International Journal of Assessment Tools in Education
2019, Vol. 6, No. 3, 461–475
https://dx.doi.org/10.21449/ijate.561154

Published at http://www.ijate.net http://dergipark.org.tr/ijate *Research Article*

Measuring Nature of Science Views of Middle School Students

Yalçın Yalaki [1,*], Nuri Doğan[2], Serhat İrez[3], Nihal Doğan[4], Gültekin Çakmakçı[5], Başak Erdem Kara[6]

[1]Hacettepe University, Department of Primary Education, Ankara, Turkey
[2]Hacettepe University, Department of Educational Sciences, Ankara, Turkey
[3]Marmara University, Department of Mathematics and Science Education, Istanbul, Turkey
[4]Bolu Abant Izzet Baysal University, Department of Mathematics and Science Education, Bolu, Turkey
[5]Hacettepe University, Department of Mathematics and Science Education, Ankara, Turkey
[6]Hacettepe University, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: 07 May 2019
Revised: 03 August 2019
Accepted: 22 August 2019

KEYWORDS

Nature of science,
Middle school,
Assessment,
Multiple choice

Abstract: Developing scientific literacy for all students is the most often stated purpose of contemporary science education. Nature of science (NOS) is seen as an important component of scientific literacy. There are various perceptions of NOS in the science education community and NOS itself is an ever-changing construct. This makes it challenging to develop instruments for measuring understanding of NOS at different levels. Many instruments have been developed and are being developed to assess NOS learning, which indicates the importance attributed to this subject. In this study, we developed a multiple-choice test to measure NOS understanding of middle school students. The instrument was applied to 1397 middle school students. The 24 item multiple-choice test had KR-20 reliability coefficient of 0.74. A 12 item multiple-choice test created as a subset of the 24 items of the original test. This test was easier and had higher discrimination, which can provide useful measurement data about students' understanding of NOS for diagnostic or formative purposes.

1. INTRODUCTION

Developing scientific literacy for all students is a frequently stated purpose of contemporary science education curricula in many countries around the world. Nature of science (NOS) is seen as an important component of scientific literacy, and the importance of teaching NOS is emphasized by important educational policy documents and by many scholars (American Association for the Advancement of Science, 1990, 1993; Lederman, 1992; Matthews, 1998; Next Generation Science Standards Lead States, 2013; National Research Council, 1996). The importance given to the teaching and learning of NOS has increased steadily over the last 40 years and so the discussions about what NOS is and how to teach it (Lederman, 2007). Accordingly, there are different perceptions of NOS in the science education community and NOS itself has been an ever-changing construct (Matthews, 1998; Lederman, 2007; Abd-El_Khalick, 2014). This creates a problem for teaching and learning of NOS, as to teach

CONTACT: Yalçın Yalaki ✉ yyalaki@hacettepe.edu.tr Hacettepe Universitesi, Eğitim Fakültesi, Temel Eğitim Bölümü, 06800 Beytepe, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

something, a perception of it must be acknowledged so that teaching practices can be planned accordingly. The same problem also affects the assessment of NOS. To develop assessment instruments for NOS learning, a certain perception of NOS must be agreed.

Among alternative perceptions of NOS that are discussed in science education, the *consensus view*, *features of science view* and *family resemblance view* can be mentioned here as examples (Lederman, 1992; Matthews, 1994, Nola & Irzik, 2010). The consensus view argues that there is sufficient consensus on certain tenets of science that should be taught in schools. The features of science view argue that the list of consensus view tenets is arbitrary and there may be more than one list of consensus view tenets or the list can be extended. The family resemblance view argues that a more comprehensive view of science can be established based on similarities among different science fields. However, as Abd-El_Khalick (2014) says, "…the construct (or constructs) in currency in the field of science education is the NOS construct or are those constructs being assessed (p. 628)." One of the most popular and most assessed construct of NOS is the so-called consensus view construct. Lederman (2007) describes the seven tenets of NOS, which constitute the consensus view as:

1. Scientific knowledge is based on evidence: science is based on direct or indirect observation of the natural world. Science is not only based on empirical evidence, it is also based on logical inferences related to evidence. Scientific knowledge is supported through experimental data but it is never proved. Observation and inference should not be confused with each other. Scientists may have different inferences of same observations.
2. Scientific knowledge is durable but also tentative: scientific knowledge is stable but it is never certain or unequivocally true. Scientific knowledge changes through evolutionary and revolutionary processes. Scientific knowledge may change with new data or reevaluation of existing data.
3. Scientific knowledge involves subjectivity: Scientists' prior knowledge, experience, values, beliefs, education and expectations influence their study and the conclusions they reach. As a field of science matures, the level and amount of disagreements among scientists may decrease.
4. Scientific knowledge involves creativity and imagination: Scientists use their creativity and imagination in every stage of their scientific work. Creativity and imagination is an important factor that differentiates scientists from one another.
5. Science is a social activity thus it is influenced by the sociocultural environment: Political establishment, social values, economic conditions, and cultural structure influence how, what and to what degree scientists study a subject and how they apply their findings.
6. Scientific theories and laws: Theories are scientific explanations while laws are scientific descriptions of the natural phenomenon. They serve different purposes in science and there is no hierarchical relationship between them.
7. Scientific method: There is no one universal scientific method that all scientists follow that guarantees scientific discovery. Many different fields of science use many different methods to produce scientific knowledge.

The availability of various assessment instruments that are designed based on the consensus view of NOS is among the reasons for the popularity of this view in the field of science education.

Abd-El_Khalick (2014) provides a detailed landscape of NOS assessment instruments of the past 60 years. His analysis shows a trend of shifting towards open ended instruments from forced choice instruments (which include multiple-choice, Likert and agree-disagree type of instruments). In his analysis of the literature, Abd-El-Khalick shows that three NOS assessment instruments Test on Understanding Science (TOUS) (Cooley & Klopfer, 1961); Views on

Science-Technology-Society (VOSTS) (Aikenhead & Ryan, 1992); and Views of Nature of Science (VNOS) including its alternative forms (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002) dominated all the NOS assessment instruments developed in the last 60 years. He argues that these instruments collectively constituted more than 50% of the used instruments in published research in this period. Of these instruments, TOUS was a theoretically developed forced-choice instrument (meaning its items were developed from a theoretical perspective of NOS), VOSTS was an empirically developed forced-choice instrument (meaning its items were developed based on empirical data) and VNOS was an open ended instrument with corresponding interviews. This trend supports author's claim that there is a shift towards open ended instruments for NOS assessment in recent years. Abd-El-Khalick also argues that the VNOS instruments are the most popular in the field in recent years.

The forced choice instruments of NOS assessment are criticized for their shortcomings as open-ended instruments became more prevalent (Abd El-Khalick, 2014; Aikenhead, 1988; Lederman et al., 2002; Lederman & O'Malley, 1990). The stated shortcomings of forced-choice instruments can be summarized as:

- Assumption that respondents understand the forced choice item the same way as developers.
- Validity of these instruments is threatened because of difficulties in interpreting respondents' choices.
- These instruments often embody a specific theoretical model of NOS which reflect developers' philosophical positions and preferences which are imposed on respondents by the choices provided.
- Respondents' NOS views are often fragmented and lacking, which makes it difficult to capture their views through forced choice instruments.
- Likert scale instruments are particularly problematic because they generate higher levels of ambiguity.

These arguments are fair, although some of them are also valid for open ended questionnaires. For one, every questionnaire, whether open ended or forced choice, is designed with a philosophical position in mind. It would be quite difficult to create a value free instrument. Secondly, open ended questionnaires do not gurantee clearer and less fragmented responses. We can argue out of experience that it can be very challenging to interpret written answers to open ended qestions, especially if they are short or unclear. Thirdly, respondends can still understand open ended questions very differently than the developers intended similar to the forced choice questions. Another problem with the open ended questionnaires is the scorer bias, which can be an important problem if necessary precautions are not taken. On the other hand, despite the above shortcomings, forced choice instruments have some major advantages. These advantages include scalability, ease of administration, ease of scoring, and ease of data analysis. Of course these benefits do not excuse the above shortcomings, but we believe that a relatively short multiple-choice test developed based on common respondent views about NOS that are reported in the literature rather than a theoretical view of NOS could still be useful for diagnosis of student views while keeping in mind its limitations.

Parallel to the developments around the world about NOS teaching and learning and assessment of NOS views, the last two primary school science curricula prepared by the Turkish Ministry of National Education (MEB, 2013; 2018) emphasize NOS as a component of scientific literacy. To contribute to the purpose of achieving scientific literacy in schools in Turkey, a long-term professional development program for science teachers about teaching and learning of NOS was organized as a research project and it was implemented with funding from The Scientific and Technological Research Council of Turkey (TUBITAK). Our project was titled "Continuing Teacher Professional Development to Support the Teaching about Nature of

Science (BIDOMEG)" which was carried out in cooperation with Abant Izzet Baysal University, Hacettepe University, Marmara University, Ministry of Education General Directorate of Teacher Training and Development, and Bolu Provincial National Education Directorate. In this article, we reported about the ScienTest study, a multiple-choice test and its development process which was aimed at measuring the NOS views of middle school students.

For the main data collection in the study, we used the VNOS-D instrument (Lederman & Khishfe, 2002), to measure seventh grade middle school students' views on NOS. The targeted NOS themes in the instrument were: 1- scientific knowledge is based on empirical evidence, 2- observation and inference are different from one other, 3- scientific knowledge is reliable but open to change, 4- creativity and imagination play an important role in the emergence of scientific knowledge, 5- scientists can be subjective during scientific studies, 6-scientific models are abstract and approximate versions of the reality. The sixth theme was added to the instrument as an extra dimension. The VNOS-D instrument is an open ended instrument and requires written answers. The written answers then need to be coded by various scorers. In large scale applications, the application of this instrument bears many difficulties. The open ended questions require young students to express their opinions in writing, which is often challenging and most of them prefer to write short answers as we very often observed during the study. Also coding written answers, especially short answers, can be challenging as it is often not clear which category the answer falls into. With close to 1400 participants, conducting interviews were not practical to clarify students' ideas. In addition, scorer errors, which can occur when the test items are evaluated and coded by different individuals, can affect the reliability of the instrument. Given the difficulties of implementing the VNOS-D in large scale, we decided to develop a multiple-choice instrument that measured the same six themes which can be applied from fifth to eighth grade levels. Development of this instrument was not a planned outcome of the project from the beginning, but rather the idea of developing such an instrument appeared with the challenges of large scale measurement.

2. METHOD

A 24 item test for the six NOS themes mentioned above was developed (see Appendix 1). Table 1 shows the targeted NOS themes and the corresponding items that were designed to measure student views about these themes. The items had three choices, each representing a different NOS understanding. One of these options represented an understanding at the targeted level (informed level in VNOS terms). The other choices were selected based on alternative conceptions of NOS that were reported in the literature. In order to confirm the validity of the test, four experts (the authors) have reviewed items and proposed changes, which were implemented.

Table 1. Targeted NOS themes and corresponding items in ScienTest

Items		Related NOS theme
1	13	Science is based on empirical evidence.
2	14	
3	15	
4	16	Scientific knowledge is tentative.
5	17	
6	18	Scientific knowledge involves creativity and imagination.
12	24	
7	19	Scientific knowledge involves observation and inference.
8	20	
9	21	Scientific knowledge involves subjectivity.
10	22	
11	23	Scientific models do not reflect exact reality.

2.1. Study Group

For the pilot study of ScienTest, two different forms of the instrument were prepared at the beginning of the study in an effort to find out which form is more reliable and these forms were applied to a total of 183 middle school students. One of the forms had 12 items and six choices and students were asked to mark all of the choices they preferred. The other form had 24 items and three choices, one of them being the desired choice. The analysis of data showed that the three-choice 24 item form was more reliable than the 12 item form (KR-20 0.641 vs 0.615). After the item analysis and validity assessments were made on the 24 item multiple-choice form, some items were revised and the final form was applied to 1397 middle school students in the spring semester of 2013. The reliability coefficient of the test (KR-20 value) was determined to be 0.740 in this large-scale application. The detailed data analysis is explained below.

2.2. Data Analysis

ScienTest's item and test parameter estimates were made according to both Classical Test Theory (CTT) and Item Response Theory (IRT). The analysis results were cross-checked based on two theories. In the analysis of data, TAP 14.7.4 software was used for analysis based on the Classical Test Theory, and IRTPRO software was used for the analysis based on the Item Response Theory. Parallel analysis based on the tetrachoric correlation matrix for factor analysis was performed with FACTOR 10.6.01 software program. Finally, Confirmatory Factor Analysis was performed with MPlus7 software program.

Data analysis took place in several stages. Firstly, exploratory factor analysis was carried out with Parallel Analysis method for the 24-item ScienTest to examine the structure of the data. After factor analysis, model-data fit of the IRT models (Rasch, 2PL and 3PL) were examined. The -2Log-Likelihood values were used to examine the fit of the IRT models' fit with data and the -2Log-Likelihood value differences for each model were compared with the chi-square difference test. Afterwards, CTT and IRT analysis were made. At the last stage, the best performing items in the two halves of ScienTest were selected, taking into account the fact that each dimension would be measured, and a 12-item final test was established and a confirmatory factor analysis was applied on this test.

3. RESULT / FINDINGS

In this section, the results of factor analysis, model-data fit, item and test analysis based on CTT and IRT, and confirmatory factor analysis of the final test are presented.

3.1. Factor Analysis

As a result of the Parallel Analysis based on the tetrachoric correlation matrix applied to the 24-item ScienTest to determine the test structure, when the values of KMO and Bartlett were analyzed it was found that the data structure was suitable for factor analysis (KMO = 0.845, Bartlett's test of sphericity $\chi 2$ = 2957.5, p = 0.00010). The results of the factor analysis are presented in Table 2.

The eigenvalue for the first factor is approximately 3.2 times the eigenvalue for the second factor. Parallel analysis result also suggests a one-factor structure. According to this result, it is accepted that the data has a one-factor structure. One factor explains the 22.3% of the variance of test scores. Factor loadings were found to be in the range of 0.102 and 0.668. Büyüköztürk (2012) suggested that factor loadings should be at least .30. Items 2, 13, 16 and 17 have factor loadings below 0.30.

Table 2. Factor analysis results for ScienTest

Item	*Factor loadings*	*Item*	*Factor loadings*
1	0.418	13	0.150
2	0.183	14	0.592
3	0.376	15	0.653
4	0.400	16	0.102
5	0.459	17	0.260
6	0.556	18	0.668
7	0.455	19	0.512
8	0.382	20	0.474
9	0.400	21	0.587
10	0.499	22	0.445
11	0.440	23	0.300
12	0.613	24	0.369
Explained variance ratio	%22.3		

3.2. Model-Data Fit

Three different IRT models (Rasch, 2-Parameter Logistics and 3-Parameter Logistics) were used for data analysis based on IRT. The -2Log-Likelihood, AIC and BIC values were used to determine which model has the best fit on ScienTest data. The model with the smallest AIC and BIC values is interpreted as the best model (Wang & Liu, 2005). The obtained values were presented in Table 3.

Table 3. Model-data fit indexes

	-2Log-Likelihood	*AIC*	*BIC*
Rasch Model	41901.37	41949.37	42075.18
2 PLM	41400.89	41496.89	41748.51
3 PLM	41141.70	41285.70	41663.13

The model with the lowest Log-Likelihood, AIC and BIC values is the 3 parameter logistic model (3 PLM). In addition, for each model, the difference of -2loglikelihood values were compared with the chi-square difference test to investigate model fit. At this point, the χ2 value on the χ2 table was found first ($\chi2_{(24,\ 0.05)} = 36.415$) and the value of -2Log Likelihood was compared with that χ2.

- 1PLM-2PLM: χ2 = (-2Log-Likelihood$_{1PLM}$) – (-2Log-Likelihood$_{2PLM}$) = 500.48> 36.42, the 2-Parameter Logistic Model (2 PLM) is more significant than the Rasch model, that is, 2PLM shows better fit with the data.
- 2PLM-3PLM: χ2 = (-2Log-Likelihood$_{2PLM}$) – (-2Log-Likelihood$_{3PLM}$) = 259.19> 36.42, the 3-Parameter Model provides a better fit than the 2-Parameter Model. The fact that this value is above the critical value indicates that the analysis of data with 3 PLM will make a significant difference.

As a result of the analysis, it is found that the best fitting model with data is 3-Parameter Model.

3.3. Descriptive Statistics

The total number of questions in the test was 24 and the number of respondents was 1397. Table 4 presents the descriptive statistics calculated for the ScienTest in the framework of CTT. The highest score that can be taken from this test is 24, where the correct answers are marked as '1' and the wrong answers are marked as '0'. The reliability coefficient calculated with the KR-20 method is .740. The reliability calculated by the Sperman-Brown split-half method (odd-even) was calculated as 0.763 and the McDonald's Omega was calculated as 0.83. McDonald's Omega is a reliability coefficient used for congeneric measurements, which is defined as measurements that items' factor loadings are different (McDonalds, 1985). When the factor loadings in Table 2 are examined, it can be said that the factor loadings differ, so the congeneric measurement can be applied. In this case, the appropriate reliability coefficient to use was McDonald's Omega, which had a value of 0.83. This value indicates that the test is reliable at an acceptable level (> .70) (Nunnaly, 1973). The test scores' mean was calculated as 13.081 (standard deviation = 4.355). The average difficulty value of the test was 0.545, while the average discrimination value was 0.378. When the values of the skewness and kurtosis were examined, it was found that they varied between -2 and +2; this is considered to be a sign of the normal distribution test scores (Pallant, 2005).

Table 4. Descriptive statistics for ScienTest

Sample size	1397	Kurtosis	-0.538
Mean	13.081	Reliability	0.740 (KR-20)
Average Difficulty	0.545		0.763 (Split Half)
Average Discrimination	0.378		0.83 (McDonald's Omega)
Median	13.00	Variance	18.965
Skewness	0.160	Standard Error	2.22

3.4. Item Analysis

The item difficulty and discrimination parameters for 24 items in the test are presented in Table 5. The item discrimination given in the framework of CTT has been interpreted using the point biserial correlation value. In the context of IRT analysis, 'a' parameter means item discrimination, 'b' parameter means item difficulty, and 'c' parameter is interpreted as guessing parameter.

When the analysis results in Table 5 are examined, it is seen that the item difficulties according to the CTT are between 0.26 and 0.74, and the average difficulty of the test is 0.545. This value can be regarded as an indicator that the test is at medium difficulty (Haladyna, 2004). The item discrimination values are found to be in the range of 0.17 and 0.44. Ebel (1965) stated that items with discrimination values smaller than 0.20 should be thrown away or completely replaced, while items with discrimination values between 0.20-0.29 should be corrected. Items with a discrimination value of .30 and above have a sufficient level of discrimination. According to this criterion, items numbered 2, 13 and 16 should be reexamined.

The IRT item analysis was interpreted only considering the 3 PLM estimates. The values in Table 5 show that the value of "a", which represents item discrimination, varies between 0.38 and 5.32. The "b" values, representing item difficulty, were in the range of -1.09 to 3.40; the "c" parameter, which is the indicator of guessing, varied between 0.10 and 0.36. According to the difficulty parameter, the items are spread over a wide range; it can be said that the test contains questions from all levels. Given the average difficulty parameter (b_{mean} = 0.51), it was found that the test's average difficulty was above the mean (b> 0). The estimated average guessing parameter was 0.23.

Hambleton and Swaminathan (2010) recommend paying attention to the items in cases that standard error value of item parameters exceed 1. When the standard error values for the parameters are examined, the standard error value (sh_{13} = 9.38, sh_{16} = 9.38) for parameter "b" for item 13 and parameter "a" value for item 16 is found to be higher than 1. According to the results of the IRT analysis, items 13 and 16 should be revised.

Table 5. Item analysis results based on CTT and IRT

	CTT		IRT					
			1 PL	2 PL		3 PL		
	Difficulty	Discrimination	b	a	b	a	b	c
1	0.55	0.41	-0.23	0.74	-0.30	1.73	0.78	0.36
2	0.26	0.23	1.21	0.33	3.24	2.96	1.85	0.21
3	0.74	0.31	-1.20	0.66	-1.73	0.72	-1.09	0.21
4	0.61	0.37	-0.52	0.72	-0.69	0.86	0.05	0.20
5	0.7	0.38	-0.96	0.87	-1.10	0.97	-0.64	0.18
6	0.68	0.44	-0.88	1.18	-0.82	1.66	-0.18	0.28
7	0.51	0.41	-0.05	0.75	-0.06	1.07	0.60	0.22
8	0.45	0.37	0.24	0.64	0.36	2.43	1.18	0.33
9	0.55	0.37	-0.23	0.70	-0.31	0.92	0.35	0.20
10	0.67	0.42	-0.82	0.94	-0.89	1.02	-0.53	0.15
11	0.58	0.39	-0.40	0.78	-0.50	1.10	0.23	0.24
12	0.66	0.48	-0.76	1.34	-0.67	1.63	-0.29	0.19
13	0.4	0.21	0.47	0.23	1.83	0.32	3.40	0.19
14	0.72	0.43	-1.08	1.23	-0.99	1.29	-0.79	0.11
15	0.7	0.49	-0.97	1.53	-0.79	1.66	-0.61	0.10
16	0.35	0.17	0.71	0.17	3.60	2.62	2.26	0.33
17	0.31	0.29	0.91	0.45	1.84	1.65	1.87	0.23
18	0.69	0.5	-0.94	1.65	-0.74	2.11	-0.38	0.20
19	0.53	0.44	-0.13	0.96	-0.15	1.36	0.37	0.20
20	0.49	0.41	0.06	0.84	0.07	1.73	0.74	0.26
21	0.57	0.47	-0.33	1.20	-0.31	1.75	0.16	0.21
22	0.55	0.4	-0.22	0.80	-0.27	1.90	0.70	0.34
23	0.35	0.33	0.70	0.53	1.22	5.38	1.29	0.27
24	0.5	0.36	-0.02	0.63	-0.03	1.62	1.00	0.34
AVERAGE	*0.545*	*0.378*				*1.69*	*0.51*	*0.23*

In the ScienTest, the first 12 questions and the second 12 questions are designed to measure the same attributes. In other words, it can be said that the first 12 items and the next 12 items were designed as a parallel test. In addition to the analysis and results obtained on the 24-item form of the test, the detailed analysis made on the 12-item two halves may be more informative. Since the two halves of the test measure the same attributes, two parallel tests were analyzed with independent exploratory factor analysis (Parallel Analysis based on the Tetrachoric Correlation Matrix) and the findings are presented in Table 6.

First, the results of the factor analysis for the test consisting of the first 12 questions were examined. When KMO and Bartlett values were examined, it was found that the data structure was appropriate for factor analysis (KMO = 0.764, Bartlett's test of sphericity $\chi 2$ (66) = 971.3, p = 0.000010). When we look at the eigenvalues, it is seen that the eigenvalues of three factors have a value higher than 1, but the eigenvalue of the first factor is about 2.6 times the eigenvalue of the second factor. Parallel analysis result also suggests a one-factor structure. The factor loadings for all items load between 0.201 and 0.588 on the first factor, which supports the one-factor structure conclusion. On the other hand, according to the assumption that 12 items are collected in one factor, the explained variance ratio was calculated as 26.2%. When Table 6 is examined, the factor loadings of only 2 items from the first 12 items are below the critical value of .30. According to Reckase (1979), explained variance ratio of 20% is enough.

When the results of the factor analysis for the test consisting of the last 12 questions were examined, the values of KMO and Bartlett were found to be at the desired levels (KMO = 0.747, Bartlett's test of sphericity $\chi 2$ (66) = 1171.1, p = 0.000010). According to these results, it can be said that the data is suitable for the factor analysis. According to the results of factor analysis for the last 12 questions, there are three factors with an eigenvalue greater than 1. However, the eigenvalue of the first factor is about 2.6 times the eigenvalue of the second factor, and the number of dimensions proposed by the Parallel Analysis method is also 1. In this case, it was decided that the structure has one factor. The variance ratio explained by one factor was 27.1% and the factor loadings related to 12 items were varied between 0.094 and 0.648. The factor loadings of the items corresponding to items 13, 16, 17 and 23 on the whole test were below .30.

Table 6. Factor analysis results for two half tests

Items	*First 12 items*	*Last 12 items*
	Factor Loads	Factor Loads
1	0.469	0.144
2	0.201	0.648
3	0.378	0.738
4	0.454	0.094
5	0.489	0.269
6	0.545	0.701
7	0.487	0.533
8	0.470	0.449
9	0.421	0.627
10	0.510	0.455
11	0.459	0.266
12	0.588	0.360
Explained variance ratio	%26.2	%27.1

When we look at Table 2, it is noted that the factor loadings of items 2, 13, 16 and 17 are low (<.30) in the factor analysis results obtained from the whole test of 24 items. In addition, according to the results of the CTT and IRT analysis, items 2, 13 and 16 need to be re-examined. The results of the exploratory factor analysis of the two 12-question half tests were also examined and it was seen that the factor loadings of the items 2, 13, 16, 17 and 23 were low. Following a concerted evaluation of all the results obtained, it was decided to form a 12-item final test consisting of the best-performing items in the two halves to investigate each behavior.

3.5. Creating a 12 Item Sub-Test

The selected items in the final test are 1, 4, 5, 8, 10, 11, 12, 14, 15, 18, 19 and 21. Factor loadings and item statistics have been considered in the selection of these items. Descriptive statistics of the 12 items selected are presented in Table 7. The total number of items in the final sub-test is 12 and the number of respondents is 1397. When the descriptive statistics of the final sub-test are examined, it is seen that the average difficulty is .617 and the discrimination is .475. The average difficulty and discrimination values of the final sub-test are both higher than the initial 24-item test. In this case, the interpretation can be made that the sub-test became easier and the test discrimination became higher. The reliability coefficient calculated according to each method is lower than the 24-item test. It is normal to encounter this situation when it is thought that the value of reliability is affected by the number of items. If the acceptable level of reliability is considered to be .70 and above, it can be said that the sub-test is reliable at acceptable levels. The skewness and kurtosis values are in the range of -2 to +2 and it is accepted that test scores' distribution is normal.

Before beginning confirmatory factor analysis, Mardia Test was conducted to check whether multivariate normality assumption was satisfied or not and it was seen that multivariate normality was not achieved (p <.05). Given that the structure of the data is categorical and not normally distributed, WLSMV and ULSMV, which are recommended estimation methods for this data (Brown, 2015), have been preferred. The results are given in Table 8.

Table 7. Descriptive statistics of the 12-item sub-test

Sample size	1397	Reliability	0.685 (KR-20)
Average difficulty	0.617		0. 716 (Split-Half)
Average Discrimination	0.475		0.798 (McDonald's Omega)
Median	8	Variance	7,409
Skewness	-0,282	Standard Error	1,528
Kurtosis	-0.685		

Table 8. Confirmatory factor analysis results for WLSMV and ULSMV techniques of final sub-test

WLSMV	df	χ2	χ2/df	RMSEA	CFI	TLI
	54	144.739*	2.68	0.035	0.961	0.952
ULSMV	df	χ2	χ2/df	RMSEA	CFI	TLI
	54	139.507*	2.58	0.034	0.960	0.952

* p < .001

At first χ2 test results were investigated among the confirmatory factor analysis results. The significance of p (p <.05) for this test is an indication that the model fit is weak. However, the chi-square statistic is a statistic that is highly influenced by the sample size. For this reason, the use of chi-square/df in large samples is recommended. If this value is between 3 and 5, acceptable fit is shown, and if it is smaller than 3, it shows perfect fit (Hair, Black, Babin & Anderson, 2009; Kline, 2015; Tabachnick & Fidell, 2013). Approximation of the goodness of fit index values (CFI and TLI) to 1 can be regarded as an indication that the model fits well with the data. For index values, 0.90-0.95 is acceptable, and above 0.95 indicates a good fit. On the other hand, if the RMSEA values indices are 0, it is perfect, and if it approaches 0, it is a good model fit. If this value is less than .03, perfect fit is accepted, if it is in the range of .03-.08, it is considered as an acceptable fit indicator (Brown, 2015; Hair et al., 2009). When all the values in Table 7 are considered together, it can be said that the one-factor model shows very good fit with the data according to the both WLSMV and ULSMV estimation methods (χ2 / sd =2.68-2.58, RMSEA =0.035-0.034, CFI =0.961-0.960, TLI = .952).

4. DISCUSSION and CONCLUSION

NOS teaching and learning is a dynamic field of study and so is the assessment of NOS learning. Many instruments are developed and continue to be developed to assess the understanding of this important construct (Abd-El-Khalick, 2014). The fact that so many instruments are being developed to assess NOS learning indicates the importance attributed to this subject. One recent example is the Nature of Science Instrument (NOSI) developed by Hacıeminoğlu, Yılmaz-Tüzün, and Ertepınar (2014). This instrument is a 13 item three point Likert scale developed to assess sixth, seventh, and eighth grade elementary students' NOS views. It focuses on four NOS themes which are "the difference between observation and inferences, tentativeness of scientific knowledge, role of imagination and creativity in scientific knowledge, and dependence of scientific knowledge on empirical evidence." The authors conducted the reliability study of this instrument with 782 students. Another example is Nature of Science View Scale (NOSvs) developed by Temel, Şen, and Özcan (2018). This instrument was developed with participation of 565 prospective teachers from different fields. The instrument is a 36 item five point Likert scale. The authors report that the final instrument measured five subscales which were 'definition and limits of science; scientific method; theory-laden and subjective nature of science; sociocultural embeddedness of science; and tentative and empirical nature of science.' All of the subscales had Cronbach alpha values above 0.70.

The above instruments were developed for an older audience than that of the ScienTest instrument. In this study, we targeted a younger audience with a multiple-choice test rather than the Likert scale which may have a higher level of ambiguity, especially with younger people. We wanted students to choose a view among given choices rather than express their degree of agreement with a view. The data analysis showed that the 24 item multiple-choice version of the test has KR-20 reliability coefficient of 0.74. As the data analysis show, a 12 item multiple-choice test created as a subset of the 24 ScienTest items created a final test with better mean difficulty and discrimination values. This 12-item sub-test was easier and had higher discrimination. This version of the test has sufficient reliability, albeit lower than the original test, and it still measures all of the NOS themes in the test. The shorter version of the test can be used with relatively younger students as it involves less reading.

In conclusion, we believe that this test can be used to collect data bout middle school students' NOS views. Multiple-choice tests have many disadvantages and also many advantages. As no measurement tool is perfect, this instrument is also not perfect, but it can provide useful measurement data about students' understanding of science for diagnostic or formative purposes.

Acknowledgements

This study was conducted under the "Continuing Teacher Professional Development to Support the Teaching about Nature of Science (BIDOMEG)" project supported by the The Scientific and Technological Research Council of Turkey (TUBITAK) with project number 111K517.

Conflict of interest

The authors declare no conflict of interest.

ORCID

Yalçın Yalaki https://orcid.org/0000-0003-0939-4766

5. REFERENCES

American Association for the Advancement of Science. (1990). *Science for All Americans*. New York: Oxford University Press.

American Association for the Advancement of Science. (1993). *Project 2061: Benchmarks for science literacy*. New York: Oxford University Press.

Abd-El-Khalick, F. (2014). The evolving landscape related to assessment of nature of science. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of Research on Science Education, Volume II* (pp. 635-664). New York: Routledge.

Aikenhead, G. S. (1988). An analysis of four ways of assessing student beliefs about STS topics. *Journal of research in science teaching*, *25*(8), 607-629.

Aikenhead, G. S., & Ryan, A. G. (1992). The development of a new instrument: 'Views on Science—Technology—Society'(VOSTS). *Science education, 76*(5), 477-491.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research (2nd ed.)*. New York, NY: Guilford Press.

Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabı [Data analysis handbook for social sciences]*. Ankara: Pegem Akademi.

Cooley, W. W. & Klopfer, L. E. (1961). *TOUS: Test on understanding science*. Princeton, NJ: Education Testing Service.

Hacıeminoğlu, E., Yılmaz-Tüzün, Ö., & Ertepınar, H. (2014) Development and validation of nature of science instrument for elementary school students. *Education 3-13, 42*(3), 258-283

Hair, J. F., Black, W. C, Babin, B.J. & Anderson, R. E. (2009). *Multivariate data analysis* (7. ed.). Upper Saddle River, NJ: Pearson Education.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Swaminathan, H. (2010). *Item response theory: principles and applications*. Norwell, MA: Kluwer Nijhoff Publishing.

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Lederman, N.G. (1992). Students' and Teachers' Conceptions of the Nature of Science: A Review of the Research, *Journal of Research in Science Teaching, 29*(4), 331- 359.

Lederman, N.G. (2007). Nature of science: Past, present, and future. In S.K. Abell &N.G. Lederman (Eds.), *Handbook of research in science education* (pp. 831–880). Mahwah, NJ: Lawrence Erlbaum Associates.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. (2002). Views of nature of science questionnaire (VNOS): Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching, 39*(6), 497-521.

Lederman, J. S., & Khishfe, R. (2002) Views of nature of science, Form D. Unpublished paper: Illinois Institute of Technology, Chicago, IL.

Lederman, N. G., & O'Malley, M. (1990). Students' perceptions of tentativeness in science: Development, use, and sources of change. *Science Education*, *74*(2), 225-239.

Matthews, M. R. (1998). In defense of modest goals when teaching about the nature of science. *Journal of Research in Science Teaching 35*(2), 161–174.

Milli Eğitim Bakanlığı [MEB]. (2013). *İlköğretim kurumları (ilkokullar ve ortaokullar) fen bilimleri dersi (3, 4, 5, 6, 7 ve 8. sınıflar) öğretim programı. [Primary schools (elementary and middle) science lesson (3, 4, 5, 6, 7 and 8th grades) curriculum]*. Ankara: MEB

Milli Eğitim Bakanlığı [MEB]. (2018). *İlköğretim kurumları (ilkokullar ve ortaokullar) fen bilimleri dersi (3, 4, 5, 6, 7 ve 8. sınıflar) öğretim programı. [Primary schools (elementary and middle) science lesson (3, 4, 5, 6, 7 and 8th grades) curriculum]*. Ankara: MEB

Nunnally, J. C. (1973). Research strategies and measurement methods for investigating human development. In J. R. Nesselroade & H. W. Reese, *Life-span developmental psychology: Methodological issues*. Oxford, England: Academic Press.

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

Next Generation Science Standards Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

Pallant, J. (2005). *SPSS survival manual: a step by step guide to data analysis using SPSS*. Maidenhead: Open University Press/McGraw-Hill,

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207-230.

Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Upper Saddle River, NJ: Pearson Education.

Temel, S., Şen, Ş. & Özcan, Ö. (2018) The development of the nature of science view scale (NOSvs) at university level. *Research in Science & Technological Education, 36*(1), 55-68

Wang, Y. & Liu, Q. (2005). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock–recruitment relationships. *Fisheries Research*, *77*, 220–225.

Appendix 1. Final version of the ScienTest (English translation)

Turkish version can be downloaded at: http://www.bilimindogasi.hacettepe.edu.tr/Biltest.pdf

1. There are statements about science below. Circle the one you think is correct.

a) Science is about the knowledge we see in the science courses.
b) Science is the new technologies that are invented and developed.
c) Science never produces a hundred percent certain knowledge, but it produce valid and reliable knowledge.

2. Which of the following do you think is a scientific discipline that is based on real experiments and observations?

a) Ufology (investigates the unknown objects seen in the sky that are called UFOs)
b) Biology (investigates living things, their structure and behavior)
c) Turkish (investigates the rules, use, reading and writing of the Turkish language)

3. Which of the following do you think is about a scientific study?

a) Creating a computer model of how a star is formed based on available data
b) Designing a new automobile model
c) Searching on the internet to learn how influenza spreads

4. Which of the following statements about scientific knowledge is true?

a) Scientific knowledge is objective; it does not change from person to person.
b) Scientific knowledge (theories, laws, hypotheses, etc.) can change with new studies and data.
c) There is only one way of producing scientific knowledge and that is the scientific method.

5. Which of the following statements about the knowledge you learned in the science and technology courses do you agree?

a) The knowledge in the Science and Technology textbooks are obtained through years of research and are unlikely to change.
b) The fact that new inventions like tablet computers and smart phones happen shows that the knowledge we read in textbooks may change one day.
c) The knowledge in Science and Technology textbooks are reliable and valid but this does not mean that this knowledge will never change in the future.

6. Do you think scientists use their creativity and imagination when they do research and experiments?

a) Whether scientists use their creativity and imagination or not depends on their field of study.
b) Science does not change from person to person; therefore, scientific studies are not influenced by creativity and imagination.
c) Scientists use their creativity and imagination in scientific studies and that is why sometimes they arrive at different conclusions.

7. It is known that all matter is made up of atoms. However, atoms' internal structure is too small to be seen even with the most powerful electron microscopes. Which of the following statements do you agree about the knowledge that scientists obtained about atoms?

a) Since we cannot see atoms, all of the diagrams and models created about atoms may not be entirely correct.
b) Knowledge about atoms are obtained through studies that have been conducted for a long time and became certain in present-day.
c) Atoms' structure can only be understood if powerful enough microscopes can be made that show their internal structure in the future; otherwise we cannot know anything about atoms.

8. Dinosaurs have lived on earth for a long time and they disappeared 65 million years ago. T-Rex is one of the most predatory dinosaurs known. How do you think scientists know that dinosaurs like T-Rex really existed and how much can they be sure of how they looked?

a) They are sure of their existence and appearance thanks to fossils and bone fragments that they found.
b) They can combine bone fragments to guess the body shape of a dinosaur, but they cannot be sure of their real look.
c) As there are pictures, models, films and documentaries about dinosaurs, scientists are sure of what they have looked like.

9. Turkey is a country that experience earthquakes often. As a result of scientific studies, scientists think that there may be an earthquake in Istanbul region in the near future. However, they expressed different opinions about the time and intensity of such an earthquake. Even though scientists have the same information, why do you think they have different opinions about this issue?

a) They have different opinions because there is no valid theory about this subject.
b) They have different opinions because they have not come together and thoroughly discussed the issue.
c) They have different opinions because they have different backgrounds, experience, knowledge and means.

10. The relationship of technological developments such as cell phones with cancer is being discusses. Studies about this relationship provided conflicting results. Some experts report that extensive use of cell phones increase the risk of cancer, while others could not find a relationship between cell phones and cancer. What do you think is the reason fort these conflicting results?

a) These kinds of conflicts may appear in the beginning of research, but eventually they are definitely resolved.
b) Scientists' preferred methods, their inferences and judgements may be different which may lead to conflicting results.
c) Science is objective and these kinds of conflicts should not exist. So one of the studies must be wrong.

11. There are different models that you use in schools (for example a model that shows internal organs, a cell model, and a DNA model, etc.). Scientists also use models when they investigate the nature. How much do you think these models reflect reality?

a) These models help us understand science subjects, but they are not real, they are only simplified versions of reality.
b) Models of very complex systems may not reflect reality, but models of simple things reflect reality.
c) If a model is well prepared, it reflects the reality.

12. At which stage/stages of their research (for example planning, doing an experiment, analyzing data, interpreting data, reporting the results, etc.) do you think scientists use their creativity and imagination?

a) All stages of a research can be done in different ways and creativity and imagination can play a role in all of the stages.
b) Creativity and imagination may play a role when planning a research but other than that it is not important.
c) I don't think they use their creativity and imagination in any stage of their research.

13. Which of the following statements about science do you think is correct?

a) Science provides certain, accurate knowledge.
b) Science allows us to reach the reality as a result of many studies conducted.
c) Science is based on experiments, observations, and logical inferences based on them.

14. Which of the following is a scientific field that is based on experiments and observations?

a) Mathematics (investigates numbers, shapes, geometry, operations, functions, etc.)
b) Chemistry (investigates matter, properties of matter, and how matter changes)
c) History (investigates the past events, people, institutions, and their relationships)

15. Which of the following do you think is a scientific study?

a) Conducting a controlled experiment on subjects to find out the effect of a medicine on cancer
b) Solving a very hard mathematical problem
c) Preparing educational TV programs about scientific topics such as genetic engineering

16. Which of the following statement about scientific knowledge do you think is correct?

a) Scientific knowledge is type of knowledge that is proven to be certain by experiments.
b) The differentiating feature that separate scientific knowledge from other knowledge is its testability.
c) All scientific knowledge becomes law after being proven in time.

17. Which of the following statements about knowledge in science textbooks do you think is right?

a) Only proven knowledge enters science textbooks, unproven knowledge cannot be in these books.
b) Some knowledge in textbooks may change in the future, but knowledge that became law never change.
c) All of the information in the science textbooks can possibly change in the future.

18. Do you think scientists use their creativity and imagination in their research and experiments?

a) Some scientists obtain better results in their research than others because of their creativity and imagination.
b) As long as scientists use the scientific method, they do not need to use their creativity and imagination.
c) Creativity and imagination are ambiguous concepts and they have no place in science.

19. Atoms are building blocks of all matter, but it is not possible to see atoms' internal structure. So which of the following statements do you agree about scientists' knowledge about atoms?

a) As atoms' pictures can be drawn and their models are made, they know the exact structure of atoms.
b) Even if atoms are very small, scientists discover their real structure with the experiments they conducted.
c) Even if atoms cannot be seen, thanks to experiments and observations, information about their structure can be obtained.

20. After living on earth for a long time, dinosaurs disappeared 65 million years ago. Which of the following statements about how much scientists are sure of their real appearance do you agree?

a) Scientists can be sure about the appearance of well-known dinosaurs like T-Rex and dinosaurs whose bones are found in abundance.
b) Based on bone and fossil findings and also with some imagination, they can only make comments about how dinosaurs looked.
c) Thanks to advancements in technology, the real appearance of dinosaurs will certainly be determined in the future if not today.

21. There are many fault lines that pass through Turkey. As a result of studies conducted about earthquakes, scientists think that in the near future there may be an earthquake in the Marmara Sea. However, they disagree on the time and severity of a possible earthquake. Why do you think scientists have different opinions even though they have the same information?

a) Scientists have different creativity and imagination and because of this, they always have differences in their opinions.
b) Earthquake research is relatively new and because of this, they have different opinions.
c) They have different opinions, because there aren't enough seismographs (tools that measure severity of an earthquake).

22. Whether cell phones cause cancer or not is being debated. Some researchers argue that extensive use of cell phones may cause cancer, while others could not find a relationship between cancer and cell phones. What do you think is the reason for this conflicting situation?

a) If scientists compare and discuss the data they collected, they will always arrive at the same conclusion and the conflicts will disappear.
b) It is normal for scientists to come to different conclusions about a subject. New studies may support one of these conclusions more so than others.
c) If scientists apply the scientific method correctly, they will always arrive at the same conclusions and these types of conflicts will not happen.

23. There are models such as cell model, DNA model, and atom model that are being used in science courses. Scientists use and produce various models as they investigate the nature. How much do you think these models reflect the reality?

a) The models being used in schools may be simple, but the models that scientists make exactly reflect the reality.
b) If enough attention is given to details, models will perfectly align with the reality.
c) Models are limited with the assumptions, creativity and means of people who created them and they never exactly reflect the reality.

24. In which of the stage/stages of research such as planning, experimenting, observing, analyzing data, interpreting data, reporting results do you think scientists use their creativity and imagination?

a) Scientists use their creativity and imagination more or less in every stage of their research.
b) Creativity and imagination is used in technological work and development of new products rather than science.
c) Scientific method is evident and there is no need for creativity and imagination in its application.