, the misconception rate of 69.26 per cent (100%–30.74%) according to the single response behaviour is a significant contribution to the research on TD̈C̈T$_{(II)}$.

In single response behaviour, the rates of FP and FN for TD̈C̈T are 15 per cent and 46 per cent respectively, these rates being above the critical point of 10 per cent (Hestenes & Halloun, 1995; Şen, Yılmaz, & Geban, 2018). It is difficult to reduce the rate of FP. Due to the nature of the tests, students may choose the right alternative according to the content of the test even if they have misconceptions (Peşman & Eryılmaz, 2010). This study showed that the explanation or causal reasoning in TD̈C̈T$_{(II)}$ was more difficult than TD̈C̈T$_{(I)}$. The TD̈C̈T FP rate in this study is similar to the results of the Rasch model in TD̈T conducted by Xiao, Han, Koenig, Xiong, and Bao (2018). In this study, it is considered normal for the rates of FP and FN to be higher than the critical value. The FP and FN for $\widehat{\text{TDT}}$ can be evaluated as parameters in logistic analysis. If the FP and FN rates are taken together according to RCB, more concordant results can be achieved.

In this study, the TD̈C̈T and MCCT findings regarding TC and TIC rates are compared according to RCB. The rate of TC was lowest in TD̈C̈T$_{(II)}$. It is seen that students change their preferences from an incorrect choice to a correct choice in RCB. This trend is highest in the MCCT and very low in TD̈C̈T$_{(II)}$. The relationship between TD̈C̈T$_{(I)}$ and TD̈C̈T$_{(II)}$ can distract students from an incorrect to a correct choice. It is seen that the TIC rate was highest in MCCT. Moreover, students' TIC rate was lower level than their TC rate for TDCT and MCCT. The TIC rate was lowest in TD̈C̈T$_{(II)}$. This trend's rate is higher in MCCT than in TD̈C̈T$_{(II)}$. TD̈C̈T$_{(I)}$ and TD̈C̈T$_{(II)}$ provide students with clues.

5. Conclusion

In general, the following results were obtained from the research. Firstly, although the majority of the students

completed the test with RCB, it was seen that they preferred to answer the majority of the items in the test with single response behaviour. In $\widehat{TDT}$s, this response tendency differs from other tests. Secondly, single response behaviour did not produce a significant difference between the $T\ddot{D}\ddot{C}T_{(I)}$ and MCCT scores. It was concluded that $T\ddot{D}\ddot{C}T_{(II)}$ is more difficult than $T\ddot{D}\ddot{C}T_{(I)}$. Thirdly, $T\ddot{D}\ddot{C}T_{(I)}$ and $T\ddot{D}\ddot{C}T_{(II)}$ were found to incorporate a guiding feature according to the single response behaviour, but the direction could not be determined. This could be because the test item difficulty of the two tiers of $\widehat{TDT}$ is different, and the item structures are different and are contrary to the theories. Furthermore, the majority of the choices in $\widehat{TDT}_{(II)}$ consist of possible misconceptions, and so have a significant advantage in terms of the chance factor. In this respect, $\widehat{TDT}$s can be criticized. Fourthly, it was found that RCB did not provide an advantage or disadvantage in terms of scoring. RCB percentage has showed that students have misconceptions definitely. Qualitative interviews with these students can lead to detailed results. Finally, although the response time for $T\ddot{D}\ddot{C}T$ was longer than MCCT due to $T\ddot{D}\ddot{C}T_{(II)}$, the scores of $T\ddot{D}\ddot{C}T_{(I)}$ were equivalent to the scores of MCCT, indicating that there were no negative aspects of $\widehat{TDT}$s in terms of time.

This study was limited to the data of $T\ddot{D}\ddot{C}T$ and MCCT. The misconception in $T\ddot{D}\ddot{C}T$ can be better confirmed by the questioning of confidence stage compared to $T\ddot{D}\ddot{C}T$. In the four-stage test compared to the three-stage test, the scoring system becomes more complex and difficult to apply in QTP. For these reasons, $T\ddot{D}\ddot{C}T$ and MCCT are preferred. The study can be expanded by testing with $T\ddot{D}T$ in better computerized testing environments for future research.

Acknowledgements

I would like to thank my graduate student Meltem Şeker for her help in data collection process. Data were collected in computer laboratory in the Faculty of Education of Dokuz Eylül University. In addition, the ethical rules of the study were evaluated by the Institute of Educational Sciences of Dokuz Eylül University. I would like to thank the administrators in the Faculty of Education and the Institute of Educational Sciences of Dokuz Eylül University for their contributions.

References

Abraham, M. R., Grzybowski, E. B., Renner, J. W., & Marek, E. A. (1992). Understandings and misunderstandings of eighth graders of five chemistry concepts found in textbooks. *Journal of Research in Science Teaching, 29*(2), 105-120. https://doi.org/10.1002/tea.3660290203

Adodo, S. O. (2013). Effects of two-tier multiple choice diagnostic assessment items on students' learning outcome in basic science technology (BST). *Academic Journal of Interdisciplinary Studies, 2*(2), 201-210. https://doi.org/10.5901/ajis.2013.v2n2p201

Al-Hamly, M., & Coombe, C. (2003). An investigation into answer changing practices on multiple choice questions with gulf Arab learners in an EFL context. In J. S. Johnson (Ed.), *Span Fellow Working Papers in Second or Foreign Language Assessment* (Volume 1, pp. 83-96). Michigan, USA: Spaan Fellow Working Papers in Second or Foreign Language Assessment.

Arslan, H. Ö., Çiğdemoğlu, 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. https://doi.org/10.1080/09500693.2012.680618

Aybek, E. C. (2012). *A Comparison of Psychometric Properties of a General Ability Test which Administered in Paper-Pencil and Computer Based Form* (Unpublished Mr.Sci. Thesis), Institute of Educational Sciences, Ankara University, Ankara, Turkey.

Bademci, V. (2006). To put an end to the oiscussion: cronbach's alpha coefficient can be used with oichotomously scored items [0,1]. *Journal of Kazım Karabekir Education Faculty, 13*, 438-446.

Baştürk, R. (2011). Impact of answer-switching behavior on multiple-choice test scores in higher education. *Journal of Measurement and Evaluation in Education and Psychology, 2*(1), 114-120.

Beck, M. D. (1978). The effect of item response changes on scores on an elementary reading achievement test. *The Journal of Educational Research, 71*(3), 153-156. https://doi.org/10.1080/00220671.1978.10885059

Bektaş, M., & Kudubeş, A. A. (2014). As a measurement and evaluation tool: written exams. *Dokuz Eylul University E-Journal of Nursing Faculty, 7*(4), 330-336.

Bernhard, J. (2000). *Improving engineering physics teaching: Learning from physics education research*. Paper presented at the meeting titled by "Physics Teaching in Engineering Education (PTEE 2000)", 13-17 June,

Budapest, Hungary.

Çakır, M., & Aldemir, B. (2011). Developing and validating a two tier Mendel genetics diagnostic test. *Mustafa Kemal University Journal of Social Sciences Institute, 8*(16), 335-353.

Casteel, C. A. (1991). Answer changing on multiple-choice test items among eighth-grade readers. *The Journal of Experimental Education, 59*(4), 300-309. https://doi.org/10.1080/00220973.1991.10806568

Chiang, W.-W., & Chiu, M.-H. (2015). Using an online assessment system to diagnose student mental models in chemistry education. *The Turkish Online Journal of Educational Technology, 14*(1), 163-178.

Cox-Davenport, R. A., Haynes, P. B., & Lawson, T. G. (2014). A mixed-methods approach to evaluating student nurses changing answers on multiple choice exams. *Journal of Nursing Education and Practice, 4*(2), 132-139. https://doi.org/10.5430/jnep.v4n2p132

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics, 69*(9), 970-977. https://doi.org/10.1119/1.1374249

Driver, R. (1981). Pupils' alternative frameworks in science. *European Journal of Science Education, 3*(1), 93-101. https://doi.org/10.1080/0140528810030109

Fraenkel, J. R., & Wallen, N. E. (2000). *How to design and evaluate research in education* (4th ed.). London: McGrawHill.

Gönen, S., Kocakaya, S., & Kocakaya, F. (2011). A study on developing an achievement test which has reliability and validity on dynamics subject. *Van Yuzuncu Yil University Journal of Education, 8*(1), 40-57.

Gürçay, D., & Gülbaş, E. (2015). Development of three-tier heat, temperature and internal energy diagnostic test. *Research in Science & Technological Education, 33*(2), 197-217. https://doi.org/10.1080/02635143.2015.1018154

Gürel, D. K., Eryılmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(5), 989-1008. https://doi.org/10.12973/eurasia.2015.1369a

Hestenes, D., & Halloun, I. (1995). Interpreting the Force Concept Inventory. *The Physics Teacher, 33*, 502-506.

Kim, Y. (2019). Partial identification of answer reviewing effects in multiple-choice exams. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12259

Li, M. N., & Yang, D. C. (2010). Development and validation of a computer-administered number sense scale for fifth-grade children in Taiwan. *School Science and Mathematics, 110*(4), 220-230. https://doi.org/10.1111/j.1949-8594.2010.00024.x

Lin, J.-W. (2016). Development and evaluation of the diagnostic power for a computer-based two-tier assessment. *Journal of Science Education Technology, 25*, 497-511. https://doi.org/10.1007/s10956-016-9609-5

Linden, W. J., & Glas, G. A. W. (2002). *Computerized adaptive testing: Theory and practice*. Dordrecht, Netherlands: Springer.

Linke, R. D., & Venz, M. I. (1979). Misconceptions in physical science among non-science background students. *Research in Science Education, 8*, 183-193. https://doi.org/10.1007/BF02359149

Lynch, D. O., & Smith, B. C. (1972). *To change or not to change item responses when taking tests: Empirical evidence for test takers*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Maier, U., Wolf, N., & Randler, C. (2016) Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Computers & Education, 95*, 85-98. https://doi.org/10.1016/j.compedu.2015.12.002

McMillan, J. H., & Schumacher, S. (2010). *Research in education: Evidence-based inquiry* (7th ed.). Boston, MA: Pearson.

McMorris, R. F., Schwarz, S. P., Richichi, R. V., Fisher, M., Buczek, N. M., Chevalier, L., Meland, K. A. (1991). *Why do young students change answers on tests?* Research Report to the State University of New York at Albany.

Milenković, D. D., Hrin, T. N., Segedinac, M. D., & Horvat, S. (2016). Development of a three-tier test as a

valid diagnostic tool for identification of misconceptions related to carbohydrates. *Journal of Chemical Education, 93*(9), 1514-1520. https://doi.org/10.1021/acs.jchemed.6b00261

Mutlu, A., & Şeşen, B. A. (2015). Development of a two-tier diagnostic test to assess undergraduates' understanding of some chemistry concepts. *Procedia - Social and Behavioral Sciences, 174*, 629-635. https://doi.org/10.1016/j.sbspro.2015.01.593

Noorbala, M. T., & Mohammadi, S. (2011). A survey on the habit to change the answers in multiple choice questions (MCQ) exams: Does the examinee benefit? *Journal of Pakistan Association of Dermatologists, 21*(4), 253-259.

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. https://doi.org/10.1002/tea.3660320106

Osborne, R. J., & Cosgrove, M. M. (1983). Children's conceptions of the changes of state of water. *Journal of Research in Science Teaching, 20*, 825-838. https://doi.org/10.1002/tea.3660200905

Özbayrak, Ö., & Kartal, M. (2012). Determination of misconceptions regarding "compounds" chapter in secondary education 9th grade by two-tier conceptual understanding test. *Buca Faculty of Education Journal, 32*, 144-156.

Özdemir, G. Y., & Kocakülah, M. S. (2016). The effect of order of different teaching activities related to mechanical waves on conceptual change. *Journal of Research in Education and Teaching, 5*(3), 150-163.

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

Sadıç, A., & Çam, A. (2015). Eight grade students' epistemological beliefs with pisa success and their scientific literacy. *Journal of Computer and Education Research, 3*(5), 18-49.

Şen, S., & Yılmaz, A. (2017). The development of a three-tier chemical bonding concept test. *Journal of Turkish Science Education, 14*(1), 110-126. https://doi.org/10.12973/tused.10193a

Şen, Ş., Yılmaz, A., & Geban, Ö. (2018). Development of three-tier electrochemistry concept test. *Karaelmas Science and Engineering Journal, 8*(1), 324-330. http://dx.doi.org/10.7212%2Fzkufbd.v8i1.1088

Tabakcıoğlu, M. B., Çizmeci, H., & Ayberkin, D. (2016). Neurosky EEG biosensor using in education. *International Journal of Applied Mathematics, Electronics and Computers, 4*, 76-78. https://doi.org/10.18100/ijamec.265371

Taber, K. S. (1999). Alternative conceptual frameworks in chemistry. *Education in Chemistry, 36*(5), 135-137.

Taber, K. S. (2017). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273-1296. https://doi.org/10.1007/s11165-016- 9602-2

Tamir, P. (1971). An alternative approach to the construction of multiple choice test items. *Journal of Biological Education, 5*(6), 305-307. https://doi.org/10.1080/00219266.1971.9653728

Treagust, D. F. (1988). Development of use of diagnostics tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159-169. https://doi.org/10.1080/0950069880100204

Uyulgan, M. A., Akkuzu, N., & Alpat, Ş. (2014). Assessing the students' understanding related to molecular geometry using a two-tier diagnostic test. *Journal of Baltic Science Education, 13*(6), 839-855.

Van Der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics, 37*(1), 180-199. https://doi.org/10.3102/1076998610396899

Weiss, D. J., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Wise, S., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

Xiao, Y., Han, J., Koenig, K., Xiong, J., & Bao, L. (2018). Multilevel Rasch modeling of two-tier multiple choice test: A case study using Lawson's classroom test of scientific reasoning. *Physical Review Physics Education Research, 14*(2), 020104/1- 020104/14. https://doi.org/10.1103/physrevphyseducres.14.020104

Yang, D. C., & Lin, Y. C. (2015). Assessing 10- to 11-year-old children's performance and misconceptions in

number sense using a four-tier diagnostic test. *Educational Research, 57*(4), 368-388. https://doi.org/10.1080/00131881.2015.1085235

Yang, D. C., Li, M. N., & Lin, C. I. (2008). A study of the performance of 5th graders in number sense and its relationship to achievement in mathematics. *International Journal of Science and Mathematics Education, 6*(4), 789-807. https://doi.org/10.1007/s10763-007-9100-0

Yang, D.-C., & Sianturi, I. A. J. (2018). Assessing students' conceptual understanding using an online three-tier diagnostic test. *Journal of Computer Assisted Learning, 35*, 678-689. https://doi.org/10.1111/jcal.12368

Yang, T.-C., Hwang, G.-J., Yang, S. J. H., & Hwang, G.-H. A. (2015) Two-tier test-based approach to improving students' computer-programming skills in a web-based learning environment. *Educational Technology & Society, 18*(1), 198-210.

Yen, Y.-C., Ho, R.-G., Liao, W.-W., & Chen, L.-J. (2012). Reducing the impact of inappropriate items on reviewable computerized adaptive testing. *Journal of Educational Technology & Society, 15*(2), 231-243.