International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 515–533
https://doi.org/10.21449/ijate.993805
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Comparison of Inter-Rater Reliability Techniques in Performance-Based Assessment

Sinem Arslan Mancar[1,*], **H. Deniz Gulleroglu**[2]

[1]Independent Researcher
[2]Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Educational Measurement and Evaluation, Ankara, Türkiye

ARTICLE HISTORY
Received: Sep. 10, 2021
Revised: Jan. 29, 2022
Accepted: May 17, 2022

Keywords:
Inter-rater reliability, Performance-based assessment, Generalizability theory, International baccalaureate diploma programme, Scientific literacy.

Abstract: The aim of this study is to analyse the importance of the number of raters and compare the results obtained by techniques based on Classical Test Theory (CTT) and Generalizability (G) Theory. The Kappa and Krippendorff alpha techniques based on CTT were used to determine the inter-rater reliability. In this descriptive research data consists of twenty individual investigation performance reports prepared by the learners of the International Baccalaureate Diploma Programme (IBDP) and also five raters who rated these reports. Raters used an analytical rubric developed by the International Baccalaureate Organization (IBO) as a scoring tool. The results of the CTT study show that Kappa and Krippendorff alpha statistical techniques failed to provide information about the sources of the errors causing incompatibility in the criteria. The studies based on G Theory provided comprehensive data about the sources of the errors and increasing the number of raters would also increase the reliability of the values. However, the raters raised the idea that it is important to develop descriptors in the criteria in the rubric.

1. INTRODUCTION

The characteristics that individuals should possess in the 21st century have become highly differentiated and diversified, compared to previous centuries. A new generation of learners should be capable of collaborating and managing the complexities of the global world. Getting ahead in 21st century society requires acquiring a set of critical skills and adopting specific characteristics. Apart from the general knowledge and skills that learners should have, they are expected to be "global citizens" who have the ability to use basic sciences to solve the problems encountered in daily life by applying their advanced critical thinking, problem-solving, productivity, creativity, communication, awareness of ethical rules, information literacy, technology literacy, global awareness, innovation, and collaboration skills effectively (Ananiadou & Claro, 2009; MEB, 2016; National Research Council, 2012; OECD, 2017; Partnership for 21st Century Learning, 2007). This means that learners need to communicate effectively, think critically, analyse local and global issues, challenges and opportunities, become information literate, reason logically, interpret scientific data in terms of cognitive

*CONTACT: Sinem Arslan Mancar snmars89@gmail.com Independent Researcher

e-ISSN: 2148-7456 /

competencies, play a key role as a team member, cooperate with others, be aware of the importance of social impact in terms of interpersonal competencies, be aware of the significant impact of ethics, and have intellectual openness and self-regulation in terms of intrapersonal competencies (Collins, 2014; IBO, 2014a; IBO, 2014b; IBO, 2014c; Marzano & Heflebower, 2012; National Research Council, 2012; Schleicher, 2015; Trilling & Fadel, 2009; Uçak & Erdem, 2020).

Today's learners have started to live in the information age because of growing up in a fast-paced digital world. Moreover, technological innovations have accelerated the transmission and processing of information. These aspects related to information have also revealed the "information literacy" and the concept of the term has been determined as one of the important learner skills. Information literacy was defined by Paul G. Zurkowski, the president of the Information Industry Association in 1974 as "the person who uses scientific information resources effectively to reach a knowledge-based solution related to problems, and who has the skills to use various information sources" (p.6). Information literate individuals who have developed scientific thinking skills and who could use science for personal and social purposes are candidates for being a scientific literate. The definition and components of information and scientific literacy concepts have evolved together with the times as one of the most fundamental and continuous parts of the scientific process is information literacy (Klucevsek, 2017).

Scientific literacy is defined as the use of scientific knowledge by a global citizen in order to identify science-related issues, draw conclusions with a scientific method, and utilize that knowledge for the benefit of society and the individual (Bybee, 1997; Holbrook & Rannikmae, 2009; Hurd, 1998; Maienschein, 1998; Nbina & Obomanu, 2010; OECD, 2017; Turgut 2007). Being a qualified scientific literate requires being able to explain the facts and concepts scientifically, develop and evaluate scientific inquiry methods and interpret the findings logically (MEB, 2016; OECD, 2017; Rychen & Salganik, 2003). It is known that scientific literacy skills are tested in standardised tests globally such as the International Mathematics and Science Trends Research (TIMSS) and the International Student Assessment Program (PISA) and in national or international educational programmes such as International General Secondary Education Certificate (IGCSE), Advanced Level (A-Level) and International Baccalaureate (IB) (IBO, 2014a; IBO, 2014b; IBO, 2014c; Mullis & Martin, 2017; OECD, 2017; Syllabus Cambridge IGCSE Global Perspectives, 2015). In terms of scientific literacy, TIMSS tests are based on a comprehensive analysis of mathematics and science curricula and mainly focus on facts and processes while PISA tests measure mathematics and scientific literacy skills, as well as the application of these skills to real-life situations (OECD, 2017). In science literacy skills of A–level and the International Baccalaureate (IB) Diploma Programme (DP) (also known as IBDP) learners are assessed by both performance assessment and final exams (IBO, 2014a; IBO, 2014b; IBO, 2014c; Cambridge International Examinations, 2015).

Scientific literacy is assessed by national or international tests and educational programmes, as abovementioned. While the characteristics today's learners should have are so diverse, it is inevitable that the assessment and measurement tools be used to assess the relevant characteristics and also change, transform or diversify. Dietel, Herman, and Knuth (1991) define assessment as "any process and test used to learn more about the current level of knowledge possessed by the learner" (online document). Testing is defined as a "single-occasion, unidimensional, time- based" usually in the form of a multiple choice or short answer (Law & Eckes, 1995). Learners were assessed only by true-false tests, multiple choice tests and short-answer tests for a long time. Currently, due to the nature of the 21st century learner, it is realized that there is not only one way of gathering information about learner learning as alternative assessment tools are supportive approaches to the assessment of learner's higher-order skills with the traditional assessment tools (Coombe *et al., 2012)*. Furthermore, testing is

viewed as just one aspect of assessment, and the term "assessment" is widely used (Kulieke *et al.*, 1990).

In alternative assessment, there are three approaches: Authentic, performance based, and constructivist (Simonson *et al.,* 2000). Similarly, Reeves (2000) suggests that three key approaches be used in assessment; namely, cognitive, performance, and portfolio. As researchers and educators use the terms "performance based assessment," "alternative assessment", and "authentic assessment" interchangeably, performance based assessment will be used to refer to alternative assessment and discussed throughout this study. Tasks and context in performance based assessment are more closely aligned with learners' context in the classroom and in real life situations. In other words, the nature of the task and context in which assessment takes place represents real life problems or issues (Coombe *et al.,* 2012*)*. Therefore, performance based assessment is a valuable tool to observe learners' skills as to how to use science knowledge to solve problems encountered in daily life as it is compatible with the nature of scientific literacy (Kutlu *et al.,* 2008). Performance tasks and contexts enable learners to apply their skills to various simulations related to real life simulations.

Performance based assessment tools are also based on the process of learning which focuses on the growth and the performance of the learner. According to Law and Eckes (1995), if learners fail to perform a given task or context at a specific time, they can still demonstrate their abilities at a later stage and in a different situation as it is not a one-time test. Furthermore, performance based assessment focuses more on the process than on the product (pass or fail), which makes assessment formative. As a result, teachers may monitor and assess their learner's strengths and weaknesses in a variety of scenarios and can improve their syllabi based on the needs of the learners (Law & Eckes, 1995; Reeves, 2000). For this reason, performance based assessment also tends to prioritize more individualized and constructive feedback.

The key feature of performance based assessment is that the learners need to create their own work such as projects, portfolios, reports, experiments, or performance, which is scored against specific criteria (Kutlu *et al.,* 2008; Simonson *et al.,* 2000). In this context, various assessment tools such as checklists, grading scales, and rubrics are used by educators and researchers (Aktaş & Alıcı, 2017). A rubric that includes the specification of the skill being examined and the constituents of various levels of performance success is defined as a set of achievement criteria with the highest and lowest degrees (Callison, 2000). Constructing an appropriate rubric is the core element to meaningful performance based assessment and there are two types of rubrics commonly used to score learners' performance; namely, holistic and analytical (Mertler, 2001; Moskal, 2000).

Holistic rubrics that assess a learner's overall performance and achievement on a qualitative level provide an overall description of various levels of performance and result in a single score or grade (Goodrich Andrade, 2001; Moskal, 2000). Holistic rubrics can also be developed and applied more rapidly. By contrast, analytic rubrics view performance as being made up of many components and provide separate scores, indicators, and descriptions for each component. The educators can monitor a reflector's performance against each of the well-defined assessment criteria (Mertler, 2001). Then, by collecting the scores calculated separately, the total score related to the performance is obtained (Moskal, 2000). Therefore, it provides more detailed information that may be useful when providing feedback.

Performance-based assessment raises some concerns about subjectivity, reliability, and validity. One of the crucial points of performance-based assessment is to conduct highly reliable measurement and evaluation practices to make accurate decisions about the learners. Analytic rubrics are often preferred with the advantage of dividing the performance process or product into specific sections, ensuring that these sections are scored to meet predetermined criteria. In this case, it is thought that errors caused by the person who measured during the scoring process,

in other words, by the rater, will have less impact. However, determining whether the aspect to be measured exists in the individual based on the opinion of a single rater may also decrease the reliability of the assessment. Accordingly, it is believed that assessments with more than one rater will increase reliability (Abedi *et al.*, 1995). On the other hand, the raters are considered as a significant source of error in the assessments made based on the opinion of the rater (Airasian, 1994; Anadol & Doğan, 2018). At this point, while the increase in the number of raters is crucial for the accuracy of the decisions taken, the higher number of raters is seen as a potential source of error that is thought to be involved in the measurement. Accordingly, various error sources may be encountered such as the individual characteristics of the raters, the number of the raters, the differences of the raters' opinions, and the surrounding variables affecting the rater (Turgut & Baykul, 2010). The measurements are objective to the extent that the raters are given the same score on the same answer, and only in this case the rater reliability is ensured (Shavelson & Webb, 1991; Turgut & Baykul, 2010). In performance-based assessments, before making decisions about individuals, it is necessary to examine the consistency between raters to determine the reliability of the measurements made.

There are many methods and techniques to analyse inter-rater reliability based on Classical Test Theory (CTT), Item Response Theory (IRT) and Generalizability Theory (G Theory) (Baykul, 2015). The variety of theories and techniques causes differentiation of the reliability coefficients obtained, but also provides different information from applications. In this study, the consistency between different numbers of raters was analysed by using Kappa and Krippendorf alpha statistical techniques based on CTT. Within the scope of the G Theory, between the raters of the fully crossed pattern (s x i x r), the G and Phi coefficients that emerged because of the D study were determined and inter-rater reliability analyses were conducted.

In CTT (Lord, 1959; Novick, 1966), observed scores (X) from psychometric instruments are thought to be composed of a true score (T) that represents the subject's score that would be obtained if there was no measurement error, or an error (especially random errors) component (E) that is due to measurement error, such that "Observed Score = True Score + Measurement Error", or in abbreviated symbols, "X = T + E" (Baykul, 2015). Such errors may arise from the individual's performing measurements, the properties measured, the measuring environment, and the measuring technique (Atılgan *et al.*, 2007; Shavelson & Webb, 1991). Since the reliability coefficient for only one type of error is calculated at one time with the CTT, it is necessary to analyse each possible source of error separately. In addition, the inability to calculate the interaction of error sources together seems to be a limitation for the CTT. However, the limitations of techniques developed based on CTT to be used in determining rater reliability revealed the need to examine these techniques.

In this research, the Kappa statistical technique was chosen, because it was not affected by the subcategories included in the analytic rubric and it showed consistency only because of the change in the number of raters. However, analysis was carried out using the Fleiss Kappa statistical technique, since there were more than two independent categories and the need to determine the consistency of two independent and more than two scoring points independently. Thus, the consistency between the assessments of different numbers of raters independent from each other was determined. Krippendorff's statistical technique was preferred due to its advantages such as being used in this study in different number of sample cases, being easily applied to each scale type, and being used in cases where the number of raters is more than two.

The G Theory which was founded in 1940 is a continuation of CTT and Analysis of Variance (ANOVA). The fact that various and many error sources can be determined separately with a single analysis in G Theory increases the importance and usefulness of the theory. In addition, obtaining the Coefficient of Reliability (G) that reveals errors arising from the interaction of various error sources both individually and with each other is the reason why G Theory is

preferred (Brennan, 2001; Shavelson & Webb, 1991). In this research, the variances arising from the items in the analytic rubrics and the raters were determined and possible sources of errors were interpreted. In D studies, consistency analyses are performed in cases where different numbers of raters are included, and suggestions are developed for this situation.

A review of the related literature both nationally (Atılgan, 2005; Bıkmaz Bilgen, 2017; Büyükkıdık, 2012; Güler, 2009; Güler; 2011; Özmen Hızarcıoğlu, 2013) and internationally (Abedi *et al.,* 1995; Goodrich Andrade, 2001, Gwet, 2002; Lane & Sabers, 1989; Marzano, 2002; Oakleaf, 2009) shows that there are many studies on inter-rater reliability. However, there is no related study conducted in the field of IBDP, one of the programmes that are internationally accepted and has standardized assessment and evaluation practices. "Individual Investigation" is a core part of the internal assessment for science subjects. Learners select a real-life issue from Physics, Chemistry, or Biology and investigate it in order to produce a scientific report about it. The main aim of this component is to convert a situation that learners wonder about a scientific issue and solve it by using a scientific method. Within the scope of this aim, learners are expected to become individuals who are aware of the nature of science, have analytical and critical thinking skills, have an ability to apply scientific research methods, and use their scientific knowledge effectively in solving real life issues. However, they produce a scientific report in which they demonstrate these skills under the guidance of the teacher (IBO, 2014a; IBO, 2014b; IBO, 2014c; IBO, 2015). It is clearly seen that individual investigation work and its report are exceptionally good examples of performance based assessment and context of investigation coincides with scientific literacy skills. Therefore, scientific literacy skills of the learner are assessed through an individual investigation process and learner's report. In this specific research, the analytic rubric used in the assessment of learner's performance was prepared by IBO experts (IBO, 2015).

This research is thought to reveal whether the analytic rubrics used internationally is used effectively, to ensure their deficiencies, if any, and to contribute to the IBDP internal assessment process since it will serve as an example for performance studies on the assessment of scientific literacy at national and international levels to be carried out in the future. In addition, due to the COVID-19 pandemic we experienced, IBDP final exams were cancelled. While calculating the graduation scores of learners about science courses, individual investigation reports in this research would be predominantly taken as a basis. At this point, inter-rater reliability has become even more important. For this reason, it is thought that the research would serve as an example for the importance of the number of raters in determining the performance based assessment and the reliability of the decisions taken.

In line with the main aim of the research, this research strives to address the following research questions:

- Is there a statistically significant difference between the scores obtained from two, three, and five raters according to Kappa statistical technique?
- Is there a statistically significant difference between the scores obtained from two, three, and five raters according to Krippendorff alpha statistical technique?
- In the pattern where all sources of variability are fully crossed (s x i x r), does the consistency between different numbers of raters differ significantly in the G and Phi coefficients?
- Are the reliability coefficients obtained from the analysis findings based on the CTT and G theory consistent with each other?

2. METHOD

2.1. Research Model

This research is built on the basis of applying different techniques based on CTT and G Theory in order to analyse the level of inter-rater reliability, examining their restrictions and finding out which of these techniques provides more comprehensive and reliable information. Since the purpose of the research is to reveal the existing situation, it is a descriptive research (Bailey, 1994; Büyüköztürk *et al.*, 2012).

2.2. Study Group

The study group of the research consists of five raters (or teachers) who scored twenty individual investigation reports prepared by the learners within the scope of IBDP internal assessment for Biology subject. The raters had between five and twenty years teaching experience in national education system, however, they had been teaching IBDP Biology for two to ten years and all had a Teaching Certificate.

Evidently, there will always be differences of interpretation of the criteria - and this may vary from person to person and from sample to sample. As an IB requirement, teachers need to meet to discuss the analytic rubric. The participants should agree on common standards at the start of the assessment and be consistent throughout. Teachers of the same science subject should mark two or three individual investigations each. They should then mark their colleagues' learners' individual investigations using the same process they used to mark their own learners' individual investigations. Afterwards, a standardization meeting should be held to determine the level of marking. Internal harmonization of marks is clearly seen as critical to obtaining reliable and valid results at the end of the assessment (IBO, 2018).

Therefore, in the line with IB guidance, in this study 2, 3 and 5 raters were chosen to assess learners' reports who had a similar internal assessment experience year in IBDP curriculum. Inter-rater reliability is applied in situations where different assessors or raters provide subjective judgment on the same target (Viera & Garret, 2005). For this reason, there should be at least two, if possible three raters, as the reliability value obtained determines how much the raters agree on the scoring of a particular target (Burry-Stock *et al.*, 1996). The reason for choosing five raters is to observe whether the increase in the number of raters significantly changes the reliability or not.

2.3. Data Collection Tools

Individual investigation is the core component of the internal assessment of the science subjects in IBDP. Learners choose one of the real-life issues in Physics, Chemistry or Biology and work on it to carry out their investigation and produce a report about it. In this research, the Biology individual investigation reports of learners who graduated from the programme in the same year were used (IBO, 2018).

2.3.1. *Analytic rubric*

In the study, analytic rubric, the basis for assessing individual investigation reports of IBDP Biology subject and developed by IBO experts and also used for the same purpose in schools that implement the programme in all countries, was used to assess the individual investigation reports of IBDP biology subject within the scope of internal assessment (IBO, 2014a).

The internal assessment requirements and analytic rubric are the same for biology, chemistry, and physics. The internal assessment, worth 20% of the final assessment, consists of one scientific investigation. The individual investigation should cover a topic that is commensurate with the level of the course of study. Learner work is internally assessed by the teacher and externally moderated by the IB examiners.

Assessment criteria should be specifically matched to any investigation that has been designed to be used to assess learners. For analytic rubric, several assessment criteria have been identified. There is a level descriptor that describes specific levels of achievement and performance, and a range of marks associated with those levels, for each assessment criterion. Teachers, or raters, are required to judge the learner's work against the level descriptors. Each of the performance levels is described with multiple indicators. There are many cases in which the indicators occur together at a specific level, but not always. In addition, not all indicators are present at all times. As a candidate's performance can fit in different levels, IB assessment models use bands of marks and recommend that teachers and examiners use a best-fit approach to deciding the appropriate mark for a particular criterion. In other words, compensation should be given for work that meets various aspects of a criterion at different levels. For a mark to be awarded, it is not necessary to meet every aspect of a level descriptor. The mark should reflect the achievement balanced against the criterion. The teacher should read each of the level descriptors until they find the one that most accurately describes the level of the work. The learner's work should be read again if it seems to fall between two descriptors and then the descriptor that more accurately describes the work of the learner should be chosen. If two or more marks are available within a level, teachers should award the higher mark if the learner's work displays the qualities described to a great extent. In other words, learners may be close to reaching a higher level. Marks should only be recorded as whole numbers; fractions or decimals are not acceptable. Teacher should not focus on the pass/fail boundary, but rather identify appropriate descriptors for each assessment criterion. Learners should be able to reach the highest-level descriptors if this is appropriate for the assessment. Teachers should not avoid using the extremes when appropriate for the assessment. If a learner achieves a high achievement level for one criterion, it does not mean that he/she will achieve high achievement levels for the other. Similarly, learners who achieve a low level of achievement for one criterion will not necessarily achieve similar levels of achievement for other criteria. The assessment of all the learners should not be assumed to result in a particular mark distribution for the teacher. Learners should be made aware of the assessment criteria. All explanations about how to use analytic rubric, criteria and descriptors should be available in IB Physics, Chemistry and Biology guides (IBO, 2014a; IBO, 2014b; IBO, 2014c). When it comes to the IB moderation, a sample of the marking of internally assessed work is remarked by a moderator to ensure that marking is accurate. During the process, assessors use statistical comparisons and linear regression techniques to determine the degree to which original teacher marks need to be adjusted to align with the set standards.

Scoring rubrics may be designed to contain both general and task specific components. The analytical rubric used in this research is a good example of this situation. The purpose of an individual investigation is to evaluate learners' scientific literacy skills and their scientific knowledge of the chosen topic. This analytic rubric used contains both a general component and a task specific component.

The IBDP analytic rubric uses five criteria with 24 points, in order to assess the final report of an individual investigation, with these raw marks and weightings assigned: personal engagement (up to 2 points/8%), exploration (up to 6 points/25%), analysis (up to 6 points/25%), evaluation (up to 6 points/25%), and communication (up to 4 points/17%). Personal engagement assesses the extent to which the learner has mastered her/his research, how she/he designed and applied it and how she/he presented it in the report. Exploration assesses the extent to which there is clear explanation of the learner's research question and supports with research and theories by reviewing the literature in this direction and completing its work in a safe, environmental, and ethical manner. Analysis assesses the extent to which some criteria such as collecting, analyzing data, and being aware of the impact of the results of the analysis on the research reflect the research situation of the learner. Evaluation assesses the

extent to which the research is supported by relevant theories, defining its strengths and weaknesses, expressing the limitations and errors, interpreting the data obtained, discussing comprehensively, and presenting suggestions based on these data. Communication assesses the extent to which the research is well structured and focuses on the research question and the clear expression of relevant information accordingly.

2.4. Data Analysis

Kappa statistical technique is the first technique applied in this study in order to determine the inter-rater reliability. Although it is often mentioned in the literature about Cohen's Kappa, Fleiss Kappa technique is preferred in cases where there are more than two raters (Cohen, 1960). In this study, Fleiss Kappa technique was used. For the Kappa statistics, SPSS syntax (stats fleiss kappa [v4]. sps)" script was used in SPSS 21.0 software. Then, for the analysis, the reliability of 2, 3 and 5 raters for five different criteria in the analytic rubrics was examined, respectively.

Krippendorff alpha technique was preferred as it can be applied to any scale level. For the Krippendorff alpha technique, "SPSS syntax (kalpha.sps)" was used in SPSS 21.0 software. Then, 2, 3 and 5 raters were calculated for both criteria and total score in the analytic rubrics to observe the consistency.

In studies based on G Theory, each rater in the rater group consisting of two, three, and five raters score each performance report in the research in a way that corresponds to the items in the analytic rubrics. In this study, the raters assessed twenty learner biology reports written for an internal assessment. In this context, the pattern used in the study is a fully crossed pattern and is expressed as (s x i x r). Accordingly, analyses were conducted in order to determine how the variance components and the percentages of these components in the total variance changed with the number of raters. EduG 6.1 software was used for statistical analysis based on G Theory in the analysis. In this context, G and Phi coefficients were determined and D study was included. In cases that occur with the change in the number of raters, the change of G coefficient is observed by conducting D coefficient study.

3. RESULTS

The findings are presented in order in which the subproblems of the research are given and interpreted. The Kappa and Krippendorff alpha statistical techniques were interpreted by calculating the inter-rater reliability values both separately for each criterion and in terms of total scores. When scoring with two, three, and five raters within the scope of the first and second sub-problems of this research, the consistency of the scores obtained was analysed by Kappa and Krippendorff alpha statistical techniques and the findings are summarized in Table 1.

When Table 1 is examined, the negative and positive values of the findings related to Kappa statistics are seen in the scores obtained from different numbers of raters. That the Kappa value (κ) is negative indicates that the agreement between two or more raters is less than expected by chance, A (–1) value for Kappa indicates no observed agreement (i.e., the raters agree on nothing), and (0) (zero) value indicates no agreement. According to Agresti (2013), negative reliability values rarely occur; however, these values were observed in this research. Even though the reliability values (Fleiss kappa coefficients) between the raters are significant, it is worth noting that these values are very low. It can be because of the fact that both low inter-rater agreement and a lack of clearly defined criteria in the rubric lead to low and negative values (Fleiss, 1971).

Table 1. *Kappa and Krippendorff's Alpha Statistical Values Regarding the Scores of Different Number of Raters.*

Number of raters	Criteria	Kappa statistical value (κ)	Krippendorff's alpha value
2	Personal Engagement	0.076*	0.026*
	Exploration	-0.026*	0.059*
	Analysis	0.133*	0.372*
	Evaluation	0.281*	0.571*
	Communication	-0.028*	-0.258*
	Total score	0.228*	0.440*
3	Personal Engagement	-0.006*	-0.073*
	Exploration	-0.049*	-0.039*
	Analysis	-0.078*	0.252*
	Evaluation	0.112*	0.337*
	Communication	0.106*	0.098*
	Total score	0.120*	0.288*
5	Personal Engagement	0.074*	0.066*
	Exploration	-0.014*	0.125*
	Analysis	0.054*	0.303*
	Evaluation	0.158*	0.503*
	Communication	0.108*	0.150*
	Total score	0.163*	0.373*

* $p<0.001$

In the condition that there are two raters, Kappa values change between -0,028 and 0,281. In this case, the lowest level of agreement is in the "communication" criterion (κ = -0.028); the highest level of agreement is estimated in the "evaluation" criterion (κ = 0.281). In the "exploration" criteria learners are expected to establish the scientific context and also they need to put a clear research question, as well as ideas or skills explored in the syllabus. Another criterion where raters scored differently from each other was "personal engagement." In this criterion, the learner is expected to reflect on the subject: why she/he chooses the subject, and how she/he uses the individual characteristics and skills she/he has while exposing the subject. These two criteria differ from one learner to another, as well as from one rater to the other. In this case, it can be thought that the criteria are perceived differently by the raters and create different expectations. In the "evaluation" criterion, the learners are expected to interpret the analysis results, make inferences, and analyse the results together with their previous knowledge. Accordingly, it is observed that learners' research is designed to meet the expectations of the raters of this section, even partially. When Table 1 is analysed, it is seen that, Kappa values are not too high or do not even get negative values. Negative values indicate low inter-rater agreement and raters make different evaluations from each other (Agresti, 2013; Fleiss, 1971). However, with the Kappa technique, no information can be obtained about the sources of errors causing no-agreement between raters. When looking at the overall inter-rater agreement across the overall score, the Kappa value (κ = 0.228) indicates low agreement (Landis & Koch, 1977).

It was determined that the mismatch regarding "personal engagement" and "analysis" criteria increased in the measurement involving three raters. This may be because the relevant items are not correctly understood by the raters, or the raters' expectations for these criteria are different. However, overall inter-rater agreement is lower than the situation where two raters

are present. The biggest difference in the negative direction was in the “analysis” criterion. "Analysis” is one of the criteria that should be prepared comprehensively by supporting various data in scientific studies (IBO, 2015). A criterion in the relevant criteria may differ from one rater to another in some way. When looking at the overall agreement among the three raters, the Kappa value indicates a low agreement (Landis & Koch, 1977). Kappa values appear to decrease as the number of raters increases.

The criteria where the five raters diverged the most were the “exploration” criterion. The “evaluation” criterion was the criteria in which raters agreed, albeit partially. However, when looking at the overall agreement between the five raters, the Kappa value indicates a low level of agreement (Landis & Koch, 1977).

According to the Krippendorff’s alpha values in Table 1, it is seen that there is a relatively high level of agreement between two raters in the “analysis” with (α= 0.372) and also “evaluation” (α= 0.571) criteria. It is much higher than other criteria. For the “personal engagement” with (α= 0.026) criterion, it is seen that the raters scored quite far from each other. The reason for this is the criterion in which these criteria reveal the individual characteristics of the learners and scores whether their studies are designed and expressed well or not (IBO, 2015).

According to the IBO (2018), scientific reports are produced at a particular time by learners. As teachers have been moderated by IBO each year, they are aware of how to use the analytic rubric in a good standard and try to standardise their assessment of learners’ work to ensure reliable results in accordance with IB guidelines. However, there are still error sources such as learners, raters, the development process of the performance task, and the analytic rubric. For example, descriptors for some of the criteria may not be sufficiently expressed in the analytic rubrics or raters struggle to use analytic rubrics though they have used them before. Therefore, it is not possible to determine these potential situations and errors with the Krippendorff alpha statistical technique (Krippendorff, 2004).

When the Krippendorff’s alpha values calculated for the three raters are examined in Table 1, it is seen that the agreement rate of the “analysis” with (α= 0.252) and “evaluation” with (α= 0.337) criteria is higher than the other criteria. All the criteria except the “communication” with (α=0.098) criterion were negatively affected by the increase in the number of raters. Accordingly, it can be stated that the scores obtained from the criteria are not reliable (Krippendorff, 2011). In addition, the divergence of Krippendorff alpha values can be based on the level of objectivity of the criteria.

According to Table 1, when the Krippendorff’s alpha values calculated for the five raters are analysed, the highest agreement can be seen at “evaluation” with (α= 0.503) criterion. In the case of five raters, no negative values were found. It can be thought that the raters do not score differently enough to reach a negative level. In the ranking of the rater reliability of the criteria; as in the two and three raters, a higher level of agreement was seen in “analysis” and “evaluation” criteria than that of the others. When looking at the overall inter-rater agreement, it was found that this ratio could not reach even fair agreement (Krippendorff, 2011).

Regarding the third research question of the study, the analyses were carried out in a fully crossed pattern (s x i x r) and the variance components estimated for the learner. However, in this part, student, and s, refers to the learner, student (s), item (i) (called as criteria) and rater (r) as given in Table 2. When the variance and total variance explanation percentages as a result of the G study in Table 2 are examined, it is seen that the variance component of the main effect of the students corresponds to 9% of the total variance. The variance component of the students gives an estimate of how students' performance studies change from one student to another. The variance component of the students is ($\sigma^2b = 0.227$) and it is expected to be at a high rate as the differentiation of the students’ characteristics affects consistency. Since performance studies

are the studies that students manage the process themselves and produce a product at the end of the research, errors arising from students, or the measured feature may interfere in measurement and evaluation practices (Brennan, 2001).

The variance component ($\sigma^2m = 1.121$) estimated for the main effect of the item has the highest variance value in the total variance with 44% of the total variance and is identified as the most important source of variability among all variance sources. In this case, it is believed that students may not be able to provide the necessary and effective performance report for each item and that the ratings of the criteria differ among raters. However, since each item in the analytic rubrics measures the skills related to performance, this rate is expected to be high (Güler & Taşdelen, 2015). It should be noted, however, that these criteria try to measure skills that are not distant from each other.

Table 2. *The Variance Components and Total Variance Percentages Obtained as a Result of the G Study of the Pattern (s x i x r).*

Variance Source	Square Total	df	Mean of Squares	Variance	Percentage of Variance (%)
s	152.952	19	8.051	0.227	9.0
i	466.656	4	116.638	1.121	44.4
r	29.472	4	7.368	0.037	1.5
si	150.480	76	1.981	0.260	10.3
sr	79.968	76	1.052	0.075	3.0
ir	51.368	16	3.210	0.126	5.0
sir	205.19.2	304	0.675	0.675	26.7
Total	1.135.992	499			100%
G = 0.90 Phi = 0.90					

It is observed that variance from the rater constitutes 1.5% of the total variance. The variance value ($\sigma^2p = 0.037$) calculated for the rater effect was found low. The variance component of the raters provides the opportunity to make an estimate of how the raters give their scores on performance studies. It shows that the raters have a low role in the differentiation of scores. The low percentage of total variance explanation of the variance component of the raters can be interpreted as independent raters make scoring consistent with each other.

(student x item) interaction provides information on whether students' performance reports differ according to the criteria in the analytic rubrics (Shavelson & Webb, 1991). As can be seen in Table 2, the student x item interaction has the highest variance value in total variance. This situation can be interpreted as students' performance reports differ from one criterion to another. It also shows that each criterion measures different skills. A student may qualify for one criterion, but not for another (IBO, 2014a; IBO, 2014b; IBO, 2014c). The criteria are composed of a range of related skills that candidates should be able to demonstrate at various levels of accomplishment. The requirement of each criterion is different from each other in the analytic rubric. The achievement level descriptors for each criterion, which describe the typical ways in which a candidate can be assessed in accordance with the criterion, are used to describe differences in candidate achievement that result in a different mark. The final mark is determined by adding up the maximum levels of achievement for each criterion. Internal consistency measures of reliability are not considered appropriate because each component (assessment tool) may deliberately contain varied forms of task, or sometimes a small number of tasks (IBO, 2018). However, as an item has a/significant effect on reliability with the higher

variance value (44%), increasing the number of items may increase the impact of "student x item" interaction (Brennan, 2001).

"item x rater" is the variance of the common interaction ($\sigma^2mp = 0.126$), which creates a 5.0% effect in the total variance. This indicates that there is no significant level of difference in the scoring consistency between the raters. As can be seen in Table 2, the (student x item x rater) variance component indicates 26.7% of the total variance. This common effect is the second-high variance in total variances. That the G and Phi coefficients are 0.90 means that the scoring reliability is high; in other words, the inter-rater agreement is high (Atılgan, 2005; Brennan 2001; Shavelson & Webb, 1991).

D study investigates the impact of variability among the scores from different numbers of raters. D study conducted within the scope of G Theory analysis allows researchers to calculate two different reliability coefficients that are effective in making both relative decisions based on students' performances and absolute decisions regarding students' performances (Shavelson & Webb, 1991). Researchers benefit from G coefficient in making relative decisions, and from Phi coefficient in making absolute decisions. The study findings carried out to examine the effect of the D study and the numbers of raters on the G and Phi coefficients are given in Table 3.

Table 3. *G and Phi Coefficients of Pattern (S X I X R) Estimated by D Study.*

Measurement pattern	Number of Items	Number of raters				
		nr=1	nr=2	nr=3	nr=4	nr=5
s x i x r	5	G=0.65 Phi=0.64	G=0.79 Phi=0.78	G=0.85 Phi=0.84	G=0.88 Phi=0.88	G=0.92 Phi=0.91

As can be seen in Table 3, the increase in the number of raters causes an increase in G and Phi coefficients. It can be concluded that the G and Phi coefficients are estimated higher, and the number of raters has a significant impact on scoring reliability in cases created using a different number of raters and the same pattern. In addition, it is clearly seen in Table 3 that the Phi coefficient, which is important for this study, was positively affected by the rater increase. In assessing performance studies, although it seems ideal in theory, it may not always be possible to reach five raters in practice. In this case, making assessments with three or four raters, if possible, can lead to more reliable results and accurate decisions about students.

According to the results of analysis based on CTT and G Theory, the fourth research problem of the study was interpreted within the scope of the findings obtained. The reliability coefficients obtained from the analysis were not consistent with each other. Kappa and Krippendorff alpha statistical techniques used for the analysis based on CTT showed a low level of agreement between raters. In both techniques, consistency between raters showed negative values on many criteria in the analytic rubric. According to the analysis based on CTT, it is not possible to determine the ideal number of raters because the values vary from two raters to five raters. However, the results which were obtained by two and five raters were close to each other. Moreover, it is not possible to determine the sources of errors causing this incompatibility with these techniques. Analyses based on G theory provide the opportunity to interpret many variables both separately and also together. In the research, it is advantageous for the researchers to observe the variances arising from the learners, the items and the raters. In this research, in line with the IBO guide (2018), learners need to have 10 hours to complete their scientific reports with the teacher guidance. They also need to produce a project plan before they carry out their investigations and experiments. Meanwhile, as an IB requirement, teachers support the learner in the line of analytic rubric. As it is a standardised process, in all schools

where this programme is implemented and their educators must follow the same stages, all environmental conditions, limitations and error sources are minimised.

4. DISCUSSION and CONCLUSION

Within the scope of the research, analyses were made by using techniques based on CTT and G Theory and the results obtained were compared in determining the inter-rater reliability levels. Values obtained from two, three, and five raters based on the Kappa statistical technique indicate a low level of agreement when examined in each criterion and in the total score. According to the results of the analysis, the highest agreement between the raters was determined as the situations where two raters were included, and the lowest level of agreement was determined as the situations where three raters were included. Negative values are seen in some criteria, in other words, inconsistency between raters indicate that raters do not make consistent assessments when scoring. According to the analysis findings based on the Kappa statistical technique, the increase in the number of raters has decreased the Kappa value relatively (Nying, 2004). This situation is thought to be an indication that Kappa statistics are affected by the increase in the number of raters. Accordingly, it can be said that it is sufficient to include two raters. These findings coincide with the finding that the increase in the number of raters in the measurement of performance of Abedi *et al.,* (1995) studies decrease reliability by increasing the level of variability in scores. However, it can be stated that the Kappa statistical technique is insufficient in determining the ideal number of raters in determining the performance-based assessment.

Based on the Krippendorff alpha statistics technique, the consistency between raters indicates a low level of agreement when the analyses obtained from different numbers of raters are analysed in each criterion and total score. The highest values indicating the compatibility between raters from the analysis made with Krippendorff alpha technique were calculated in cases related to the situation of two raters, as in the Kappa statistics. This finding coincides with the findings of Bıkmaz Bilgen and Doğan (2017), where the highest agreement was found when there were two raters. However, in both techniques, the analyses in the case of three raters indicate that the inter-rater agreement is at the lowest level. In the Krippendorff alpha technique, as in the Kappa technique, as the number of raters increased, the alpha value changed; however, this change was not as significant as in the Kappa statistic and displayed a relatively more stable structure.

The Kappa and Krippendorff alpha values, the basics of which were developed based on the CTT, calculated the levels of inter-rater reliability exceptionally low. Although there are sources of students, item, and scoring variability in this study, it can be said that these techniques based on CTT are insufficient in reaching the variable that causes negative values and incompatibility. According to the findings of the study, it is seen that Kappa and Krippendorff alpha techniques are insufficient in deciding the ideal number of raters and error sources in assessing the performance reports reflected by the learner characteristics.

Based on the G studies, the effect of the variance originating from the students' items and raters in the measurement process was calculated in the total variance. The study findings related to sources of variance from G study showed that the main source of variance across all criteria was items, while raters represented a relatively small source of variance. It shows that raters were not a significant source of error. Moreover, it means that items measured different kinds of skills and raters could not create a significant impact on the assessment process. Additionally, the common effect resulting from students and items has a high value. This indicates that students show different competencies in different items. This result also shows that each of the items measures different skills, and it is an expected result. It also means the analytic rubric is a reliable measurement tool. According to the D study, it is seen that increasing the number of

raters increases the reliability positively. Especially in cases where five raters are not reached in practice, it can be said that assessing with three or four raters increases the scoring reliability, so that accurate decisions can be made. In the analysis based on G theory, the inter-rater reliability coefficient was higher than that of the Kappa and Krippendorff alpha techniques. Moreover, it was concluded that increasing the number of raters with the D study would increase reliability (Büyükkıdık, 2012; Deliceoğlu, 2009; Güler, 2009; Öztürk, 2011). These results provide more comprehensive data against the limitations of the CTT. However, Kamış and Doğan (2017) state that the G and Phi coefficients, which are obtained in real situations where the raters are not randomly selected from the population and estimated as a result of different decision studies, differ even though they take values close to each other. The G and Phi coefficients obtained as a result of the D studies require that the relevant sources of variability (raters, items, etc.) be randomly selected from the population in the new application to be carried out. Random selection of raters from the population is possible in large-scale measurement applications, but it is practically not possible in-class measurement applications. It is not clear whether the G and Phi coefficients obtained as a result of the D study accurately predict the real situation. The analysis results in this research show that the actual values obtained and predicted in the D studies are similar but differentiated. It is recommended raters should be selected randomly from the population and determine which coefficient will be more accurate to use afterwards.

In addition to the quantitative studies, opinions were received from the raters. These views are primarily the performance reports developed by the teacher and the learner together, while the teachers try to improve themselves to provide reliable feedback to the learner while trying to apply the scientific research methods and steps in the most correct way. However, they think that it is important to elaborate on the descriptions in the criteria, in other words, to make the expressions used to measure the targeted feature clearer. In this context, it would be an appropriate decision to expand the explanations of the criteria in an analytic rubric.

Based on the findings of the research, it can be stated that it is important to use performance based assessment and evaluation approaches in order to observe the learners' characteristics in all aspects. It is seen that the individual investigation steps carried out in the field of biology science within the scope of IBDP and the measurement and evaluation practices of these studies may be examples of the studies to be carried out at the national level. Furthermore, in this study, the rater group had a similar teaching experience and background (e.g., rating experience) in IBDP. In future research, raters with different teaching and/or rating experiences and backgrounds in IBDP should be preferred. Researchers or educators, therefore, compare the relationship between rater experience and the assessment of scientific reports.

Acknowledgments

This paper was produced from part of the first author's master's thesis prepared under the supervision of the second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Ankara University/Social Sciences Institute, 17/07/2019-09-259.

Authorship Contribution Statement

Sinem Arslan Mancar: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **H. Deniz Gulleroglu**: Methodology, Supervision, and Validation.

Orcid

Sinem Arslan Mancar https://orcid.org/0000-0002-2031-2189
H. Deniz Gulleroglu https://orcid.org/0000-0001-6995-8223

REFERENCES

Abedi, J., Baker, E.L., & Herl, H. (1995). *Comparing reliability indices obtained by different approaches for performance assessments* (CSE Report 401). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). https://cresst.org/wp-content/uploads/TECH401.pdf

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). John Wiley & Sons.

Airasian, P.W. (1994). *Classroom assessment* (2nd ed.). McGraw-Hill.

Aktaş, M. & Alıcı, D. (2017). Kontrol listesi, analitik rubrik ve dereceleme ölçeklerinde puanlayıcı güvenirliğinin genellenebilirlik kuramına göre incelenmesi [Examination of scoring reliability according to generalizability theory in checklist, analytic rubric, and rating scales]. *International Journal of Eurasia Social Sciences*, *8*(29), 991-1010.

Anadol, H.Ö., & Doğan, C.D. (2018). Dereceli puanlama anahtarlarının güvenirliğinin farklı deneyim yıllarına sahip puanlayıcıların kullanıldığı durumlarda incelenmesi [The examination of realiability of scoring rubrics regarding raters with different experience years]. *İlköğretim Onlin*e, 1066-1076. https://doi.org/10.17051/ilkonline.2018.419355

Ananiadou, K., & Claro, M. (2009), 21st century skills and competences for new millennium learners in OECD countries. *OECD Education Working Papers*, *41*. OECD Publishing, Paris,https://doi.org/10.1787/218525261154

Atılgan, H.E., (2005). Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama [Generalizability theory and a sample application for inter-rater reliability]. *Educational Sciences and Practice, 4*(7), 95-108. http://ebuline.com/pdfs/7Sayi/7_6.pdf

Atılgan, H., Kan, A., & Doğan, N. (2007). *Eğitimde ölçme ve değerlendirme* [Assessment and evaluation in an education] (2nd ed.). Anı Yayıncılık.

Bailey, D.K. (1994). *Methods of social research* (4th ed.). Free-Press.

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: klasik test teorisi ve uygulaması* [Measurement in education and psychology: classical test theory and practice] (3rd ed.). Pegem Yayıncılık.

Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması [The comparison of ınterrater reliability estimating techniques]. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(1), 63-78. https://doi.org/10.21031/epod.294847

Brennan, R.L. (2001). *Generalizability theory*. Springer-Verlag.

Burry-Stock, J.A., Shaw, D.G., Laurie, C., & Chissom, B.S. (1996). Rater-agreement indexes for performance assessment. *Educational and Psychological Measurement*, *56*(2), 251-262. https://doi.org/10.1177/0013164496056002006

Büyükkıdık, S. (2012). *Problem çözme becerisinin değerlendirilmesinde puanlayıcılar arası güvenirliğin klasik test kuramı ve genellenebilirlik kuramına göre karşılaştırılması.* [Comparison of interrater reliability based on the classical test theory and generalizability theory in problem solving skills assessment] [Master's Thesis, Hacettepe University]. Hacettepe University Libraries.

Büyüköztürk, Ş., Kılıç Çakmak E., Akgün Ö.E., Karadeniz Ş., & Demirel F. (2012). *Bilimsel araştırma yöntemleri* [Scientific research methods] (11th ed.). Pegem Yayıncılık.

Bybee R.W. (1997). Towards an understanding of scientific literacy. In: W. Gräber & C. Bolte. (Eds.). Scientific literacy. An international symposium (p. 37-68). Institut für die Pädagogikder Naturwissenschaften (IPN): Kiel, Germany.

Callison, D. (2000). Rubrics. *School Library Media Activities Monthly*, *17*(2), 34-6,42.

Cambridge International Examinations (2015). *Cambridge IGCSE global perspectives 0457. Syllabus for examination in 2018, 2019 and 2020*. https://www.cambridgeinternational.org/Images/252230-2018-2020-syllabus.pdf

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Collins, R. (2014). Skills for the 21st Century: teaching higher-order thinking. *Curriculum & Leadership Journal, 12*(14). http://www.curriculum.edu.au/leader/teaching_higher_order_thinking,37431.html?issueID=12910

Coombe, C.A., Davidson, P., O'Sullivan, B., & Stoynoff, S. (Eds.). (2012). *The Cambridge guide to second language assessment*. Cambridge University Press.

Deliceoğlu, G. (2009). *Futbol yetilerine ilişkin dereceleme ölçeğinin genellenebilirlik ve klasik test kuramına dayalı güvenirliklerinin karşılaştırılması* [The comparison of the reliabilities of the soccer abilites' rating scale based on the classical test theory and generalizability]. [Doctoral dissertation, Ankara University, Ankara]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Dietel, R.J., Herman, J.L., & Knuth, R.A. (1991). What does research say about assessment? NCREL, Oak Brook. http://www.ncrel.org/sdrs/areas/stw_esys/4assess.htm

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378- 382. https://doi.org/10.1037/h0031619

Goodrich Andrade, H. (2001) The effects of instructional rubrics on learning to write. *Current issues in Education, 4*. http://cie.asu.edu/ojs/index.php/cieatasu/article/view/1630

Güler, N. (2009). Genellenebilirlik kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması [Generalizability theory and comparison of the results of G and D studies computed by SPSS and GENOVA packet programs]. *Eğitim ve Bilim*, *34*(154). http://eb.ted.org.tr/index.php/EB/article/view/551/45

Güler, N. (2011). Rasgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramına göre güvenirliğin karşılaştırılması [The comparison of reliability according to generalizability theory and classical test theory on random data]. *Eğitim ve Bilim. 36*(162), 225-234. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/993

Güler, N., & Taşdelen, G. (2015). Açık uçlu maddelerde farklı yaklaşımlarla elde edilen puanlayıcılar arası güvenirliğin değerlendirilmesi [The evaluation of rater reliability of open-ended items obtained from different approaches] *Journal of Measurement and Evaluation in Education and Psychology*, *6*(1). 12-24. https://doi.org/10.21031/epod.63041

Gwet, K. (2002), Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Series: Statistical Methods for Inter-Rater Reliability Assessment*, *1*(1).1-5. https://www.agreestat.com/papers/kappa_statistic_is_not_satisfactory.pdf

Holbrook, J., & Rannikmae, M. (2009). The meaning of scientific literacy. *International Journal of Environmental & Science Education*, *4*(3), 275-288. https://files.eric.ed.gov/fulltext/EJ884397.pdf

Hurd, P. D. (1998) Scientific literacy: new minds for a changing world. *Science Education*, *82*, 407-416.

International Baccalaureate Organization (IBO). (2014a). *International Baccalaureate Diploma Programme Biology Guide First Assessment 2016*. https://internationalbaccalaureate.force.com/ibportal/IBPortalLogin?lang=en_US

International Baccalaureate Organization (IBO). (2014b). *International Baccalaureate Diploma Programme Chemistry Guide First Assessment 2016*. https://www.ibchem.com/root_pdf/Chemistry_guide_2016.pdf

International Baccalaureate Organization (IBO). (2014c). *International Baccalaureate Diploma Programme Physics Guide First Assessment 2016*. https://ibphysics.org/wp-content/uploads/2016/01/ib-physics-syllabus.pdf

International Baccalaureate Organization (IBO). (2015). *International Baccalaureate Diploma Programme: From principles into practice*. International Baccalaureate Organization.

International Baccalaureate Organization (IBO). (2018). *International Baccalaureate Organization (IBO). (2018). The IB Diploma Programme Statistical Bulletin, May 2018 Examination Session*. https://www.ibo.org/contentassets/bc850970f4e54b87828f83c7976a4db6/dp-statistical-bulletin-may-2018-en.pdf

International Baccalaureate Organization (IBO). (2018). *Assessment principles and practices-Quality assessments in a digital age*. https://www.ibo.org/contentassets/1cdf850e366447e99b5a862aab622883/assessment-principles-and-practices-2018-en.pdf

Kamış, Ö., & Doğan, C. (2017). *Genellenebilirlik kuramında gerçekleştirilen karar çalışmaları ne kadar kararlı?* [How consistent are decision studies in G theory?]. *Journal of Education and Learning, 7*(4). https://dergipark.org.tr/en/download/article-file/336342

Klucevsek, K. (2017). The intersection of information and science literacy. *Communications in Information Literacy, 11*(2), 354-365. https://files.eric.ed.gov/fulltext/EJ1166457.pdf

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Quality and Quantity*, *38*(6), 787-800. https://doi.org/10.1007/s11135-004-8107-7

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* Sage.

Krippendorff, K. (2011). Computing Krippendorff‟s alpha reliability. http://repository.upenn.edu/asc_papers/43

Kulieke, M., Bakker, J., Collins, C., Fennimore, T., Fine, C., Herman, J., Jones, B.F., Raack, L., & Tinzmann, M.B. (1990). Why should assessment be based on a vision of learning? [online document] NCREL, Oak Brook: IL. Available online: http://www.ncrel.org/sdrs/areas/rpl_esys/assess.htm

Kutlu, Ö., Doğan, D.C., & Karakaya, İ. (2008). *Performansa ve portfolyoya dayalı durum belirleme* [Assessment and evaluation determination based on performance and portfolio] (5th ed.). Pegem Yayıncılık.

Landis, J,R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1) 159-174. https://doi.org/10.2307/2529310

Lane, S., & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays, *Applied Measurement in Education*, *2*(3). 195-205. https://doi.org/10.1207/s15324818ame0203_1

Law, B., & Eckes, M. (1995). *Assessment and ESL*. Peguis publishers.

Lord F.M. (1959). Statistical inferences about true scores. *Psychometrika, 24*(1), 1–17. https://doi.org/10.1007/BF02289759 .

Maienschein, J. (1998). Scientific literacy. *Science*, *281*(5379), 917. https://www.proquest.com/openview/568e8a30ee2b1c68d787bbcb39e3f94e/1?pq-origsite=gscholar&cbl=1256

Marzano, R. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education, 15*(3). 249-268. https://doi.org/10.1207/S15324818AME1503_2

Marzano, R.J., & Heflebower, T. (2012). *Teaching & assessing 21st century skills*. Marzano Research Laboratory.

Mertler, C.A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research and Evaluation*, *7*(25), 1-8. https://doi.org/10.7275/gcy8-0w24

Millî Eğitim Bakanlığı (MEB) (2016). *PISA 2015 Ulusal Raporu [PISA 2015: National Report for Turkey*]. Millî Eğitim Bakanlığı, Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı, Ankara. https://odsgm.meb.gov.tr/test/analizler/docs/PISA/PISA2015_Ulusal_Rapor.pdf

Moskal, B.M. (2000) Scoring rubrics: What, When, How? *Practical Assessment Research and Evaluation*, *7*(3), 1-11. https://doi.org/10.7275/a5vq-7q66

National Research Council. (2012). *Education for life and work: developing transferable knowledge and skills in the 21st century*. The National Academies Press. https://doi.org/10.17226/13398

Nbina, J., & Obomanu, B. (2010). The meaning of scientific literacy: A model of relevance in science education. *Academic Leadership: The Online Journal, 8*(4). https://scholars.fhsu.edu/alj/

Novick M.R. (1966) The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1-18. https://doi.org/10.1016/0022-2496(66)90002-2

Nying, E. (2004). *A comparative study of interrater reliability coefficients obtained from different statistical procedures using monte carlo simulation techniques* [Doctoral dissertation, Western Michigan University]. https://scholarworks.wmich.edu/dissertations/1267

Oakleaf, M. (2009). The information literacy instruction assessment cycle: a guide for increasing student learning and improving librarian instructional skills. *Journal of Documentation*, *65*(4). 539-560. https://doi.org/10.1108/00220410910970249

Organisation for Economic Cooperation and Development (OECD). (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

Özmen Hızarcıoğlu, B. (2013). *Problem çözme sürecinde dereceli puanlama anahtarı (Rubrik) kullanımında puanlayıcı uyumunun incelenmesi* [Examining scorer's coherence of using rubric in the problem solving process] [Master's dissertation, Abant Izzet Baysal University]. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=-9VIu1xAI6tVn8H1Pmf2Mg&no=XE36zEJKy4iJQQ-bARoPnA

Öztürk, M.E. (2011). *Voleybol becerileri gözlem formu ile elde edilen puanların, genellenebilirlik ve klasik test kuramına göre karşılaştırılması* [The comparison of points of the volleyball abilities observation form (VAOF) according to the generalizability theory and the classical test theory] [Unpublished doctoral dissertation, Hacettepe University]. National Thesis Centre. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=K9erNYiV2Ks_xzov1XrfSQ&no=5OJsxJV1JE2E3hGJDkB8lQ

Partnership for 21st Century Learning. (2007). *Framework for 21st century learning.* https://files.eric.ed.gov/fulltext/ED519462.pdf

Reeves, T.C. (2000). Alternative assessment approaches for online learning environments in higher education. *Educational Computing Research, 3*(1), 101-111.

Rychen, D.S., & Salganik, L.H. (Eds.). (2003). *Key competencies for a successful life and a well functioning society*. Cambridge.

Schleicher, A. (2015), *Schools for 21st-Century Learners: Strong Leaders, Confident Teachers, Innovative Approaches*, International Summit on the Teaching Profession, OECD Publishing. https://doi.org/10.1787/9789264231191-en

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: a primer*. Sage.

Simonson, M., Smaldino, S, Albright, M., & Zvacek, S. (2000). *Assessment for distance education* (ch 11). T*eaching and learning at a distance: foundations of distance education*. Prentice-Hall.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment frameworks*. http://timssandpirls.bc.edu/timss2019/frameworks/

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. John Wiley & Sons

Turgut, H. (2007). Scientific literacy for all. *Ankara University Journal of Faculty of Educational Sciences (JFES), 40* (2), 233-256. https://doi.org/10.1501/Egifak_0000000176

Turgut, M.F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme* [*Assessment and evaluation in an education*]. Pegem Yayınları.

Uçak, S., & Erdem, H.H. (2020). Eğitimde yeni bir yön arayışı bağlamında 21. Yüzyıl becerileri ve eğitim felsefesi [On the skills of 21st century and philosophy of education in terms of searching a new aspect in education]**.** *Uşak Üniversitesi Eğitim Araştırmaları Dergisi, 6*(1), 76-93. https://doi.org/10.29065/usakead.690205

Viere, A.J., & Garrett, J.M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine, 37*(5), 360-362.

Zurkowski, P.G. (1974). *The Information Service Environment Relationships and Priorities. Related Paper No. 5.* National Commission on Libraries and Information Science, Washington, D.C. National Program for Library and Information Services. https://files.eric.ed.gov/fulltext/ED100391.pdf