Online Science Education Journal *http://dergipark.gov.tr/ofed*

2022; 7(2): 72 - 84.

Performance Scale Development Study for Science-Chemical Laboratory Applications

Özge Göktürk, *Mersin University, Faculty of Education, Department of Mathematics and Science Education, Department of Science Education, Mersin, Turkey*
Nezaket Bilge Uzun, *Mersin University, Faculty of Education, Department of Educational Sciences, Department of Measurement and Evaluation in Education, Mersin, Turkey*
Mehtap Aktaş*, *Kafkas University, Faculty of Education, Department of Educational Sciences, Department of Measurement and Evaluation in Education, Kars, Turkey*

*Corresponding Author: mhtpaktas@gmail.com

To cite this article: Göktürk, Ö., Uzun, N. B., & Aktaş, M. (2022). Performance scale development study for science-chemical laboratory applications. *Online Science Education Journal, 7(2),* 72-84.

Article Info

Article History

Received:
26 Nov 2022

Accepted:
15 Dec 2022

Keywords

Rubric
Analytical Rubric
Science Laboratory Applications

Abstract

This study aimed to develop an analytical rubric that measures performance levels for Science/Chemistry Laboratory Applications and to examine its reliability with generalizability theory. The study group consists of 18 grade 5-6 secondary school students in formal education, taking science or chemistry laboratory courses in a science and art center in the 2020-2021 academic year. In addition, the laboratory performances of students were scored simultaneously by three expert raters, using an analytical rubric. In addition, during the development process of the rubric, the opinions of eight experts were obtained when determining and arranging performance indicators and performance levels. It was determined that the scale developed with the findings obtained from the performance scale development study yielded valid, reliable, and generalizable results in determining the performance of students who attended science/chemistry laboratory practices. In this context, it is thought that the use of this scale, which was developed to evaluate the students' performance in secondary school science/chemistry laboratory practice courses, will provide valid and reliable measurement results and will make the evaluation process more objective.

INTRODUCTION

Science tries to describe and interpret the events taking place in nature to find the facts of nature. In this respect, science can be interpreted as human beings' attempt to understand themselves through nature (Collette, 1989). In the self-recognition process, people tend to do research to describe and interpret events. Novak (1964) defined research as an effort to find logical explanations for events that individuals are curious about. Therefore, doing research is an effort to eliminate curiosity. Laboratories are the environments that provide students with the opportunity to practice this effort by systematizing it. For this reason, laboratory applications have an important place in science education (Lunetta, 1998; Saunders, 1992).

Laboratory practices are practical learning environments where the concepts desired to be learned are transferred to the learners through first-hand or demonstration methods (Tezcan & Aslan, 2007). Since the main purpose of laboratory applications is the realization of meaningful

learning, the active participation of students in the process, their taking responsibility, and the realization of learning by doing (Aksoy & Doymuş, 2011; Nakiboğlu & Meriç, 2000), science educators suggest that frequent laboratory activities will have many benefits on learning (Hofstein & Lunetta, 1982).

The techniques used in science laboratory applications have an important place in learning the target information (Leach, 1998). As a classical approach in science laboratory practices, there are traditional laboratory practices in the "recipe type" where verification-type experiments, that is, high cognitive levels, do not need to be employed (Jackson, 2004). Kaptan (1999) stated that traditional laboratory practices are used to prove the information found in books. However, it was also stated that it did not contribute to the structuring of scientific knowledge in students (Renner, 1986; Aktamış, 2007). On the contrary, it is stated that in applications where alternative laboratory approaches are used, students construct their knowledge by evaluating and constructing the knowledge they have learned through experiments and developing scientific thinking and critical thinking skills (Wyatt, 2005; Rehorek, 2004; Jackson, 2004; Lunetta & Tamir, 1978; Ergin, Şahin-Pekmez, Öngel-Erdal, 2005). Therefore, for permanent and effective learning to take place, using alternative approaches by creating environments where student-focused activities can be designed and implemented becomes of great significance (Lapadat, 2000; Costa, 1985; Birinci, Sezen, & Tekbıyık, 2010). Therefore, effective science education necessitates laboratory practices and environment prepared according to the constructivist approach, in which students can learn the desired learning outcomes through practice and will be responsible for their own learning related to their daily life. In the constructivist learning approach, which is the basis of alternative learning approaches in science education, students are guided to find solutions to problems by interacting with their environment (İlhan, 2013; Geraldo, Jofili, & Watts, 1999). According to this approach, students should be at the center of the learning process, while teachers should guide students on how to construct knowledge (Liang & Gabel, 2005).

As a result of the literature research on laboratory applications in science education, it is seen that various studies have been carried out on the subject. These include evaluation of science laboratory applications (Uluçınar, Cansaran & Karaca, 2004; Ayas, Karamustafaoğlu, Sevim & Karamustafaoğlu, 2002), opinions about the usability of teaching model in science laboratory (Bozdoğan & Altunçekiç, 2007), self-efficacy in laboratory utliziation (Kılıç Mocan, Keleş, & Uzun, 2015; Kaya, Böyük, 2011; Boyuk, Demir, Erol, 2010; Akdemir, 2006; Yurdatapan, 2013), evaluation of science laboratory use (Güneş, Dilek, Topal, & Nesrin, 2013), opinions on the use of science laboratories (Demir, Böyük, Koç, 2011; Kocakülah, Savaş, 2011; Kılıç, Aydın, 2018), difficulties encountered in the chemistry laboratory (Aydoğdu, 1999), competencies in using science laboratory materials (Çoştu, Ayas, Çalık, Ünal, Karataş, 2005; Korkmaz, 2000), views on science laboratory applications (Uluçınar, Doğan, Kaya,2008; Karamustafaoğlu, 2012; Uzal, Erdem, Önen, Gürdal, 2010), attitude towards science experiments (Yıldız, Akpınar, Aydoğdu, & Ergin, 2006; Alkan, Erdem, 2013; Karatay, Doğan, Şahin, 2014; Taşlıdere, Korur, 2012), success in science experiment applications (Alkan, Erdem, 2013; Tezcan & Bilgin, 2004), the effect of science laboratory on academic achievement (Ayvacı & Durmuş, 2016), and the use of V Diagram in science laboratory (Meriç, 2003; Nakiboğlu, Meriç, 2000; Nakiboğlu, Benlikaya, & Karakoç, 2001).

It is seen that there is a tendency towards the use of the traditional evaluation approach in studies conducted for the evaluation of science or laboratory studies, but the adoption of alternative measurement methods is seen as an important method or tool in terms of process evaluation (MEB, 2005). It is argued that traditional assessment measures low-level cognitive knowledge,

whereas performance assessment, which is one of the alternative assessment types, aims to measure high-level cognitive knowledge and competencies (Aydın & Karaçam, 2015).

Performance can be expressed as a service, situation, or idea that is put forward for the realization of a task that is expected or desired to be performed in line with the determined criteria (Pugh, 1991). When examined on the basis of education, performance can also be defined as psychomotor skills such as playing a musical instrument, doing sports, and using a microscope (Turgut & Baykul, 2010; Helvacı 2002). When measuring and evaluating the performance, it is expected that the relevant behavior is done, not explained. In behavioral measurements, in order to reveal the behavior, the measurement is made by observing all the performance steps or the product revealed as a result of the behavior (İşman, 2001). In performance measurement and evaluation processes, checklists revealing whether performance exists, rating scales that reveal the degree as well as the existence of the behavior (Dalkıran, 2006), and holistic and analytical rubrics that allow more objective measurement-evaluation can be used. Rubric is one of the most widely used measurement tools in performance evaluation. It is seen as an important tool in minimizing the biases that may occur during scoring and obtaining more realistic results regarding performance (Parlak & Doğan, 2014). According to Popham (1997), rubric consists of three parts: evaluation criteria, criterion definitions, and scoring strategy. Evaluation criteria are used to distinguish between acceptable and unacceptable answers, criteria definitions are used to identify qualitative differences in students' answers, and a scoring strategy is used to determine the path followed for scoring (analytical or holistic rubrics). Since holistic rubrics focus on the whole process, they are used to make a general judgment about the quality of the performance and are more convenient for evaluating the results (Jonsson & Svingby, 2007). Analytical rubrics, on the other hand, have restrictive performance characteristics when compared to holistic rubrics and can make process evaluation in more detail (Sezer, 2005). Each performance criterion or skill that an individual is expected to demonstrate is evaluated independently within the framework of defined criteria (Çepni, 2011).

Laboratory practices are also activities in which students are active and have to demonstrate performance, as they include stages such as planning, observation, data collection, conclusion, and evaluation. Therefore, it is necessary to evaluate the performance of students. One of the biggest shortcomings when evaluating the performances exhibited in these applications is the use of appropriate measurement tools, and the second one is not using a reliable and valid measurement tool. For example, evaluating laboratory applications through paper and pencil exams prevents students' involvement in scientific research stages. This is because there is a big difference between explaining how to do an experiment in writting and applying it. Evaluation of laboratory applications while practicing is considered important in terms of allowing students to analyze their results (Hilosky, Sutman, Schmuckler, 1998; Goh, Toh, & Chia, 1989). Silberman, Day, Jeffers, Klanderman, Phillips, and Zipp (1987) stated that in order to evaluate the performance of laboratory applications, performance measurement can ensure that the laboratory practitioner is successful in reaching the goal and that the student can achieve permanent learning by improving their higher-order skills in practice. They determined that using a tool for evaluating laboratory practices makes students more willing to actively participate in practices.

Suits (2004) emphasized the importance of performance assessment tools in the assessment of high-level research skills as they provide useful feedback on the quality of laboratory practice exams, and developed an assessment rubric consisting of six components to evaluate laboratory practice. When Panadero and Jonnson's (2013) studies on the use of rubrics are examined, it is

observed that using rubrics eliminates evaluation bias, gives more accurate feedback, reduces assessment anxiety in learners, and helps improved performance.

When the international and national literature on laboratory practices and evaluation is examined, it is seen that various methods are used to make laboratory practices more educative and effective (Arnold, 2003; Exstrom and Mosher 2000; Harle, Leber, Hess, Yoder, 2003; Selco, Roberts, Wacks, 2003; Criswell, 2006) but few of them have been observed to develop and use a performance scale for evaluation (Silberman et al., 1987; Hilosky et al., 1998; Suits, 2004; Arı, 2008). However, it has been determined that these scales are limited to certain applications. It was found that academic achievement tests or V diagrams are often used as an evaluation method (Meriç, 2003; Nakiboğlu, Meriç, 2000; Nakiboğlu, Benlikaya, & Karakoç, 2001), but there was no study on developing a valid and reliable performance scale for measuring laboratory practices in science education that incorporates the whole laboratory practice process into the assessment. Researchers suggest that performance-based assessment and evaluation methods should be used in order to fully evaluate the process (Darling-Hammond, 1994; Shepard, 2008).

Therefore, in this study, which addresses laboratory applications as performance evaluation processes, the development of a "rubric" for laboratory applications may close an important gap in the field. While developing measurement tools used in performance evaluation, such as rubrics, reliability between raters is generally taken as a basis (Atılgan, 2005; Atmaz, 2009; Deliceoğlu, 2009; Güler & Taşdelen Teker, 2015). In this study, the Generalizability Theory (GT), which gives information about both random and systematic sources of error and provides comprehensive reliability analyses in performance evaluation; was employed.

Unlike the classical test theory, GT provides a single reliability value by examining the effects of multiple error sources at the same time. The generalizability theory focuses not on the observed score or a specific measurement result, but on how the measurement results can be generalized to a much larger universe than a specific sample (Güler, 2009). In generalizability theory, it is possible to reach a single reliability value by considering the interaction of many error sources, and error sources within the scope of the study (Brennan, 2001). According to Shavelson and Webb (1991), the GT has four important features. 1) It deals with multiple sources of variance with a single analysis. 2) It determines the size of each variance source 3) It provides the calculation of two different reliability coefficients related to the relative decisions of the individuals (G coefficient) and the absolute decisions about the performance of the individuals (Phi coefficient) 4) It enables decision studies that can offer suggestions for arranging measurements in which the measurement errors are minimized.

GT provides comprehensive analyses by simultaneously evaluating the error from many sources of variability based on the within-group correlation coefficient, which enables the evaluation of reliability in behavioral measurements, the design, and research of reliable observations. GT means that the reliability of an observation depends on the universe from which conclusions are drawn. It deals with how sources of variability reflect the universe. In generalizability theory, the universe is a construct that the researcher considers, but it also provides findings on how well the observed scores represent the universe score. Interpreting how accurately the universe score is estimated from the observed score within the framework of these findings can also be considered as construct validity studies. For this purpose, GT-based analyses, which eliminate the difference between reliability and validity and focus on the generalizability of the measurement tool, were conducted in this study.

There are two basic stages in GT, namely Generalizability study (G-study) and the Decision study (D-study) (Goodwin, 2001). In the G-study, all sources of variability (variance components) and interactions involved in the study are estimated. In the D-study, these estimated variance components are optimized and the conditions for the most appropriate sources of variability are tried to be determined. In addition to the G-studies of this developed measurement tool, S-studies were also conducted to determine the most suitable conditions for the sources of variability.

In light of all these explanations, the aim of this study is to develop a reliable and valid analytical rubric that can enable the evaluation and comparison of application processes and results of students taking the science and chemistry laboratory applications course. In line with the aim, "developing analytical rubric that measures performance levels for Science/Chemistry Laboratory Applications and examining its reliability with generalizability theory and "conducting the decision study by manipulating the number of conditions belonging to the rater variability source", answers were sought to the following questions.

1. What are the variance components estimated for the individual, rater, task, and their interactions for the Science/Chemistry Laboratory Applications analytical rubric?
2. What are the reliability (G and Phi) coefficients of the scores obtained as a result of the analytical rubric evaluation of Science/Chemistry Laboratory Applications?
3. What is the effect of manipulating the number of raters with the D-study on the G and phi coefficients?

METHOD

Working Group

The study group consists of 18 grade 5-6 secondary school students in formal educations, who take science or chemistry laboratory courses in a science and art center in the Southeastern Anatolia Region in the 2020-2021 academic year. In addition, the laboratory performances of students were scored simultaneously by three expert raters using the analytical rubric, which was tried to be developed by the researchers through the process described in detail below.

In addition, during the development process of the rubric, eight experts' opinions were obtained when determinating and arranging performance indicators and performance levels, and these opinions were evaluated by the researchers.

Data Collection Tools

The following steps were carried out on the basis of the steps suggested by Goodrich (2001) and Andrade (1997) in the analytical rubric development process, which was prepared in order to evaluate the process of science and chemistry laboratory applications. The procedures related to this process are detailed below.

1. Literature review was conducted for the evaluation of science and chemistry laboratory process.
2. A review of the literature on performance evaluation was conducted.
3. The following steps have been followed in determining performance indicators and performance levels.

The performance indicators to be used in determining performance, 11 performance indicators were determined for six main tasks for the rubric of science and chemistry laboratory

applications. These tasks are preparation for the experiment, preliminary knowledge of the experiment, preliminary preparation of the experiment, the use of chemicals, the execution of the experiment, and the test result and report. The indicators determined depending on these tasks were determined by considering the behaviors expected from the learner in such laboratory practices. In determining the rubric to be used, it was preferred to use analytical rubrics in the research to be able to conduct the process evaluation in more detail by scoring. In the determination of performance levels, the performance levels were determined as excellent, acceptable, inadequate, and unobserved performance. A score of 3 was given when the performance expected to be observed was fully realized, and 0 was not realized at all.

4. A draft rubric was created.
5. The following steps were followed to receive and evaluate expert opinions and to finalize the draft rubric.

After the analytical rubric was prepared, the opinions of eight experts working at various universities were consulted. One of the experts works in the field of chemistry education (an associate professor), two of them conduct laboratory-based studies in science education and chemistry departments (professor and associate professor), one works in the field of analytical chemistry and teaches analytical chemistry laboratory applications, one works in the field of physical chemistry, teaches the physical chemistry laboratory applications course, one conducts studies in the field of food chemistry, one works in the field of organic chemistry and teaches the organic chemistry laboratory applications course, and one works in the field of measurement and evaluation and has scale development studies. The prepared analytical rubric was evaluated by eight experts in terms of both the tasks and the behavioral indicators of these tasks in terms of suitability and clarity to the target audience. As a result of these evaluations, a consensus was reached and a draft rubric was obtained.

6. Piloting the draft rubric
7. Conducting validity and reliability studies as a result of the pilot application
8. Finalizing the rubric.

Data Collection

The study data were obtained from 5^{th} and 6^{th}-grade students attending chemistry classes in the 2020-2021 academic years in a Science and Art Center located in the Southeastern Anatolia Region of Turkey. The subject of "Solution Preparation", which is among the activities of the students' chemistry laboratory applications, was chosen as a performance determination application. In line with the tasks in the prepared performance evaluation scale, the performance of each student was scored and evaluated independently by three different expert teachers, and the data were collected.

Analysis of Data

In the study, Generalizability Theory was used, which enables the determination of reliability by evaluating all error sources at the same time. Within the scope of GT, a G study was conducted on the main effects of student, task, and rater and the interaction effects depending on these sources of variability, and the variance components were estimated by G studies. In addition, G and Phi coefficients, which are important in making absolute and relative decisions regarding the developed rubric, were calculated. Then, the number of raters was manipulated and the D study was carried out, and suggestions were made for the appropriate number of raters.

Within the scope of the study, the sxtxr (student x task x rater) pattern in which all sources of variability are crossed; is used. In this design, all students (s); are in line with all the tasks (t) included in the rubric; rated by all raters (r). 18 students were evaluated simultaneously by three raters while performing 11 tasks. The G and D studies for the fully crossed "SxTxR" pattern were carried out using the EduG program.

FINDINGS AND DISCUSSION

Examination of the predicted variance components of the individual, rater, task, and their interactions for the analytical rubric of Science/Chemistry Laboratory Applications

In the fully crossed SxRxT pattern, the variance components estimated according to the three raters scoring 18 students in line with 11 tasks in the Science/Chemistry Laboratory Practices analytical rubric are given in Table 1.

Table 1. Components of variance obtained as a result of the G study

Source of Variance	Sum of Squares	df	Mean of Squares	Explained Variance %	Standard Error
S	293.86700	17	17.28629	68.5	0.16995
T	22.38047	10	2.23805	3.7	0.01726
R	0.16498	2	0.08249	0.0	0.00096
ST	44.22559	170	0.26015	4.9	0.01008
SR	1.53199	34	0.04506	0.0	0.00141
TR	11.98316	20	0.59916	3.3	0.01006
STR	50.31987	340	0.14800	19.5	0.01132
Total	424.47306	593		100%	

In Table 1, a total of seven variance components were estimated, namely the main effects of student (S), rater (R), task (T) of the student, rater and task variability sources, the interaction effects of studentxtask (ST), studentxrater (SR) and taskxrater (TR), and residual effects (STR, e). The findings of the variance components obtained from these sources of variability were interpreted as main effects, interaction effects, and residual effects.

Considering the main effects, among all variance percentages, the estimated variance component percentage of students has the highest value (68.5%). The variance estimated for the students reveals the differences between students in terms of performance levels shown during the laboratory applications. In this design, students are the main object of measurement, and it is desirable that all sources of variability, their interactions, and other sources of variability not taken into account in the design, and the percentage estimated for random errors are higher (Güler, 2009). This finding can be interpreted that the main difference in the measurement results related to the performance of science / chemistry laboratory applications is due to students. The high percentage value obtained can be interpreted as heterogeneous in terms of performance among individuals and the measurement tool used was successful in revealing this heterogeneity.

Among the main effects, the second highest (3.7%) component of variance belongs to tasks (T). This shows that the tasks that make up the steps of performance differ in terms of difficulty. It points out that while some tasks are difficult to perform by students as they require complex operations, some tasks are prepared in a way that individuals with basic laboratory knowledge can perform. The variance component (R) related to the source of rater variability gives information about the generosity and rigidity of the raters in terms of scoring. This introduces a systematic error in the existing measurement situation. However, the fact that the percentage of variance for the raters is 0.0% indicates that there is no systematic error originating from

raters. The raters reported very consistent results in their scoring in terms of the feature being measured.

When the percentages of variance components of the interaction effects are examined, it is seen that the highest variance component belongs to the student-task interaction (ST) (4.9%). Students differ in terms of their performance on the basis of tasks. This high variance component may have been caused by confounding variables such as students' being familiar with the laboratory performance in question, their previous interest in laboratory practices, and the difference in difficulty between the steps. In addition, this high variance component in the main effects that make up the interaction, may have caused the variance component of the student-task interaction effect to be high. When the rater and task interaction effect is examined (TR), the main reason for the percentage of variance (3.3%) can be interpreted as the main effect of difficulty and convenience between tasks. The scoring status of the raters varies from item to item. The fact that the variance component of the rater and student interaction (SR) is 0.0% indicates that there is no interaction between the raters and students that may affect the scoring and cause a systematic error.

When the residual variance (SRT) result is examined, it is thought that there are random errors in the measurement and different sources of variability (rater gender, student skill, interest, experience) that are not present in the design also affect the result. However, in this study, contrary to many studies, the fact that the percentage of residual variance was relatively lower than the students, who are the main object of measurement, can be interpreted as the most important sources of variability that can affect this performance are included in the design and the measurement tool used in terms of the feature subject to measurement reveals the differences between students well.

Examining the Reliability (G and Phi) Coefficients of the Scores Obtained as A Result of the Analytical Rubric Evaluation of Science/Chemistry Laboratory Applications

The reliability (G and Phi) coefficients of the scores obtained as a result of the analytical rubric evaluation of Science/Chemistry Laboratory Applications were $G_{relative}$=.99 and $G_{absolute}$=0.98. The coefficients calculated for the G (0.99) relative evaluation and Phi (0.98) absolute evaluation calculated according to the estimated variance components were quite high for this accepted population of observations. Based on these coefficients, it can be interpreted that the performance levels in the measurement tool and the degrees corresponding to the levels defined are correctly determined, the tasks in the rubric are relatively different from each other in terms of difficulty, they distinguish students well in terms of their laboratory performances, and the rater reliability of the raters who carry out the scoring is high.

Examination of G and Phi Coefficients Obtained as A Result of Manipulating the Number of Raters

Based on the variance components examined within the scope of the G study, the results of the decision studies carried out to determine the ideal rater coefficient in a similar measurement situation are given in Table 2. The number of raters was manipulated as 2, 4, 5 and G and Phi coefficients were obtained for each variation.

Table 2. Scoring decision study

	Number of Raters (R)			
	3*	**2**	**4**	**5**
G Coefficient	0.985	0.980	0.987	0.988
Phi Coefficient	0.978	0.974	0.981	0.982

When the results of the G and phi coefficients obtained based on the D studies in Table 2 are examined, it is concluded that there is no significant change in the coefficients when the number of raters is increase or decrease and the measurement situation can be carried out with two raters because of being practical.

CONCLUSION AND RECOMMENDATIONS

It was determined that the scale developed with the findings obtained as a result of the performance scale development study yielded valid, reliable, and generalizable results in determining the performance of the students attending science/chemistry laboratory practices. In this context, it is thought that using this scale, developed to evaluate the students' performance in secondary school science/chemistry laboratory practice courses, will provide valid and reliable measurement results and make the evaluation process more objective.

The variables that make up the main effect in the performance scale were determined as student, task, and rater. While the expected situation in a performance scale is that the effect originating from students is high, the effect from tasks and raters is low (Shavelson & Webb, 1991; Brennan, 2001; Güler, 2009). It can be concluded that the main source of change in the measurement results obtained from the developed performance scale is the students, as the estimated variance value of students has a very high value (Table 1). This indicates that individuals differ considerably from each other in terms of the performance in question and that they are heterogeneous. In other words, individual differences regarding performance were revealed with the measurement tool. This indicates that the reliability and validity of the said measurement tool are high.

As the relative and absolute reliability coefficients ($G_{relative}$=.99 and $G_{absolute}$=0.98) obtained as a result of the analyses for the Science/Chemistry Laboratory Applications analytical rubric are quite high, it can be said that the evaluation results obtained from this tool will give very reliable and valid results. These coefficients show that the tasks that are the subject of performance are defined correctly in the measurement tool and that they distinguish the students well in terms of the feature that is the subject of the measurement because their difficulty levels are different. The Student-Task variance (3.3%) indicates that students may be somewhat related to their experiences on assignments. The task-rater variance (4.9%) suggests that it may have been caused by factors such as the raters' gender, interests, and interactions with the task.

The fact that the variance of the raters is 0.0% (Table 1.) means that there is no error caused by the raters, that is, the raters make objective scoring in terms of the feature being measured and can make consistent evaluations with each other. The low student-rater variance (0.0%) may also mean that the error caused by the raters is low. In other words, it can be said that raters do not exhibit different behaviors from student to student. However, as a result of the D study, it was revealed that the ideal number of raters to be recommended in the performance scale created to measure the performance of science/chemistry laboratory practice is two.

Valid and reliable measurement tools are needed, especially when it comes to evaluating practical and performance-requiring situations (Darling-Hammond, 1994; Shepard, 2008). Therefore, the widespread use of reliable and valid measurement tools for performance evaluation in the literature will increase the accuracy of the evaluation results.

Studies on popularizing the use of Science/Chemistry Laboratory Applications Analytical Rubrics as measurement tools in schools (vocational high schools, science high schools, science

and art centers, secondary schools) affiliated to the National Education, in science and chemistry laboratories in education faculties of universities, in chemistry laboratory practices of pharmacy, science, literature, and engineering faculties can be conducted

In future studies on this subject, it may be recommended to carry out generalizability studies with different sources of variability such as students' gender, whether they have participated in laboratory practices before, and raters' gender.

REFERENCES

Akdemir, Ö. (2006). *İlköğretim II. kademede fen bilgisi öğretmenlerinin laboratuar uygulamalarindaki yeterlikleri ve uygulamalar sirasinda karşilaştiklari sorunlar (Competency of secondary school science teachers in laboratory practices and the problems they face with applications).* Yayımlanmamış yüksek lisans tezi, Fırat Üniversitesi, Elazığ.

Aksoy, G., & Doymuş, K. (2011). Fen ve teknoloji dersinin laboratuar öğretiminde işbirlikli öğrenmenin etkisi. *Erzincan Eğitim Fakültesi Dergisi, 13*(1), 107-122.

Aktamış, H. (2007). *Fen eğitiminde bilimsel süreç becerilerinin bilimsel yaratıcılığa etkisi: ilköğretim 7. sınıf fizik ünitesi örneği*. Yayınlanmamış Doktora Tezi, Dokuz Eylül Üniversitesi, Eğitim Bilimleri Enstitüsü.

Alkan, F., & Erdem, E. (2013). Kendi kendine öğrenmenin laboratuvarda başari, hazirbulunuşluk, laboratuvar becerileri tutumu ve endişeye etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 44*(44), 15-26. Retrieved From https://Dergipark.Org.Tr/Tr/Pub/Hunefd/İssue/7792/101938

Andrade, H. G. (1997). Understanding rubrics. *Educational leadership*, *54*(4), 14-17.

Arı, E. (2008). *Yapılandırmacı yaklaşim ve öğrenme stillerinin genel kimya laboratuar uygulamalarinda öğrencilerin başarisi bilimsel işlem becerileri ve tutumlari üzerine etkisi*. Yayımlanmamış doktora tezi, Marmara Üniversitesi Eğitim Bilimleri Enstitüsü İlköğretim Ana Bilim Dalı, İstanbul.

Arnold, R. J. (2003). The water project: a multi-week laboratory project for undergraduate analytical chemistry. *Journal of Chemical Education*, *80*(1), 58-60.

Atılgan, H. (2005). Genellenebilirlik kurami ve puanlayicilar arasi güvenirlik için örnek bir uygulama. *Eğitim Bilimleri ve Uygulama*, *4*(7), 95-108.

Atmaz, G. (2009). *Puanlama Yönergesi (Rubrik) Kullanılması durumunda puanlayici güvenirliğinin incelenmesi*. Yayınlanmamış yüksek lisans tezi, Mersin Üniversitesi, Mersin.

Ayas, A., Karamustafaoğlu, S., Sevim, S., & Karamustafaoğlu, O. (2002). Genel kimya labaratuvar uygulamalarinin öğrenci ve öğretim elemani gözüyle değerlendirilmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 23*(23), 50-56.

Aydın, F., & Karaçam, S. (2015). Gruplar için teknolojik tasarim uygulamalarini değerlendirmeye yönelik bir analitik rubrik çalışması. *Mersin University Journal of the Faculty of Education*, *11*(1), 132-147.

Aydoğdu, C. (1999). Kimya laboratuvar uygulamalarinda karşilaşilan güçlüklerin saptanması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 15*, 30-35.

Ayvacı, H, Durmuş, A. (2016). Tga yöntemine dayali laboratuvar uygulamalarinin fen bilgisi öğretmen adaylarinin "isi ve sicaklik" konusunda akademik başarilarina etkisi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 39*(39), 101-118. Retrieved From https://Dergipark.Org.Tr/Tr/Pub/Pauefd/İssue/33882/375174

Birinci, K. K., Sezen, G., & Tekbıyık, A. (2010). Fen ve teknoloji derslerinde yapilandirmaci yaklaşima dayali etkinliklerde öğretim teknolojilerinin kullanilabilirliğine yönelik öğretmen görüşleri. *Eğitim Teknolojileri Araştırmaları Dergisi*, *1*(2).

Bozdoğan, A. E., & Altunçekiç, A. (2007). Fen bilgisi öğretmen adaylarinin 5e öğretim modelinin kullanilabilirliği hakkindaki görüşleri. *Kastamonu Eğitim Dergisi, 15*(2), 579-590.

Böyük, U, Demir, S, & Erol, M. (2010). Fen ve teknoloji dersi öğretmenlerinin laboratuvar çalişmalarina yönelik yeterlik görüşlerinin farkli değişkenlere göre incelenmesi. *Tübav Bilim Dergisi*, *3*(4)

Brennan, R. L. (2001). Generalizability theory. Iowa City, IA: ACT PublicationsCollette, E.L., & Chiapetta, A. (1989). *Teaching Science in Middle and Secondary Schools*. Berril Publishing Company, Toronto.

Costa, A. (1985). *Developing Minds: Programs for Teaching Thinking.* USA: ASCD

Criswell, B., (2006). The extraction and isolation of saltpeter from nitered soil. a curriculum alignment project for a first-year high school chemistry course. *Journal of Chemical Education*, 83, 241.

Çepni, S. (2011). "Performansların Değerlendirilmesi". Emin Karip (Ed.). Performansların Değerlendirilmesi (pp.261-262) Ankara: Pegem Akademi.

Çoştu, B., Ayas, A., Çalık, M., Ünal, S., & Karataş, F. (2005). Fen öğretmen adaylarinin çözelti hazirlama ve laboratuvar malzemelerini kullanma yeterliliklerinin belirlenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 28*(28), 65-72. Retrieved From Https://Dergipark.Org.Tr/Tr/Pub/Hunefd/İssue/7808/102421

Dalkıran, E. (2006). *Keman eğitiminde performansın ölçülmesi*. Yayımlanmamış Doktora Tezi, Ankara: Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Darling-Hammond, L. (1994). Setting standards for students: The case for authentic assessment. *The Educational Forum, 59*(1), 14-21.

Deliceoğlu, G. (2009). *Futbol Yetilerine İlişkin Dereceleme Ölçeğinin Genellenebilirlik ve Klasik Test Kuramına Dayalı Güvenirliklerinin Karşılaştırılması.* Yayınlanmamış Doktora Tezi, Ankara Üniversitesi, Ankara.

Demir, S., Böyük, U., & Ayşe, K. O. Ç. (2011). Fen Ve Teknoloji Dersi Öğretmenlerinin Laboratuvar Şartları Ve Kullanımına İlişkin Görüşleri İle Teknolojik Yenilikleri İzleme Eğilimleri. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 7*(2), 66-79.

Ergin, Ö., Şahin-Pekmez, E., & Öngel-Erdal, S. (2005). *Kuramdan uygulamaya deney yoluyla fen öğretimi*. İzmir: Dinazor kitapevi.

Exstrom, C. L., & Mosher, M. D., (2000), A novel high school chemistry camp as an outreach model for regional colleges and universities. *Journal of Chemical Education, 77*(10), 1295-1297.

Geraldo, A., Jofili, Z., & Watts, M. (1999). A course for critical constructivism through action research: a case study from biology. *Research in Science & Technological Education, 17*(1), 518.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercises Science, 5*(1), 13-14.

Goh, N. K., Toh, K. A., & Chia, L. S. (1989). Use of modified laboratory instruction for improving science process skills acquisition. *Journal of Chemical Education, 66*, 430-432.

Güler, N. (2009). Generalizability Theory and comparison of the results of g and d studies computed by spss and genova packet programs. *Education and Science, 34*(154).

Güler N., & Taşdelen Teker G. (2015). Açık uçlu maddelerde farklı yaklaşımlarla elde edilen puanlayıcılar arası güvenirliğin değerlendirilmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6*(1), *12-24.*

Güneş, M. H., Dilek, N. Ş., Topal, N., & Nesrin, C. A. N. (2013). Fen ve teknoloji dersinde laboratuar kullanimina yönelik öğretmen ve öğrenci değerlendirmeleri. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi,* (20), 1-11.

Harle, H. D., Leber, P. A., Hess, K. R., & Yoder, C. H., (2003), A Concept-Based Environmental Project for the First-Year Laboratory: Remediation of Barium Contaminated Soil by In Situ Immobilization. *Journal of Chemical Education, 80*(5), 561-562.

Helvacı, M. A. (2002). Performans yönetimi sürecinde performans değerlendirmenin önemi. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 35*(1-2), 155-169

Hilosky, A., Sutman, F., & Schmuckler, J. (1998). Is laboratory based instruction in beginning collegelevel chemistry worth the effort and expense? *Journal of Chemical Education, 75*, 100-104.

Hofstein, A., & Lunetta, V. N. (1982). The Role of the Laboratory in Science Teaching: Neglected Aspects of Research. *Review of Educational Research, 52*(2), 201-217.

İlhan, H. (2013). *Fen ve teknoloji dersi laboratuvarlarinda öğrenme ortamlarinin yapilan-dirmaci yaklaşima uygunluğunun değerlendirilmesi (Erzurum ili örneği).* Yayımlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

İşman, A. (2001). *Türk eğitim sisteminde ölçme ve değerlendirme*. Adapazarı: Değişim Yayınları.

Jackson, D. J. (2004). *Scaffolding experiments in secondary chemistry to improve content delivery.* Unpublished Master's Thesis, Michigan State University.

Jonsson, A., & Svingby, G. (2007). The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences. *Educational Research Review, 2*, 130–144.

Kaptan, F. (1999). *Fen Bilgisi Öğretimi. İstanbul*. Milli Eğitim Basımevi.

Karamustafaoğlu, S. (2012). Sınıf öğretmeni adaylarinin fen bilgisi laboratuvar uygulamalari- 1 dersi kazanimlarinin kimya deneyleri açisindan incelenmesi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 31*(31), 163-174. Retrieved From Https://Dergipark.Org.Tr/Tr/Pub/Pauefd/İssue/11112/132859

Karatay, R., Doğan, F., & Şahin, Ç. (2014). Determination of attitudes of preservice teachers towards laboratory practices. *Eğitimde Kuram Ve Uygulama, 10*(3), 703-722. Retrieved From Https://Dergipark.Org.Tr/Tr/Pub/Eku/İssue/5461/74098

Kaya, H., & Böyük, U. (2011). Fen bilimleri öğretmenlerinin laboratuvar çalişmalarina yönelik yeterlikleri. *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi, 27*(1), 126-134. Retrieved From Https://Dergipark.Org.Tr/Tr/Pub/Erciyesfen/İssue/25571/269742

Kılıç Mocan, D., Keleş, Ö., & Uzun, N. (2015). Fen bilimleri öğretmenlerinin laboratuvar kullanimina yönelik özyeterlik inançlari: laboratuvar uygulamalari programının etkisi. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi, 17*(1), 218-236.

Kılıç, M., & Aydın, A. (2018). Öğretmenlerin fen bilimleri dersi kapsamında laboratuvar uygulamalari hakkindaki görüşlerinin planlanmiş davraniş teorisi yardimiyla incelenmesi. *Kastamonu Eğitim Dergisi, 26*(1), 241-246. Doi: 10.24106/Kefdergi.378575

Kocakülah, A., & Savaş, E. (2011). Fen bilgisi öğretmen adaylarinin deney tasarlama ve uygulama sürecine ilişkin görüşleri. *Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi, 30*(1), 1-28. Retrieved From Https://Dergipark.Org.Tr/Tr/Pub/Omuefd/İssue/20250/214851

Korkmaz, H. (2000). Fen öğretiminde araç gereç kullanimi ve laboratuvar uygulamalari açisindan öğretmen yeterlikleri. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 19*(19), Retrieved From Https://Dergipark.Org.Tr/Tr/Pub/Hunefd/İssue/7819/102762

Lapadat, J. E. (2000). Construction of science knowledge: scaffolding conceptual change through discourse. *Journal of Classroom Interaction, 35*(2), 1-14.

Leach, J. (1998). Teaching about the world of science in the laboratory: the influence of student' ideas. In J. Wellington (Ed.), *Practical work in school: which way we now?* (pp.52-68). London and Newyork: Routledge.

Liang, L. L., & Gabel, D. L. (2005). Effectiveness of a constructivist approach to science instruction for prospective elementary teachers. International

Lunetta, N. V., & Tamir, P. (1978). An analysis of laboratory activities: project physics and PSSC. *Journal of Biological Education, 40*, 635-642.

Lunetta, V. N. (1998). The School Science Laboratory. Historical Perspective and Centers for Contemporary Teaching. In P. Fensham (Ed.) *Developments and Dilemmas in Science Education*. Falmer Pres, London.

MEB. (2005). *İlköğretim 6., 7. ve 8. sınıf fen ve teknoloji dersi öğretim programları.* Ankara: Milli Eğitim Bakanlığı Talim ve Terbiye Kurulu Başkanlığı.

Meriç, G. (2003). Bir değerlendirme ve laboratuar araci olarak v-diyagraminin tarihi, kullanimi ve fen eğitimine sağlayacağı katkilar üzerine bir inceleme. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 13*(13), 136-149. Retrieved From Https://Dergipark.Org.Tr/Tr/Pub/Pauefd/İssue/11130/133123

Nakiboğlu, C., & Meriç, G. (2000). Genel kimya lâboratuvarlarinda V-diyagrami kullanimi ve uygulamalari. *BAÜ Fen Bilimleri Enstitüsü Dergisi, 2*(1), 58-75.

Nakiboğlu, C., Benlikaya, R., & Karakoç, Ö. (2001). Ortaöğretim kimya derslerinde V-diyagrami uygulamalari. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 21*(21).

Novak, A. (1964). *Scientific inquiry*. Bioscience, (14), 25-28.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: a review. *Educational Research Review*, *9*, 129-144.

Parlak, B., & Doğan, N. (2014). Dereceli Puanlama Anahtarı ve Puanlama Anahtarından Elde Edilen Puanların Uyum Düzeyleri. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *29*(2), 189-197.

Popham, J. W. (1997). What's wrong and what's right with rubric. *Educational Leadership, 55*(2), 12.

Pugh, D. (1991). *Organizational Behaviour*. Prentice Hall Interneational (UK) Ltd.

Rehorek J. S. (2004). Inquiry-based teaching: an example of descriptive science in action. *American Biology Teacher, 66*(7), 493-500.

Renner, J. W. (1986). Rediscovering the lab. The Science Teacher, 44–45.

Saunders, W. (1992). The constructivist perspective: implications and teaching strategies for science. *School Science and Mathematics, 92*(3), 136-141.

Selco, J.I., Roberts, J. L., & Wacks, D. B., (2003). The analysis of seawater: a laboratory-centered learning project in general chemistry. *Journal of Chemical Education, 80*(1), 54-57.

Sezer, S. (2005). Öğrencinin Akademik başarisinin belirlenmesinde tamamlayici değerlendirme araci olarak rubrik kullanimi üzerinde bir araştirma. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 18*(18), 61-69.

Shepard, L. (2008). The role of assessment in a learning culture. *Journal of Education, 189*(1/2), 95–106.

Shavelson, R. J., & Webb, M. N. (2003). "Generalizability theory" Encyclopedia of Social Measurement, ed. Kempf-Leonard, Kimberly. Academic Pres, San Diego.

Silberman, R., Day, S., Jeffers, P., Klanderman, K., Phillips, M. G., & Zipp, A. (1987). Unusual laboratory practical examinations for general chemistry. Journal of Chemical Education, 64, 622.

Suits, J. P. (2004). Assessing investigative skill development in inquiry- based and traditional college science laboratory courses. School Science and Mathematics, 104, 6, 248-257.

Taşlıdere, E, Korur, F. (2012). Fen ve teknoloji öğretmen adaylarinin fizik laboratuvarina yönelik tutumlari: Mehmet Akif Ersoy Üniversitesi Örneği. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 1*(23), 295-318. Retrieved From Https://Dergipark.Org.Tr/Tr/Pub/Maeuefd/İssue/19396/2

Tezcan, H., & Aslan, S. (2007). Lise öğrencilerinin çözeltiler konusunu kavramaları üzerine laboratuar destekli öğretim yönteminin etkisi. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, *27*(3), 65-82.

Tezcan, H., & Bilgin, E. (2004). Liselerde çözünürlük konusunun öğretiminde laboratuvar yönteminin ve bazi faktörlerin öğrenci başarisina etkileri. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi*, *24*(3).

Turgut, M. F., & Baykul, Y. (2010). Eğitimde ölçe ve değerlendirme. Ankara: Pegem Akademi.

Uluçınar, Ş., Cansaran, A., & Karaca, A. (2004). Fen bilimleri laboratuvar uygulamalarinin değerlendirilmesi. *Türk Eğitim Bilimleri Dergisi, 2*(4), 465-475. Retrieved From Https://Dergipark.Org.Tr/Tr/Pub/Tebd/İssue/26126/275208

Uluçınar, Ş., Doğan., A, & Kaya, O. (2008). Sınıf öğretmenlerinin fen öğretimi ve laboratuvar uygulamalarina ilişkin görüşleri. *Kastamonu Eğitim Dergisi, 16*(2), 485-494. Retrieved From Https://Dergipark.Org.Tr/En/Pub/Kefdergi/İssue/49100/626541

Uzal, G., Erdem, A., Önen, F., Gürdal, A., & Gürdal, A. (2010). Basit araç gereçlerle yapilan fen deneyleri konusunda öğretmen görüşleri ve gerçekleştirilen hizmet içi eğitimin değerlendirilmesi. *Necatibey Eğitim Fakültesi Elektronik Fen Ve Matematik Eğitimi Dergisi, 4*(1), 64-84. Retrieved From Https://Dergipark.Org.Tr/Tr/Pub/Balikesirnef/İssue/3370/46519

Wyatt, S. (2005). Extending inquiry-based learning to include original experimentation. *Journal of General Education, 54*(2), 83-89.

Yıldız, E., Akpınar, E., Aydoğdu, B., & Ergin, Ö. (2006). Fen bilgisi öğretmenlerinin fen deneylerinin amaçlarina yönelik tutumlari. *Türk Fen Eğitimi Dergisi, 3*(2), 2-18.

Yurdatapan, M. (2013). Probleme dayali laboratuvar etkinliklerinin öğrencilerin bilimsel süreç becerilerine özgüvenine ve öz-yeterliliğine etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, Özel, (1), 421-435.