Anatolian Journal of Education
e-ISSN: 2547-9652

October 2020 ● Vol.5, No.2
www.e-aje.net
pp. 19-36

Comparison of Threshold Values of Three-Tier Diagnostic and Multiple-Choice Tests Based on Response Time

Suat Türkoguz
Assoc. Prof., Dokuz Eylül Universty, Turkey, *suat.turkoguz@gmail.com*

This study aimed to investigate the item *Response Time Fidelity scores (RTFs)*, *Kuder-Richardson Reliability (KR_{20})* and *Cronbach's Alpha Reliability* (α) coefficients, calculate KR_{20} coefficients with *RTFs* for 30 threshold points between 1 and 30 seconds and compare these values on plots, estimate the threshold point in the literature to support these plots and conduct risk analyses with *Cox Proportional Hazards Model (Cox-HM)* on the initial data and the new data determined by the threshold value. This study is a descriptive research. The participants were pre-service science teachers at a state university in Turkey (n:115). The study was conducted in two groups. In Group-1 (n:57), a *Three-Tier Diagnostic Test ($_3TDT$)* was performed, and in Group-2 (n:58), a *Multiple-Choice Test (MCT)* was applied. Tests consisted of 63 items and chemical concepts. As a result, it was found that item response times could be used in the validity and reliability studies of measurement tools. Moreover, examinees spent less time and effort in Tier-II of $_3TDT$. In future research, response times and *RTFs* may have roles in development of $_3TDT$s about different concepts. Moreover, these may be used in Rasch and Parametric Logistic Analysis for $_3TDT$. This study may be repeated using different psychometric scales.

Keywords: tier-diagnostic, threshold, reliability, cox-hm, chemistry

INTRODUCTION

Measurement and assessment tools are used to determine the degree of realization of pre-determined learning objectives and outcomes as a result of the teaching process. *Multiple-Choice Tests (MCTs)* and open-ended questions are widely used testing and assessment tools. With these tests, information about the learning process cannot be determined while measuring the knowledge levels of the students as a result of the teaching process (Akınoğlu, 2011). Therefore, it will be beneficial to make measurement and assessment based on the process, rather than result-oriented measurement and assessment (Bektaş & Kudubeş, 2014). In addition to measuring how many questions students answered correctly in an examination, we need to know how much they have learned. *MCTs* and open-ended test have advantages and disadvantages. *Tier Diagnostic Tests (TDTs)*, which are thought to overcome the disadvantageous aspects of these tests, have become widespread. Below, the causes and justifications of the advantageous and disadvantaged aspects of *MCT* and *TDT* are listed respectively.

The reasons for the popular preference of *MCTs* are that they can be applied to a large number of people and it is easy to prepare and assess. As known, *MCTs* consist of questions and options one which is the correct one, while the others are distracters. However, these tests do not provide an idea of why the examinee has chosen that option. The lack of reasons for choosing the options of test items may constitute a problem in terms of understanding the ways in which concepts are constructed. Furthermore, it may prevent the determination of the student's misconceptions. *MCTs* are fast in terms

Citation: Türkoguz, S. (2020). Comparison of Threshold Values of Three-Tier Diagnostic and Multiple-Choice Tests Based on Response Time. *Anatolian Journal of Education*, 5(2), 19-36. https://doi.org/10.29333/aje.2020.522a

of testing and assessment, but it is disadvantageous in terms of the chance factor and not explaining the reasons of choosing particular options. Recently, various studies have been carried out about preparation in a short time, assessment and application versatility of testing tools, as well as the level of association of them with other testing tools (Gelbal & Kelecioğlu, 2007; Karahan, 2007). Therefore, the disadvantages of *MCTs* have been debated, and the popularity of *TDTs* has increased.

It may be stated that, since the 1970s, studies on scientific misconceptions have started and increased after 1980s (Driver, 1981; Tamir, 1971). There are some techniques that are used in studies on misconceptions such as interviewing, asking open-ended questions, drawing techniques, conceptual change texts, concept cartoons, mind and concept maps and *TDTs.* There have been 4158 studies in 2015, which were selected about misconceptions in science education. 273 of them were examined by purposive sampling, and it was found that 9% of them were conducted by *TDTs*, while 91% were conducted by open-ended questions, *MCTs* and interviews (Gürel, Eryılmaz & McDermott, 2015). Nowadays, *TDTs* are used to look for misconceptions, and they are consolidated by interviews. *TDTs* may be developed in Tier-II, Tier-III and Tier-IV of test forms.

Students' current misconceptions may have an impact on their responses in tests. Therefore, *TDTs* may be used to find out the reason for the student's response as well as determining the misconception regarding the topic. With composed tiers, the misconception can be determined. *MCTs are* inadequate in terms of understanding the reasoning of the student. Hence, *MCTs* may be converted into *TDTs*, and this inadequacy of *MCTs* may be overcome (Karataş, Köse & Coştu, 2003). While a sample TDT is prepared, at first, learning outcomes may be listed, concept maps may be prepared and if there is a list of existing misconceptions, these could be added to the current list of items by crosstab, and the test may be prepared by adapting it for the Bloom Taxonomy. Tier-I of $_3$*TDTs* is prepared by question items as it is in *MCT*. Tier-II and Tier-III could be added to create *TDT.* Test which is prepared this way may comprise responses which ask for reasons for the options selected in Tier-I of $_3$*TDTs* (Caleon & Subramaniam, 2010). Tier-I of $_3$*TDTs* consists of question items which *MCTs* have; Tier-II asks the examinee to explain their reasoning, and Tier-III comprises a question item that verifies the accuracy of Tier-I and Tier-II. One of the options in Tier-II should be accurate in terms of scientific propositions; however, other options' propositions should include misconceptions. In Tier-II, open-ended questions may be asked instead of providing options, or some empty space may be provided to the student in order for them to write down their own statements. Similarly, $_4$*TDTs* may also be prepared. Only after Tier-I and Tier-III, a question proposal which provides verification of the accuracy of the responses of examinees is added. Summarily, Tier-II and Tier-IV are the acknowledgement phases of $_4$*TDT.*

Scoring of *TDT* is different from those in other testing and assessment tools. In order to get a full score from a question in *TDT*, all tiers should be completed accurately. Otherwise, one cannot get a score. It is thought to provide a more accurate assessment to have such a scoring structure. However, discussions about the reliability of *TDTs*, even *MCTs*, continue (Bademci, 2007; Taber, 2017). Each tier of *TDTs* might be assessed by Logistics and Rasch models with different parameters because of their difficulty in reliability and scoring.

Identifying of the Threshold Value

By analyzing *Item Response Time (IRT)* for a test item and the worst performance critical point determined from the distribution curve according to the bimodal (0-1) response given to the test item, easy items or difficult items may be extracted by considering the threshold value, and the rate of examinees who can estimate it may be estimated (Bolsinova, De Boeck & Tijmstra, 2017). The relationship between *IRT* and item response accuracy helps to have essential inferences about the examinee.

The examinee may have passed some test items by chance or guessed them because of the difficulty of a test item or their lack of knowledge. These passed or estimated items' inaccurate coding may cause an error in terms of reliability and validity, as well as causing negative bias depending on the aptitude of the examinee (Weeks, Davier & Yamamoto, 2016). In this context, adjustments of test scores and reliability coefficients may be achieved within a certain time by looking at *IRTs* left blank or predicted (Bulut, 2015). The reliability and validity of a test may be determined by directly associating *IRT* with the response which is given as, well as using it together with the item and only searching for the item. In accordance with this purpose, duration of computerized tests might help reduce the data to be analyzed (Guo, Rios, Haberman, Liu, Wang & Paek, 2016). While reducing data, generally a threshold value is determined according to test times. This threshold value may vary between 5 and 20 seconds depending on the research. This value may change according to the test type (Numerical, verbal, power test and speed test etc.), or on the word and character count. In some studies, this threshold value is determined by large samples and graphical distributions. However, in some studies, there are estimations based merely on the total test time and test score. In some studies, it was proposed to have a different threshold value for each testing item. In this context, discussions about threshold values' variable structure have continued. Predicted, omitted or left-blank responses depending on the threshold value are recommended to be coded and analyzed differently. In this manner, scoring biases of three or four coding stages could be prevented, and this could provide useful information in terms of reliability (Weeks at al., 2016). There might be some problems encountered in terms of item difficulty and calculation of reliability coefficient according to the scoring of two different tiers in *TDTs*. In order to determine the effects of Tier-II on Tier-I of *TDTs* of it and show estimation or elimination actions, the preferences and *IRTs* in Tier-I and Tier-II of *TDTs* may be helpful in determining the threshold value. Therefore, the threshold values may be determined by looking at the transition points between the tiers, and the effects on the reliability coefficient may be examined.

Threshold Point Determination Model

A threshold level was used to determine the solution behavior index (Wise & Kong, 2005). The purpose of these studies was to explain the actions of the examinee such as rapid guess behaviors and solution behaviors. They first used distribution frequencies based on time and they accepted the threshold point as the lowest frequency value after the maximum mode values according to the plots of this distribution. Accordingly, these threshold levels were values such as between 3 and 7 seconds.

The score of the correct response for the periods after the threshold value is "1" point, while the score of the correct response for the periods before the threshold value is coded as "0" point.

$$SB_{ij}=\begin{cases} 1 \ if \ t_{ij} \geq B_j \\ 0 \ otherwise \end{cases} \qquad (1)$$

Subsequently, it is collected the correct responses to the "i^{th}" item of the examinee with the number j and rated them in the total number of items. It is defined this calculation as *Response Time Effort score (RTEs)* of the examinee.

$$RTE_i = \frac{\sum_{j=1}^{K} SB_{ij}}{K} \qquad (2)$$

Then, it is collected the solution behavior scores of all examinees and rated them in the total number of examinees. It is defined this calculation as the *Response Time Fidelity score (RTFs)* of the test.

$$RTF_i = \frac{\sum_{i=1}^{N} SB_{ij}}{N} \qquad (3)$$

RTEs and *RTFs* range from 0 to 1. Different ways may be tried in determining the threshold value.

Reliability in Power and Speed Tests and Estimation of the Threshold Value

Power tests can show higher reliability values than speed tests. Despite the low reliability of the speed test, the data can provide some evidence of criterion validity and estimates (Semmes, Davison & Close, 2011). According to Streiner (2003), reliability is reduced in four situations. (1) In tests with a time limitation, (2) When the items are graded in a simple manner from easy to difficult, (3) When the response to a question item depends on the response to another question item and (4) When the scope of the test consists of different dimensions. The reliability coefficient of tests consisting of conceptual, algorithmic and descriptive question items may be moderate (Prisacari & Danielson, 2017). When *TDTs* are considered for a power test, power tests are successive responses, and the response of one tier within the test to the other tier may lead to a decrease in the reliability of these tests.

Reliability may be increased by increasing the number of test items in tests with time limits or by changing the time limit while keeping the number of test items fixed. Reliability may also be increased with software-based automated time-limited testing strategies to respond to test items at different time limits for each item, with the support of specialized developers (Semmes et al., 2011).

Reliability evaluations can be made with binary logistics models based on *IRT*, item difficulty and threshold values. The definitions of the applied model may be questioned, and new threshold values may be determined if reliability improvement is not achieved depending on the linear or logistic model that is applied (Meyer, 2010). Item difficulty may be analyzed by considering *IRTs* of the examinee during the development of *TDTs* (Direnga, Timmermann, Presentati & Brose, 2015). *IRTs* at each *TDTs* may be modelled by linear or logistic methods, and information about the difficulty, validity and reliability of the test may be estimated along with the parameters.

Relation to IRT with the Survival Analysis of Cox-HM

The most commonly model of survival analysis is *Cox Proportional Hazards Model (Cox-HM).* Survival analysis was developed by Cox (1972) based on the regression model and is widely used today (Yetkin, 2006). In *Cox-HM*, the joint dependent variable (Survival time) can be monitoring time (death) until the time of death of a person suffering from a disease, deterioration time of a device after a specified period and correct or incorrect *IRTs* for the examinee. Furthermore, the explanatory variables can be factors such as age, gender, type of treatment and teaching method. The regression method used to reveal the cause-effect relationship between the dependent variables based on the continuous data and the independent variables based on the categorical data with the joint dependent variables is called the Cox regression method (Yetkin, 2006). In *Cox-HM*, a covariate of the variables that are dependent on survival time is formed by binary categorical variables, and the effect of the explanatory variable is explained (Zacks, 1992).

Identification of Cox-HM

In *Cox-HM*, "$\boldsymbol{x}$" is the vector of the joint dependent variables based on the results of the observed $\boldsymbol{X_i}$ events, and the survival time is "***t***". The $\boldsymbol{X_i}$ values represent conditions such as death after a given treatment, deterioration of a device from the time of reception and correct response to a question. The hazard function may be written as $\boldsymbol{h(t;x)}$ according to the joint dependent variables based on the results of the observed $\boldsymbol{X_i}$ events. Here, the risk function $\boldsymbol{h(t;x)}$ may be defined as "survival time in the risk of death", "correct *IRT* to the risk of incorrect" and "incorrect *IRT* to the risk of correct". Accordingly, *Cox-HM* is written as

$$\boldsymbol{h(t;x)=h_0(t)exp(\beta' x)} \qquad (4)$$

In this model, $\boldsymbol{\beta'}$ is the regression coefficient vector, $\boldsymbol{h_0(t)}$ is the baseline risk function when $\boldsymbol{x=0}$ (Ata, Sertkaya-Karasoy & Sözer, 2007). The $\boldsymbol{\beta}$ coefficients, which are the unknown parameters of *Cox-HM*, can be estimated using the likelihood method. The significance of the $\boldsymbol{\beta}$ coefficients is tested through

Wald test, likelihood ratio test and Score test. The joint hypothesis ***(H_0: β=0)*** is established for all three significance tests. Different distribution functions are used in decision criteria (Lee & Wang, 2003).

Purpose of the Study

Studies on the relationships between *IRT* and response accuracy in education tests are based on the studies until 1990 (Wise & Kong, 2005). In literature, there are studies on determining the threshold value by benefiting from item response accuracy with *IRT* and studies on determining *RTFs* by benefiting from this threshold value (Bulut, 2015; Meyer, 2010; Weeks et al., 2016; Wise & DeMars, 2010; Wise & Gao, 2017; Wright, 2016). In some studies, it was stated that the reliability coefficient increases in item adjustments for the threshold value (Bugbee, 1996; Kong et al., 2007). Performance determinations were made by trying to find the threshold values in visual plots between *IRT* and item response accuracy. With the help of logistic analyses, the item response performance, *IRT* and *RTFs* dependent on the threshold value were investigated using individual parameters (Bulut, 2015; Weeks et al., 2016; Wright, 2016). The majority of studies mentioned in the literature were carried out within the scope of *MCTs*. There are no studies which investigated *RTFs* with *IRT* within the scope of *TDTs* or conceptual understanding tests. In this study, it was aimed to investigate *TDTs* within the scope of *RTFs* and the critical threshold value.

Discussions on the reliability coefficient and the scoring of *TDTs* and *MCTs* are ongoing nowadays (Bademci, 2007; Taber, 2017). In particular, uncertainties remain as to how Tier-I of *TDTs* will be scored, how Tier-II is scored and how they affect each other. Additionally, there are problems in calculation of the reliability coefficients of these tests (Gürel et al., 2015; Peşman & Eryılmaz, 2010). For this purpose, it is believed that this study may be guiding for determination of the scoring and reliability coefficient problems in *TDTs* and for finding the threshold value.

This study aimed to investigate *RTFs*, *IRTs*, *Kuder-Richardson reliability (KR_{20})* coefficient and *Cronbach's Alpha reliability* (α) coefficient, calculate KR_{20} coefficients with *RTFs* for 30 threshold points between 1 and 30 seconds and compare these values on plots, estimate the threshold point in the relevant literature to support these plots and conduct risk analyses with *Cox-HM* on the initial data and the new data determined by the threshold value.

Research Questions

The problem of this study is stated as "how do KR_{20} and α coefficients, *RTFs*, *IRT* and threshold critical point change based on $_3$*TDT* and *MCT* according to confirmatory respond option in the end of test item?" In this context, the following sub-problems were used to solve the main problem;

Considering the confirmatory respond option which the examinee is sure of the response in the end of test item;

(1) How do KR_{20} coefficients vary based on $_3$*TDT* and *MCT*?

(2) How do α coefficients vary based on $_3$*TDT* and *MCT*?

(3) How do *RTFs* vary based on $_3$*TDT* and *MCT*?

(4) How do the mean test *IRTs* vary based on $_3$*TDT* and *MCT*?

(5) How do the threshold value change based on $_3$*TDT* and *MCT*?

METHOD

Research Design

This study is a descriptive research that aimed to investigate the threshold values of *$_3$TDT* and *MCTs* in a computerized testing environment based on *RTFs*, reliability coefficients and *Cox-HM*. The study was conducted with two groups determined by random sampling. In Group-1 (n:57), *$_3$TDT* was performed, and *MCT* was applied in Group-2 (n:58). The data were collected by adding a confirmation stage to *$_2$TDT* involving 44 items developed by Mutlu and Şeşen (2016). Additionally, valid and reliable electrochemistry test items were added to *$_3$TDT* according to the chemistry course content. The tests consisted of 63 items and chemistry concepts such as acids-bases, electrochemistry, thermodynamics, chemical kinetics and equilibrium. All tests were pre-tested with different pre-service teachers (n:151) and the reliability coefficient of the test was found to be 0.61.

Participants of the Study

The participants were pre-service science teachers at a state university in Turkey (n:115). The study was conducted with two groups determined by random sampling. All tests were pre-tested with different pre-service teachers (n:151).

Data Collection Process

In Group-1, *$_3$TDT* was performed, while *MCT* was applied in Group-2. Tier-I of *$_3$TDT* included the parts question items and their distractors, Tier-II included the options of misconception which made a causative inquiry related to the distractors of Tier-I, and the final tier included the stage in which the responses were confirmed. *MCT* involved the parts of question items, their distractors and the confirmatory respond in last of test's item.

The test structure forms for Tier-I of *$_3$TDT* and *MCT* were the same. These tests were performed individually in the computer laboratory. In the tests, if the examinee responded "I am sure" in the end of test item (confirmatory respond option), they went to the next question. However, the examinee was directed to the same question if the examinee responded as "I am not sure" in the end of the test item (confirmatory respond option). However, the score that the examinee got from that question decreased when the question is returned. A *Quizer Test Program (QTP)* were developed for this study. All responses and times of the examinees to *QTP* were recorded, and the examinees were not allowed to see or change their responses at the end. In figure 2, it is showed the process of time recording in QTP for *$_3$TDT* (Group-1).

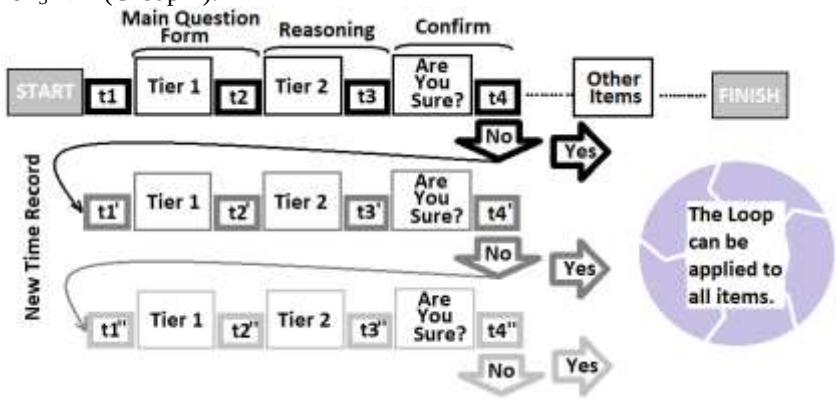

Figure 2
The Process of Time Recording in QTP for *$_3$TDT* (Group-1)

Data Analysis and Interpretation

In this study, the data were analyzed on KR_{20} and α coefficients, *RTFs* and *Cox-HM* using the SPSS and Excel programs. The data were analyzed respectively according to the sub-problem statements of the research and presented in the findings section.

FINDINGS

In the presentation of the findings, general findings are given first, and then, they were explained with detailed analysis. The findings in the tables are explained by emphasizing the important points below the tables.

General Comparison of KR_{20} Coefficients

Table 1
General Comparison of KR_{20} Coefficients for MCT and Tier-I of $_3$TDT

Decision / Group	Y	N (I.RC/II.RIC)	N′ (I.RIC/II.RC)
Group-1 (Tier-I of $_3$TDT)	0.408	0.381	0.468
Group-2 (MCT)	0.290	0.237	0.188

Y: Yes. I am sure, $N_{(I.RC/II.RIC)}$: No. I am not sure. First (I) Response is Correct, Second (II) Response is InCorrect, $N'_{(I.RIC/II.RC)}$: No. I am not sure. First (I) Response is InCorrect, Second (II) Response is Correct.

As seen in Table 1, KR_{20} coefficients for Tier-I of $_3$*TDT* were higher than that *MCT* according to confirmatory responds in the situations of Y, N and N′. The lowest KR_{20} coefficients were calculated according to confirmatory respond in the status of N for Tier-I of $_3$*TDT*, and in the status of N′ for *MCT* (For N, $KR_{20/G1}$=0.381; For N′, $KR_{20/G2}$=0.188). KR_{20} coefficient observed in the status of Y for *MCT* was higher according to confirmatory responds in the situations of N and N′ (For Y, $KR_{20/G2}$=0.290)

Table 2
General Comparison of KR_{20} Coefficients for Tiers of $_3$TDT in the Group-I

Decision / Group	Y	N (I.RC/II.RIC)	N′ (I.RIC/II.RC)
Group-1(Tier-I of $_3$TDT)	0.408	0.381	0.468
Group-1 (Tier-II of $_3$TDT)	0.470	0.526	0.549

As seen in Table 2, KR_{20} coefficients for Tier-II of $_3$*TDT* were higher than that Tier-I of $_3$*TDT* according to confirmatory responds in the situations of Y, N and N′. KR_{20} coefficients for Tier-I and Tier-II of $_3$*TDT* were high according to confirmatory respond in the status of N′ (For N′, $KR_{20/Tier-I}$=0.468 & $KR_{20/Tier-II}$=0.549).

A General Comparison of α Coefficients

In this section, the evidences of the 2nd research question was presented.

Table 3
General Comparison of α Coefficient for Tests for MCT and Tier-I of $_3$TDT

Decision / Group	Y	N (I.RC/II.RIC)	N′ (I.RIC/II.RC)	N″ Total
Group-1 (Tier-I of $_3$TDT)	0.738	0.774	0.745	0.775
Group-2 (MCT)	0.763	0.708	0.758	0.705

As seen in Table 3, α coefficient for Tier-I of $_3$*TDT* according to confirmatory respond in the status of Y was found to be slightly lower than that *MCT* (For Y, α_{G1}=0.738 & α_{G2}=0.763). The lowest α

coefficients were calculated according to confirmatory respond in the status of Y for Tier-I of *$_3$TDT*, and in the status of N″ for *MCT* (For Y, α_{G1}=0.738; For N″, α_{G2}=0.705). The α coefficient was higher according to confirmatory respond in the status of Y for *MCT* (For Y, α_{G2}=0.763).

Table 4
General Comparison of α Coefficient of IRTs for Tiers of $_3$TDT in the Group-I

Group \ Decision	Y	N (I.RC/II.RIC)	N′ (I.RIC/II.RC)	N″ Total
Group-1 (Tier-I of $_3$TDT)	0.738	0.774	0.745	0.775
Group-1 (Tier-II of $_3$TDT)	0.802	0.831	0.816	0.829

As seen in Table 4, α coefficients for Tier-II of *$_3$TDT* were slightly higher than Tier-I of *$_3$TDT* according to confirmatory responds in the situations of Y, N, N′ and N″. The highest α reliability coefficients were calculated according to confirmatory respond in the status of N″ for Tier-I of *$_3$TDT*, and in the status of N for Tier-I of *$_3$TDT* (For N″, $\alpha_{Tier\text{-}I}$=0.775; For N, $\alpha_{Tier\text{-}II}$=0.831).

General Comparison of RTFs

In this section, the evidences of the 3rd research question was presented.

Table 5
General Comparison of RTF Scores for MCT and Tier-I of $_3$TDT

Group \ Decision	Y	N (I.RC/II.RIC)	N′ (I.RIC/II.RC)	N″ Total
Group-1 (Tier-I of $_3$TDT)	0.366	0.402	0.408	0.408
Group-2 (MCT)	0.374	0.396	0.397	0.397

As seen in Table 5, *RTFs* for Tier-I of *$_3$TDT* were found to be slightly higher than *MCT* according to confirmatory responds in the situations of Y, N and N′ except in the status of Y (For Y, RTF_{G1}=0.366). The lowest RTFs were calculated according to confirmatory respond in the status of Y for Tier-I of *$_3$TDT* and *MCT* (For Y, RTF_{G1}=0.366 & RTF_{G2}=0.374). The highest RTFs for Tier-I of *$_3$TDT* and *MCT* were found according to confirmatory respond in the statuses of N′ and N″ (for N′ and N″, RTF_{G1}=0.408 & RTF_{G2}=0.397)

Table 6
General Comparison of RTF Scores for Tiers of $_3$TDT in the Group-I

Group \ Decision	Y	N (I.RC/II.RIC)	N′ (I.RIC/II.RC)	N″ Total
Group-1 (Tier-I of $_3$TDT)	0.366	0.402	0.408	0.408
Group-1 (Tier-II of $_3$TDT)	0.311	0.341	0.345	0.345

As seen in Table 6, *RTFs* for Tier-II of *$_3$TDT* were slightly lower than Tier-I of *$_3$TDT* according to confirmatory responds in the situations of Y, N, N′ and N″. *RTFs* for Tier-I and Tier-II of *$_3$TDT* were the highest according to confirmatory respond in the statuses of N′ and N″ (for N′ and N″, $RTF_{Tier\text{-}I}$=0.408 & $RTF_{Tier\text{-}II}$=0.345).

General Comparison of Mean IRTs

In this section, the evidences of the 4th research question was presented. First, to find personal mean *IRTs*, the time spent by the examinee on all test items was collected and divided by the number of questions. In order to find the mean *IRTs* in each status of *MCT* and Tier-I of *$_3$TDT*, the individual *IRTs* of the examinees were collected and divided by the number of examinees (Table 7).

Table 7
Comparison of mean IRTs for MCT and Tier-I of $_3$TDT

Decision / Group	Y	N (I.RC/II.RIC)	N′ (I.RIC/II.RC)	N″ Total
Group-1 (Tier-I of $_3$TDT)	44.40	45.03	39.62	45.70
Group-2 (MCT)	42.64	43.06	39.07	43.42

Note: Time unit is second.

As seen in Table 7, the mean *IRT*s for Tier-I of *$_3$TDT* were found to be slightly higher than that *MCT* according to confirmatory responds in the situations of Y, N, N′ and N″. In the status of N′, the mean *IRT*s for Tier-I of *$_3$TDT* and *MCT* were the lowest according to confirmatory responds in the other situations (M_{timeG1}=39.62 & M_{timeG2}=39.07). The mean *IRT*s for Tier-I of *$_3$TDT* and *MCT* were the highest according to confirmatory respond in the status of N″ (For N″ , M_{timeG1}=45.70 & M_{timeG2}=43.42).

Table 8
Comparison of mean IRTs for Tiers of $_3$TDT in the Group-I

Decision / Group	Y	N (I.RC/II.RIC)	N′ (I.RIC/II.RC)	N″ Total
Group-1 (Tier-I of $_3$TDT)	44.40	45.03	39.62	45.70
Group-2 (MCT)	24.32	25.98	22.19	26.94

Note: Time unit is second.

As seen in Table 8, the mean *IRT*s for Tier-II of *$_3$TDT* were slightly lower than in Tier-I according to confirmatory responds in the situations of Y, N, N′ and N″. The mean *IRT*s for Tier-I and Tier-II of *$_3$TDT* were the highest according to confirmatory respond in the status of N″ (For N″, $M_{time/Tier-I}$=45.70 & $M_{time/Tier-II}$=26.94).

Determination of the Threshold Value by Using the KR_{20} Coefficients and RTFs, and Verification of the Threshold with Cox-HM

In this section, the evidences of the 5th research question was presented. In this section, *IRTs* of the examinees were taken into account in response to the final tier (confirmatory respond option) of the test, and 30 different threshold values between 1 and 30 seconds were verified. *KR_{20}* coefficients and *RTFs* were found at 30 different threshold value points and plotted on a time basis. For this study, the data was updated for every 30 different threshold points. The correct responses before the threshold point were thought to be rapid guesses, and these responses were coded as "0" points. Therefore, it is possible to say that 30 different data sets were created for 30 different time periods. *KR_{20}* coefficient and *RTFs* were plotted at 30 different points, and the sudden change points were determined from the plot (Figure 3). It was assumed that the determined change point could be the threshold point. By this assumption, the initial data sets were validated by *Cox-HM* risk analysis methods with the datasets 2 second after the set threshold. The significance level of *Cox-HM* risk analysis was determined as $p<0.1$ (Table 9). According to *Cox-HM* analysis on the significance level of 0.1, when the data sets at different points in the plus 2nd second of the threshold were rearranged by assumption of rapid guessing, the omitted new data would be risky. With this logic, a difference of 2nd second would be sufficient for the threshold value point.

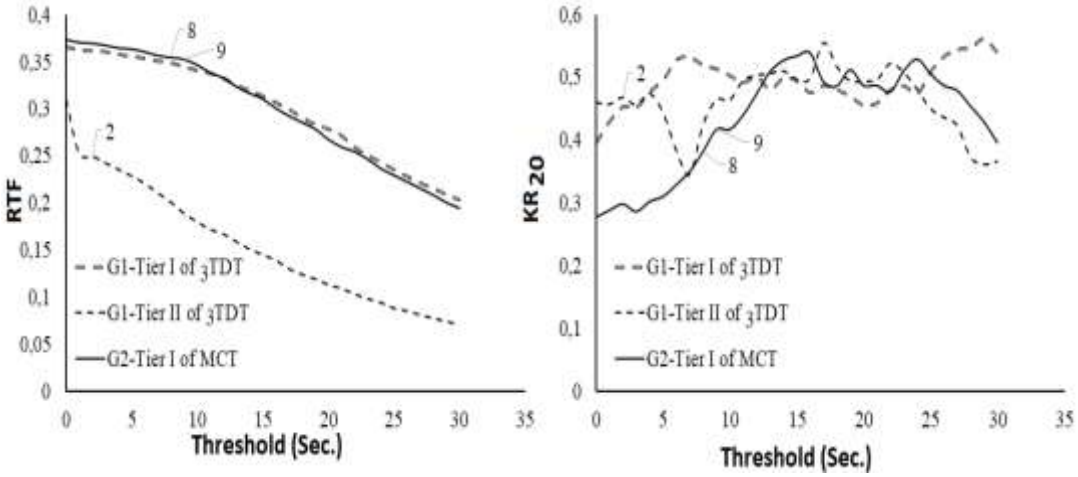

Figure 3
Threshold Plots of KR_{20} Coefficient and RTFs based on the Y State of the Tests

When the plots in Figure 3 are examined, significant changes may be observed. According to *RTFs*, the breakpoint of Tier-I of *$_3$TDT* was the 8th second, which was the 2nd second for Tier-II. Likewise, the breakpoint of *MCT* was the 9th second. Similar data could be observed for *KR_{20}* coefficients, but the significant change was in the data of *RTFs*. The plots showed a significant difference when the breakpoints of Tier-I and Tier-II of *$_3$TDT* were examined. The breakpoint in Tier-I of *$_3$TDT* was the 8th second, the breaking point in Tier-II of *$_3$TDT* was 2nd second. In this case, these points may be stated to be the threshold value. Therefore, *Cox-HM* was performed as the third validation analysis (Table 9). *Cox-HM* analysis was carried out between the initial data and the re-arranged new data in the plus 2nd second following the threshold determined by the plot. When *Cox-HM* analysis showed a significant difference in the level of 0.1, it was thought that the data in the plus period according to the threshold point posed a risk. This was believed to form the validation logic of the analyses.

Table 9
Cox-HM Analysis Results by the Y Status of MCT and Tier-I/II of $_3$TDT

Group	Wald	Sig.	β	Exp (β)	Threshold (Second)
Group-1 (Tier-I of $_3$TDT)	3.650	0.056	0.076	1.079	8+2=10th
Group-1 (Tier-II of $_3$TDT)	3.696	0.055	-0.091	0.913	2+2=4th
Group-2 (MCT)	6.097	0.013	-0.097	0.908	9+2=11th

In *Cox-HM* analysis performed by adding 2 seconds to the threshold values determined by the data on *KR_{20}* coefficients and *RTFs*, the number of data units which decreased due to the reason of rapid guesses before the 2nd second point was added to the threshold value point, may constitute error and risk in scoring. In this case, the observed thresholds are acceptable values with a 2-second margin of error. The fact that these threshold values were close to the values determined by Wise and Kong (2005) confirmed the results obtained in the study. For these data, *Cox-HM* analysis was performed for the threshold values at 30 different times. Errors and risks occurred in the subtracted data after the threshold value, but these risks were not observed in the data below the threshold value.

Similarly, *KR_{20}*, *RTF* and *Cox-HM* analyses were used for the statuses of N, N′ and N″. Due to the word limitation of the study, the findings on these statuses are not given in the tables and figures. In the status of N, according to *RTFs*, the breaking point of Tier-I of *$_3$TDT* was calculated as 8th second and the breaking point of Tier-II was calculated as the 3rd second. Similarly, the threshold point of *MCT* was found to be 9th second. It may be stated that the values of 8th, 3rd and 9th seconds obtained for

this situation were the threshold values. In the status of N′, according to *RTFs*, it was understood that the threshold point for Tier-I of *$_3$TDT* was 8th second, which was 2nd second for Tier-II. Similarly, the threshold point of *MCT* appeared to be 9th second. It may be stated that the values of 8th, 2nd and 9th seconds obtained for this status were the threshold values. In the status of N″, according to *RTFs*, it was understood that the threshold point for Tier-I of *$_3$TDT* was 9th second, which was the 3rd second for Tier-II of *$_3$TDT*. Likewise, the threshold point of *MCT* appeared to be 10th second. It may be stated that the values of 9th, 3rd and 10th seconds obtained for this status were the threshold values.

DISCUSSION

In this section, the findings are discussed with the support of the relevant literature in the order of the sub-problems of the study. The discussion of the sub-problems was carried out in separate paragraphs, respectively.

By the findings of the *KR_{20}* reliability coefficients, the coefficient of Tier-II of *$_3$TDT* was higher than the others. It was the lowest in *MCT* used as the control group. Power tests may show higher reliability values than speed tests (Semmes, Davison & Close, 2011). The level of reliability of test items containing concepts, algorithms, knowledge and definitions may be medium (Prisacari & Danielson, 2017). If *$_3$TDT* is evaluated as a power test, and *MCT* is evaluated as a speed test, it may be significant that the reliability was slightly higher in both tiers of *$_3$TDT* in comparison to *MCT*. The level of changes in the *KR_{20}* reliability coefficients was not significant by the students' response change behavior. The *KR_{20}* reliability coefficients showed a significant increase in response change behavior for Tier-II of *$_3$TDT*, but this was found insignificant for *MCT* and Tier-I of *$_3$TDT*. In measurement theories, the error rate of the measurement tools decreases due to the decrease in the chance factor, and therefore, the reliability increases (Çakır & Aldemir, 2011). In the findings, it was seen that the reliability of *$_3$TDT* was more positive than *MCT* according to the response change behavior. In this case, it may be stated that response change behavior reduces the chance factor in tests. With these results, it was observed that there is a partial dependence between the tiers of *$_3$TDT*. This partial dependency may affect the attitude to answer the test while lowering the chance factor. Therefore, the partial dependence between the tiers of the test may affect the reliability coefficient.

By the response time, in the findings of the Cronbach' Alpha reliability coefficients, the coefficient of Tier-II of the *$_3$TDTs* was slightly higher than the others. The Alpha reliability coefficient of *MCT* used in the control group was slightly higher than Tier-I of the *$_3$TDT* and slightly lower than Tier-II of *$_3$TDT*. The differences in the Cronbach's Alpha reliability coefficients according to the students' response change behavior were very little and not significant. The use of the α coefficient in the data set of *IRT* may provide significant findings in test development and internal construct validity of tests (Bugbee, 1996; Kong et al., 2007). The results of the first sub-problem were compatible with the results of the second sub-problem. With these results, it was proven again that there is a partial dependence between tiers of the *$_3$TDT* test. As Bugbee (1996) and Kong et al (2007) stated, the compatibility between the item response time and the internal structures of tests may be seen, and such a dependency is clearly understood.

By the findings on the *RTF* scores, the *RTF* rate of Tier-II of *$_3$TDT* was slightly lower than the others. Similar to the findings of this study, there are studies showing that the correct response rate of the first tier (Main Question Form) was higher than the second tier (Reasoning Form) (Li & Yang, 2010; Yang & Lin, 2015). As a reason for this, Yang and Lin (2015) stated that the first and second tiers of the tiered diagnostic tests were evaluated by the students as two separate test question forms, and that the students were more cognitively tired in the second tier of the tests. According to the students' response change behavior, the improvement in the *RTF* scores was significant. This significant development was achieved on both levels of *$_3$TDT*, provided that it was a little more than *MCT*. There are studies

showing that response change behavior does not contribute to *MCT* (Noorbala & Mohammadi, 2011; Beck, 1978; McMorris et al., 1991). In this study, it was observed that response change behavior did not contribute to the scores obtained from *MCT*, but they contributed only to *$_3$TDT*. This significant change in the *RTF* scores for *$_3$TDT* also confirmed a significant relationship between response time and correct response performance.

It was shown in the exams that the examinee's attitude, motivation and the test's significance level affected the results. Differences were found in comparisons between *IRTs* and item response accuracy of low-stakes tests and high-stakes tests (Bulut, 2015; Hambleton, Swaminathan & Rogers, 1991; Kiplinger & Linn, 1993). In this study, the test was a high-stakes test for chemistry. The students' anxiety levels were high because they were in the final exam. Therefore, the findings were obtained from students with high anxiety levels. Testing the same experiments with low-stakes tests may vary in terms of their results. Considering the findings of the mean *IRT* rates, the mean *IRT* rate of Tier-II of *$_3$TDT* was lower than the others. There was little difference between the mean *IRT* rate of Tier-I of *$_3$TDT* and the mean *IRT* rate of *MCT* used as the control group. Although there were changes in the mean *IRT* rates according to the students' response change behavior, these were not much significant. These findings were consistent with the results obtained from the solution of the third sub-problem of this study and the results of Li and Yang (2010) and Yang and Lin (2015). The low mean *IRT* rates of Tier-II of *$_3$TDT* may be attributed to the intensity of this tier with misconception options. ¾ of the options of Tier-II for each test item consisted of misconceptions. Therefore, it may have affected the answering performance. In the solution of this sub-problem, Yang and Lin's (2015) thesis that the tiers of tiered diagnostic test are considered as a separate problem form and bring a separate cognitive load to the students was confirmed.

There may be some problems in determining the threshold value with KR_{20} coefficient. When KR_{20} coefficient calculation was performed with the new data set at 30 different points between 1 and 30 seconds, an increase in the reliability coefficient was generally observed, but this technique was not sufficient to fully determine the threshold value. Reliability coefficients may be affected in updating the data with item response time and threshold value point (Wise & DeMars, 2010; Wise & Kong, 2005). Meyer (2010) suggested that the reliability coefficient can be improved with binary logistics models according to the response time, item difficulty and threshold values. This study was focused on determining the threshold point with KR_{20} rather than improving the reliability coefficient. It was observed that the reliability coefficients tested between 1 and 30 seconds changed in negative or positive ways. The KR_{20} reliability coefficient was thought to be insufficient in determining the threshold point alone. Therefore, in this study, *RTF* and *Cox-HM* data were used in addition to determining the threshold value with KR_{20}. *Cox-HM* can offer consistent information on testing the threshold value. Ranger and Ortner (2012) suggested that *Cox-HM* should be one of the routine models for *IRTs* in exams. Time-based solution behavior can be determined by the conditional dependency principle based on the time spent responding to a test item in binary (0-1) mode (Bolsinova et al., 2017; Meyer, 2010). According to the findings of the study, the examinees exhibited rapid guessing behaviors after the 8th second in Tier-I of *$_3$TDT*, and they similarly showed rapid guessing behaviors after a 3rd second in Tier-II of *$_3$TDT*. Additionally, *MCT* showed that they still exhibited rapid guessing behaviors after the approximately 8th second. When the examinees were given the option to respond to a test item more than once, there was an increase in the rapid guessing behaviors up to 8th second plus 1 second in Tier-I of *$_3$TDT*. Similarly, there were positive rapid guessing behaviors up to 3rd second plus 1 second in Tier-II of *$_3$TDT*. The threshold values determined in this study varied between 3 and 11 seconds in total. Setzer et al. (2013) determined threshold values for each test item with a graphical method in their research. In their research, an average value between 2 and 10 seconds of the threshold values was reached. The source of this difference was shown in the structure and word length of the test items. Wright (2016) emphasized that

the threshold value will depend on the type of test items and the characteristics of the exam participants, and therefore, different situations should be evaluated in determining the threshold value. Wright (2016) used the 10^{th} second as a threshold value point for the Mathematics test in their study. Wise and DeMars (2010) found a threshold value of 4 seconds on average. As it may be seen from the related literature, the threshold value point has a variable structure. However, the difference between Tier-I and -II of *$_3$TDT* was a clear and significant finding.

CONCLUSION

The results of the study are listed according to the sub-problems in order to understand the solution of the main problem.

As a result of the solution of the first sub-problem, it may be stated that giving the examinee an additional choice right from the options of test items in exams increases the reliability and internal consistency of the test. The reason for the high KR_{20} coefficient in Tier-II of *$_3$TDT* was that Tier-II consisted of short sentences and response options containing misconceptions. This may also be evidence that Tier-II of *$_3$TDT* was complementary to Tier-I of *$_3$TDT* , and this provided hinting support in the solution of test item. When the first and second response choice of the examinees were compared, a significant increase in the reliability coefficient has significantly increased in favor of the second response choice. The response choice right giving to examinees shows a positive contribution to the reliability coefficient and the internal consistency of the test. Therefore, the current status of the reliability analysis should be reconsidered in *TDTs*. Additionally, new approaches should be developed for grading *TDTs.*

In considering a result of the solution of the second sub-problem, was that *IRTs* may be used in the validity and reliability studies of measurement tools. It was previously stated that giving the right of repeating the response to the responded test increased the reliability and internal consistency of the test. However, re-responding to a test item that the examinees have already responded causes them to spend extra time and produces confusion. The reliability coefficient of Tier-II of *$_3$TDT* was higher than that Tier-I of *$_3$TDT*. Therefore, it may be stated that the responses were consistent in Tier-II of *$_3$TDT* in terms of *IRT.* The α coefficients were high in Tier-II of *$_3$TDT* because the second phase was composed of short sentences and test options containing misconceptions. In this context, despite the fact that the test item options consisted of misconceptions at Tier-II of *$_3$TDT*, the examinee did not show any abnormality in their response to test items according to *IRT.* For this reason, it may be stated that the examinees were confident in the items with misconceptions. This shows that *TDTs* are more effective in determining misconceptions because the examinee responds to Tier-II of these tests in a more stable and reasonable manner. According to *IRTs*, the reliability coefficients were affected by the repetition of the previously responded test item. Similarly, Tier-II of *$_3$TDT* was affected by the right to re-response. This finding shows that the options in Tier-I and Tier-II of *$_3$TDT* supported both the accuracy of responses and *IRT.*

However, another remarkable result of the solution of the third sub-problem, was that the examinees spent less time in Tier-II of *$_3$TDT*. Furthermore, it is important to take into account how much *IRT* will be given to the examinees in exams, as well as where *TDTs* are conducted and the differences in *IRT* during the scoring. This is because the examinee does not show equal effort in both tiers. This study provides important information about the structure of *TDTs.* Tier-II of *$_3$TDT* included the parts of misconceptions and causal association. Although the findings in Tier-II of *$_3$TDT* consisted of short sentences, the low rate of *RTF* showed that the examinees had difficulty in conceptual structures. *RTFs* of *$_3$TDT* was lower than that *MCT*. This finding showed that the examinees had difficulty at Tier-I and II tiers of *$_3$TDT*, and this led to a decrease in their *RTFs*. Re-responding to the test item, which was responded to once before, had a negative effect on *RTFs*. Granting second or third response rights to

examinees does not make a meaningful contribution to their responses. On the contrary, it leads to a disadvantage.

As a result of the solution of the fourth sub-problem, one of the other important results in this study was that the time spent in Tier-I of *$_3$TDT* was higher than that Tier-II of *$_3$TDT*. When *$_3$TDT* and *MCT* were compared in terms of *IRT*, there was a significant time difference in favor of *$_3$TDT*. This study may help researchers who aim to develop *TDT* and investigate misconceptions. If *IRTs* are examined within the scope of misconception, very significant results may be reached. Additionally, the mutual support of both tiers of *$_3$TDT* affected the spent time. The number of words in both tiers, the number of images and the question type revealed the time difference between the tiers. Additionally, the second tier generally has a denser structure in terms of its short sentence structure. Although the number of words was equivalent in both tiers, the student may spend more time understanding the question when he meets the first tier in the test. Likewise, familiarity resulting from the first tier may cause less time to pass in the second tier.

In conclusion of the solution of the fifth sub-problem, the threshold point of Tier-II of *$_3$TDT* was between 2 and 4 seconds. The threshold point of *MCT* and *$_3$TDT* ranged from 8 to 11 seconds on average. The threshold points of *MCT* and *$_3$TDT* were very close to each other. The threshold point of Tier-II of *$_3$TDT* was lower by half than Tier-I. In Tier-II, the students showed faster guessing behavior than in other situations. Additionally, the mutual support of both tiers of *$_3$TDT* affected the fast solution behaviors. Further studies on the threshold point regarding tiered diagnostic tests should be supported. In tests carried out in the computer environment, there may be disruptive variables that may affect the fast guessing behavior. These disruptive variables are exemplary situations such as test structure, order of test items, shape, visual structure and rate of arrival of the test. For this reason, descriptive and experimental studies in which these exemplary situations are controlled are needed.

MAIN FINDINGS AND IMPLICATIONS

The main findings and implications regarding the previous discussions and conclusions related to the study are presented in this section. In the study, it was found that the students showed fast guessing behavior for Tier-II of *$_3$TDT*, and the reliability coefficient of this stage was higher than that in Tier-I of *$_3$TDT*. The reason affecting these findings was that Tier-II of *$_3$TDT* consisted of short sentences and options that offered conceptual propositions. In the findings of the study, it was observed that the response change behavior did not contribute much to the students' test score, but it provided additional indicators for the validity and reliability of the test. For *$_3$TDT*, the students' confident response to the test item according to the response change behavior may provide important information for the structure and characteristics of tests.

In the study's findings, it was observed that *Cox-HM* provided valid findings in determining the threshold value of the tests, and accordingly, the fast guessing behavior of the students, but the reliability coefficients were insufficient in this regard. Nevertheless, different calculation forms such as *Cox-HM* may be used in addition to the reliability coefficients in determining the threshold value.

IRT presented important findings in the validity and reliability analysis. *IRT* may be a greater indicator in scale development studies and determination of student performances. Students may spend time answering test items based on the difficulty level of the test.

LIMITATIONS AND FUTURE RESEARCH

The results of this study should be interpreted considering various limitations. In previous studies, the accuracy of item response was examined by looking at *IRT* in cognitive tests and reading tests (Meyer, 2010; Semmes et al., 2011; Weeks et al., 2016). No time-based studies of skill tests were found. In this study, *IRTs* of *TDTs* were investigated along with item response accuracy. It is recommended to

conduct research on the types of *TDTs*, different subject areas and tier types. Studies on *RTF* and threshold determination were generally carried out with low-risk examinations over large samples. In this study, a large sample was not studied, and the sample remained limited since the study was carried out with the current students at the school. Additionally, since the tests were carried out with examinations at the end of the semester, these tests may be considered as high-risk tests. For this reason, the findings of the study may remain limited due to the fact that the examinations constituted a high-risk test. The tests were carried out with standard computers and *QTP*. Therefore, there was a limitation due to computer and test software. Since the data were collected with university students, the results and generalizable features of the study were for university students. The study may be repeated with different designs for secondary and high school students. In previous studies, logistic models related to *MCT* have been utilized. New logistic models may be established for tiered diagnostic tests.

REFERENCES

Akınoğlu, O. (2011). Öğretim ilke ve yöntemleri [Teaching principles and methods]. In Ş. Tan (Ed.), *Öğretim kuram ve modelleri [Instructional theories and models]* (pp.149-202). Ankara: Pegem.

Ata, N., Sertkaya-Karasoy, D., & Sözer, M.T. (2007). The methods used for assessment of proportional hazard assumption and an application. *Journal of Engineering & Architecture Faculty of Eskişehir Osmangazi University*, *20(1)*, 57–80.

Bademci, V. (2007). *Ölçme ve araştırma yöntembiliminde paradigma değişikliği: testler güvenilir değildir*. Ankara: Yeniyap.

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.*

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82(4),* 1126–1148. https://doi.org/10.1007/s11336-016-9537-6.

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *J. of Res. on Computing in Education*, *28(3),* 282–299. https://doi.org/10.1080/08886504.1996.10782166.

Bulut, O. (2015). Applying item response theory models to entrance examination for graduate studies: practical issues and insights. *Journal of Measurement and Evaluation in Education and Psychology*, *6(2)*, 313–330. https://doi.org/10.21031/epod.17523.

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.

Caleon, I., & 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. https://doi.org/10.1080/09500690902890130.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, *34*(2), 187–220. https://doi.org/10.1007/978-1-4612-4380-9_37.

Direnga, J., Timmermann, D., Presentati B., Brose, A., & Kautz, C. (2015, July). A statistical method for assessing teaching effectiveness based on non-identical pre-and post-tests. Paper presented at the *SEFI-2014, 42nd Annual Conference.* Birmingham, UK.

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

Gelbal, S., & Kelecioğlu, H. (2007). Teachers' proficiency perceptions of about the measurement and evaluation techniques and the problems they confront. *Hacettepe U. J. of Education, 33,* 135–145.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29(3),* 173–183. https://doi.org/10.1080/08957347.2016.1171766.

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.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. New York: Sage.

Karahan, U. (2007). *Application of alternative measurement and evaluation methods that are grid, diagnostic tree and concept maps within biology education* (Unpublished master thesis). Gazi University, Ankara, Turkey.

Karataş, F. Ö., Köse, S., & Coştu, B. (2003). Determination of students' misconceptions in science: activities through POE method. *Pamukkale University Journal of Education, 13(1),* 54–69.

Kiplinger, V. L., & Linn, R. L. (1993). Raising the stakes of test administration: the impact on student performance on the national assessment of educational progress. *Journal of Educational Assessment, 3(2),* 111-133. https://doi.org/10.1207/s15326977ea0302_1.

Lee, E. T., & Wang, J. (2003). *Statistical methods for survival data analysis*. New Jersey: John Wiley & Sons.

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.

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. (ERIC Document Reproduction Service, ED 342 803).

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement, 34(7),* 521–538. https://doi.org/10.1177/0146621609355451.

Mutlu, A., & Şeşen, B. A., (2016). Evaluating of pre-service science teachers' understanding of general chemistry concepts by using two tier diagnostic test. *J. of Baltic Science Edu., 15(1),* 79–96.

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.

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.

Prisacari, A. A., & Danielson, J. (2017). Rethinking testing mode: should i offer my next chemistry test on paper or computer? *Computers & Education*, *106*, 1–12. https://doi.org/10.1016/j.compedu.2016.11.008.

Ranger, J., & Ortner, T. (2012). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, *65(2)*, 334–349. https://doi.org/10.1111/j.2044-8317.2011.02032.x.

Semmes, R., Davison, M. L., & Close, C. (2011). Modelling individual differences in numerical reasoning speed as a random effect of response time limits. *Applied Psychological Measurement*, *35(6)*, 433–446. https://doi.org/10.1177/0146621611407305.

Setzer, J. C., Wise, S. L., Van Den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment, *Applied Measurement in Education*, 26(1), 34-49, https://doi.org/10.1080/08957347.2013.739453.

Streiner, D. L. (2003). Construct validity of the relationship profile test: a self-report measure of dependency-detachment. *Journal of Personality Assessment*, *80(1)*, 67–74. https://doi.org/10.1207/S15327752JPA8001.

Taber, K. S. (2017). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Res. in Sc Edu.*, *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.

Weeks, J. P., Von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling*, *58(4)*, 671–701.

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.

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15(1)*, 27–41. https://doi.org/10.1080/10627191003673216.

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *App. Meas. in Education*, *30(4)*, 343–354. https://doi.org/10.1080/08957347.2017.1353992.

Wright, D. B. (2016). Treating all rapid responses as errors (TARRE) improves estimates of ability (slightly). *Psychological Test and Assessment Modeling*, *58(1)*, 15–31.

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.

Yetkin, B. B. (2006). *Cox regression analysis and practice* (Unpublished master thesis). Mimar Sinan University, İstanbul, Turkey.

Zacks, S. (1992). *Introduction to reliability analysis, probability analysis and statistical methods*. New York: Springer-Verlag.