NEU PRESS

akef

Volume: 7 Issue: 2 Year: 2025 Research Article e-ISSN: 2687-1750

Inter-Rater Reliability Analysis in Performance-Based Assessment: A Comparison of Generalizability Coefficients and Rater Consistency

Mustafa KÖROĞLU[1*]

[1] Erzincan Binali Yıldırım University, Faculty of Education, Department of Educational Sciences, Erzincan, Türkiye

Article Info

Received: 19.07.2025
Accepted: 23.09.2025
Published: 30.09.2025

Keywords:
Generalizability theory,
Performance-based assessment,
Rater reliability.

ABSTRACT

This study examines the reliability of a performance-based assessment designed to measure university students' basic statistical skills within the framework of Generalizability Theory (GT). The participants were 80 students enrolled in a Guidance and Psychological Counseling program. They completed a two-hour exam consisting of 10 applied tasks, which were scored independently by two raters using a detailed analytic rubric. Data were analyzed using a fully crossed design (person × item × rater). The fully crossed design is characterized by every person providing responses to all items, with every rater scoring all student responses. Variance components were estimated via the maximum likelihood method, and 95% confidence intervals were obtained using a bootstrap procedure. Results indicated that 50.2% of the total variance was attributable to students, 25.6% to items, and 16.6% to raters. The relative generalizability coefficient was calculated as .98, while the absolute decision coefficient (Φ) was .81. Increasing the number of items and raters improved the Φ coefficient and reduced error variance. Findings demonstrate that true performance differences among students were reliably captured, although rater effects could not be entirely eliminated. The study demonstrates that rubric use, investment in rater training, and the adoption of a multi-task, multi-rater approach are critically important in performance assessments.

Performansa Dayalı Değerlendirmede Puanlayıcılar Arası Güvenirlik Analizi: Genellenebilirlik Katsayıları ve Puanlayıcı Tutarlılığının Karşılaştırılması

Makale Bilgisi

Geliş Tarihi: 19.07.2025
Kabul Tarihi: 23.09.2025
Yayın Tarihi: 30.09.2025

Keywords:
Genellenebilirlik kuramı, Performansa dayalı değerlendirme, Puanlayıcı güvenirliği.

ÖZET

Bu çalışma üniversite öğrencilerinin temel istatistik becerilerini ölçmeye yönelik performans temelli bir değerlendirme aracının güvenirliğini Genellenebilirlik Kuramı (GK) kapsamında incelemektedir. Araştırmaya Rehberlik ve Psikolojik Danışmanlık programında öğrenim gören 80 öğrenci katılmıştır. Öğrenciler, 10 uygulamalı görevden oluşan iki saatlik bir sınava girmiş ve bu görevler ayrıntılı bir rubrik kullanılarak iki bağımsız puanlayıcı tarafından değerlendirilmiştir. Veriler, tam çapraz desen (kişi × madde × puanlayıcı) üzerinden analiz edilmiştir. Çaprazlanmış desen her bireyin her bir soruya cevap verdiği ve her bir değerlendiricinin tüm öğrencilerin verdiği cevapları değerlendirdiği bir durum olarak bilinmektedir. Varyans bileşenleri maksimum olabilirlik yöntemiyle, %95 güven aralıkları ise bootstrap tekniğiyle hesaplanmıştır. Sonuçlar, varyansın %50,2'sinin öğrencilere, %25,6'sının maddelere ve %16,6'sının puanlayıcılara ait olduğunu göstermiştir. Göreli genellenebilirlik katsayısı ,98; mutlak karar katsayısı (Φ) ise ,81 bulunmuştur. Madde sayısı ve puanlayıcı sayısı artırıldığında Φ katsayısının yükseldiği ve hata varyansının azaldığı belirlenmiştir. Bulgular, öğrenciler arasındaki gerçek performans farklılıklarının güvenilir biçimde ölçülebildiğini, ancak puanlayıcı etkilerinin tamamen ortadan kaldırılamadığını ortaya koymaktadır. Çalışma, rubrik kullanımının, puanlayıcı eğitimine yatırım yapılmasının ve çoklu görev–çoklu puanlayıcı yaklaşımının performans değerlendirmelerinde kritik önem taşıdığını göstermektedir.

To cite this article:

Köroğlu, M. (2025). Inter-rater reliability analysis in performance-based assessment: A comparison of generalizability coefficients and rater consistency. *Ahmet Keleşoğlu Faculty of Education Journal (AKEF), 7*(2), 218-234. https://doi.org/10.38151/akef.2025.158

***Corresponding Author:** Mustafa KÖROĞLU, *mustafa.koroglu@erzincan.edu.tr*

INTRODUCTION

Measurement and evaluation in education are defined as integral and essential components of the instructional process, applied at every stage from planning to final assessment. Through assessment practices, it is possible to determine whether targeted behaviors have been achieved, identify disruptions in the learning process, and evaluate the effectiveness of the instructional program. However, traditional assessment methods such as multiple-choice tests often fall short of reflecting real-life situations. Contrary to the limitations of traditional measurement and evaluation methods, performance-based assessments measure 21st-century competencies particularly critical thinking, problem solving, collaboration, communication, and data literacy through authentic tasks (Mandinach & Gummer, 2016; OECD, 2019). In the Turkish context, the "Skills Framework" and "Literacy Skills" sections of the Century of Türkiye Education Model explicitly position data literacy and digital literacy as cross-curricular components; this, in turn, strengthens the alignment of performance tasks with the national curriculum (MEB, 2024). Moreover, the systematic use of analytic rubrics both yields detailed feedback and enhances inter-rater consistency, thereby supporting the reliability of the measurements (Panadero et al., 2023; Villa et al., 2020). Kutluet al. (2009) emphasize that such conventional approaches are disconnected from authentic life contexts and, therefore, fail to adequately assess what students have truly learned through formats like multiple-choice or fill-in-the-blank tests. These assessments, which rely on responses given in a limited time without access to resources, do not fully capture the skills that students are expected to acquire for everyday life. These limitations have led to a growing emphasis on performance-based assessment methods, which aim to measure higher-order skills more effectively.

Performance-based assessment is defined as the holistic evaluation of students' behaviors in contexts that resemble real-life situations (Fitzpatrick & Morrison, 1971). In this approach, performance refers to both the process and product demonstrated by an individual when faced with a problem situation. Unlike traditional tests, performance-based assessments focus on the observation of higher-order cognitive processes, skills, and abilities. Performance is considered a complex and multidimensional construct, involving advanced levels of cognitive functioning (Kutlu et al., 2009). As such, performance-based assessments are regarded as essential tools for measuring 21st-century skills and applied learning.

In the context of teaching basic statistics, it is emphasized that students should demonstrate not only procedural knowledge such as formulas but also statistical reasoning and problem-solving skills. In the information age, statistical literacy is recognized as a core competency that enables individuals to interpret and use data effectively. Therefore, the practical assessment of statistical skills in teacher education programs is valuable both for gauging students' readiness and for fostering data-driven thinking among future educators.

Performance-based assessments typically involve three fundamental steps: (i) defining an appropriate task, (ii) having students complete the task, and (iii) observing and scoring either the process or the resulting product. The most common scoring tools include checklists and rubrics. Rubrics are widely used because they enable systematic evaluation of student performance based on predefined criteria. Rubrics can be designed either analytically or holistically; in analytic rubrics, performance is broken down into distinct components, each of which is scored separately, allowing for more detailed feedback (Mertler, 2000; Nitko, 2001). Research indicates that analytic rubrics tend to provide greater reliability and more informative feedback than holistic rubrics (Jonsson & Svingby, 2007; Kutlu et al., 2009). Accordingly, appropriate rubric design and structured rater training are considered key factors in enhancing the reliability of performance assessments.

As with any form of assessment, performance-based tasks must be both valid and reliable. Reliability refers to the consistency of measurement results and their freedom from random error

(Baykul, 2000). In performance-based assessments, potential sources of measurement error include task characteristics, student conditions at the time of assessment, and scoring procedures. Among these, raters are frequently identified as a major source of error. The fact that different raters may assign different scores to the same performance introduces unwanted variance into the assessment results. For this reason, rater reliability is regarded as a critical component of overall reliability in performance assessments. Cohen et al. (1996) argue that because performance assessments largely rely on rater judgment, it is essential to involve multiple raters and to calculate inter-rater reliability.

Rater reliability refers to the extent to which measurement results remain consistent regardless of the rater and is defined as the degree of agreement among different raters (Kutlu et al., 2009). The inter-rater reliability coefficient indicates how closely the scores assigned by two or more independent raters align. High inter-rater reliability suggests that results are not significantly affected by subjective rater bias, thereby improving the objectivity of the assessment (Burry-Stock et al., 1996). Several techniques have been employed in the literature to estimate inter-rater reliability, including percent agreement, inter-rater correlations, and analysis of variance (ANOVA) based on score differences. However, simple agreement or correlation statistics have been criticized for not accounting for chance agreement. Consequently, more robust indices such as Cohen's Kappa coefficient have been proposed (Cohen, 1960; Viera & Garrett, 2005). For continuous data, the intraclass correlation coefficient (ICC) is widely used as a preferred reliability index (Shrout & Fleiss, 1979).

While rater consistency is frequently emphasized in the context of performance assessments, raters are not the sole source of measurement error. The quality and number of tasks administered can also significantly influence assessment outcomes. Dunbar et al. (1991) found that the reliability of ratings may vary considerably depending on the specific tasks used. Even when tasks are well-designed, performance scores based on a single task often lack generalizability, and students' consistent performance across tasks cannot be guaranteed. Brennan (2000) highlights that even with well-trained raters and detailed rubrics, inter-task reliability can pose a greater challenge than inter-rater reliability. Accordingly, it is argued that investigations into the reliability of performance-based assessments should consider not only rater consistency but also task quantity, task variability, and other potential sources of error.

In measurement contexts where multiple sources of error are likely to occur, Generalizability Theory (GT) offers a comprehensive framework. GT extends the concept of reliability beyond the Classical Test Theory (CTT), allowing for a detailed examination of multiple sources of measurement error (Cronbach et al., 1972). Whereas CTT conceptualizes the observed score as $X = T + E$, with a single undifferentiated error term, GT distinguishes each potential source of error as a separate facet contributing to variability. This enables the decomposition of total error variance into components such as raters, tasks, and occasions. These variance components are estimated using analysis of variance (ANOVA) procedures.

A core concept of GT is the "universe," defined as the set of all potential conditions under which observations could be made. A G study estimates the variance components within this universe using data from a specific measurement design (e.g., with a particular number of raters and tasks). Based on these estimates, a D study can then be conducted to predict how changes in the measurement design such as increasing the number of tasks or raters would influence reliability.

One of the most important outputs of GT is the generalizability coefficient (G), which reflects the consistency of individuals' scores relative to each other, also referred to as relative reliability. In addition, the phi (Φ) coefficient represents absolute reliability, indicating the degree to which observed scores reflect true scores without being distorted by error. Since Φ takes into account absolute error, it is typically lower than the G coefficient. Both coefficients range from 0 to 1, with higher values

indicating greater reliability (Shavelson & Webb, 1991). In this regard, GT offers two separate reliability coefficients tailored to different evaluation purposes, in contrast to the single reliability index used in CTT. Furthermore, GT provides a comprehensive map of the sources of error, enabling researchers to make holistic judgments regarding both reliability and validity (Güler, 2009; Volpe et al., 2009).

GT is particularly advantageous for complex measurement designs involving multiple raters and tasks. In this study, which involves a two-rater performance assessment, GT provides an ideal framework for evaluating scoring reliability. The use of two independent raters reduces the subjectivity inherent in single-rater scoring and enhances the overall consistency of the results. When a performance is independently scored by two experts, the average score tends to contain less random error than a single rater's score. Moreover, when one rater tends to score more leniently and another more strictly, their combined evaluations offer a balancing effect, leading to a more equitable assessment. Research also shows that increasing the number of independent raters, like increasing the number of items, enhances overall reliability. For example, Kim et al. (2017) reported that achieving a reliability coefficient of .90 requires both multiple tasks and multiple raters, while a reliability of .80 may be attainable with a single rater, provided that a sufficient number of tasks is used. This finding suggests that in many classroom applications, a single task and a single rater may be sufficient, but additional raters are essential when high-stakes decisions require greater precision.

Nevertheless, inter-rater consistency should still be examined statistically, even when multiple raters are employed. Brennan (2000) argues that different reliability indicators complement one another, and thus, both generalizability coefficients and inter-rater reliability indices should be reported concurrently. By evaluating both the generalizability of scores to the broader universe of conditions and the consistency among raters, researchers can offer a more holistic and valid interpretation of measurement reliability.

The unique contribution of this study lies in examining the reliability of a two-rater scoring system in the assessment of performance-based basic statistical skills, through the lens of both Generalizability Theory (GT) and classical inter-rater consistency. Performance-based assessments are increasingly used to evaluate complex skills, yet ensuring dependable scoring remains a persistent methodological need; Generalizability Theory (GT) provides a principled framework to decompose person, item, and rater variance and to inform design decisions (Andersen et al., 2021). However, in the specific domain of applied basic statistical skills scored by two raters with analytic rubrics, empirical evidence on dependability is limited, and studies seldom report GT coefficients alongside classical inter-rater indices such as ICCs (Koo & Li, 2016; Sturgis et al., 2022; Yılmaz, 2024). This study addresses that gap by estimating GT models via REML-based linear mixed-effects procedures and deriving D-study projections for rater and task sampling (Jiang et al., 2020). In parallel, inter-rater consistency is benchmarked with ICCs following current reporting guidelines, yielding complementary evidence and concrete design guidance for future implementations (Andersen et al., 2021; Koo & Li, 2016).

While GT has been widely applied in areas such as writing assessments, clinical performance evaluations, and similar domains, this study addresses a critical gap by evaluating the reliability of performance-based assessments in basic statistics education (Bacon, 2003; Shavelson & Webb, 1991). Therefore, this study aims to determine the reliability of performance tasks designed to assess pre-service teachers' applied statistical competencies, such as data analysis and interpretation.

Specifically, the study seeks to estimate the generalizability coefficient (G) and the phi coefficient (Φ) by analyzing the variance components derived from scores obtained on statistical performance tasks completed by 80 students. Additionally, the consistency between two independent raters will be examined and compared to provide a dual perspective on measurement reliability. This dual approach allows for the identification of specific sources of error in the assessment process, particularly those

arising from raters, and quantifies the extent to which using two raters contributes to the reliability of evaluation outcomes.

It is anticipated that the findings will offer valuable insights into strategies for enhancing reliability in performance-based assessment practices, demonstrate the practical application of GT, and contribute a comparative perspective to the literature on rater reliability in educational measurement. Ultimately, this study is expected to serve as a guide for researchers and educators conducting similar assessments of applied skills, particularly in determining how many raters and tasks are necessary to ensure dependable scoring and how scoring processes can be optimized.

METHOD

The study focuses on the evaluation of the reliability of a performance-based assessment designed to measure basic statistical skills among university students. Variance components were analyzed using Generalizability Theory to identify and quantify potential sources of measurement error, including person, item, and rater effects. The research was carried out with a cross-sectional design, and all data were collected within a single session under standardized classroom conditions. The study utilized analytical rubrics, multiple raters, and a series of applied tasks to capture authentic student performance. All analyses were conducted using advanced statistical techniques, including REML-based variance estimation and bootstrap procedures to ensure robustness and interpretability of the findings (Huebner & Lucht, 2019; Jiang et al., 2020).

Research Design

This study adopts a quantitative methodological research design within the framework of Generalizability Theory (GT) to investigate the reliability of performance-based assessments in higher education. Without manipulating variables, a fully crossed p × i × r design was employed to estimate variance components and compute generalizability coefficients. This design allows for a comprehensive and systematic analysis of measurement reliability under varying conditions and is widely adopted in educational measurement research (Karasar, 2020; Brennan, 2001).

Study Group

The study group consisted of 80 undergraduate students enrolled in the Department of Guidance and Psychological Counseling at a faculty of education in a state university during the 2024–2025 academic year. Participants were selected using purposive sampling and included all students enrolled in the relevant course. Most participants were between the ages of 19 and 20, with a balanced distribution of male and female students. Participation in the study was integrated into a course-related activity, and the evaluation results were also used to provide formative feedback within the course. Since the study did not involve any experimental manipulation or interventions and data were collected as part of a regular classroom activity, all procedures adhered to ethical research principles, including voluntary participation and confidentiality. Student identities were anonymized, and the data were used solely for aggregate statistical analysis.

Data Collection Procedure

Data were collected during a 120-minute in-class assessment at the end of the semester. Students were asked to complete a series of performance-based tasks aligned with the course content in statistics. The assessment session was organized as a formal exam under standardized conditions, with all students completing the tasks simultaneously and under supervision. Prior to the exam, both written and oral instructions were provided, including explanations of each task and the expected outcomes. Students

worked individually and were not permitted to use external resources or communicate with peers. At the end of the allocated time, all answer sheets were collected. The classroom environment was designed to be relaxed yet formal enough to ensure that students could demonstrate their actual competencies. Two researchers were present throughout the session to monitor compliance with instructions and to maintain the integrity of the data collection process. All students completed the assessment, and there were no instances of data loss or incomplete responses. After collecting, all answer sheets were anonymized and prepared for scoring.

Data Collection Instrument

The data were obtained through a performance-based assessment designed to measure students' applied basic statistical skills. The instrument consisted of a set of complementary statistical problems and tasks that reflected core topics in educational statistics, including measures of central tendency and variability, item-level statistical calculations, data interpretation, and graphical literacy. Each task presented a realistic scenario and accompanying data, requiring students to carry out statistical analyses and interpret the results. For example, in one task, a dataset of class exam scores was presented; students were asked to compute measures of central tendency and dispersion (arithmetic mean, median, standard deviation, interquartile range) and to briefly interpret the findings. In another task, dichotomous (correct–incorrect) responses for several items from an achievement test were provided; students were expected to compute item difficulty (p) and the corrected item–total correlation, and to justify improvement suggestions for items with low discrimination. In a third task, a histogram and box plot for the same dataset were supplied; students were required to identify outliers, interpret the distributional form, and select the appropriate measure of central tendency.

To ensure content validity, the initial version of the tasks was reviewed by two faculty members, one specializing in educational measurement and the other in statistics education. Based on expert feedback, task instructions were revised for clarity, ambiguous expressions were corrected, and the alignment of each task with the target student level was reassessed. The final instrument consisted of 10 performance tasks, each carrying equal weight in the total score.

Scoring criteria were predetermined, and a detailed analytic rubric was developed for scoring purposes. The rubric specified the evaluation criteria for each task and defined performance levels for each criterion. For example, dimensions such as “accuracy of calculation,” “appropriate use of statistical reasoning,” and “interpretation and inference” were assessed using a four-point scale ranging from 0 (unsatisfactory) to 3 (excellent). With multiple criteria per task and 10 tasks in total, the maximum possible score was scaled to 100. The rubric design was based on the performance assessment guidelines proposed by Kutlu et al. (2009) and informed by the rubric development principles found in Jonsson and Svingby (2007) and Moskal and Leydens (2000). The rubric was designed to serve both as a feedback tool for students and as a standardized reference for scoring consistency among raters. Prior to implementation, the rubric was introduced to both raters, and a calibration session was conducted to ensure inter-rater agreement and alignment in scoring.

Data Analysis

The raw scores collected from the performance-based assessment were first converted into a long-format dataset according to a fully crossed $p \times i \times r$ design, where each observation was clearly identified by its Person–Item–Rater combination. This data restructuring step is essential for conducting variance components analysis within the framework of Generalizability Theory (GT) (Cronbach et al., 1972).

In the generalizability study, Person was defined as the object of measurement, while Item and Rater were treated as random facets. Variance components were estimated using the Restricted

Maximum Likelihood (REML) method via the R statistical environment. The model included all two-way interaction terms (Person × Item, Person × Rater, Item × Rater) to account for major sources of measurement error, following GT modeling conventions (Shavelson & Webb, 1991).

Both the relative reliability coefficient (G coefficient) and the absolute reliability coefficient (Φ coefficient) were calculated based on the estimated variance components. Reporting these two indices together is essential, as they reflect different decision-making contexts: relative reliability refers to the consistency of rank ordering among individuals, while absolute reliability pertains to the precision of absolute score interpretations (Brennan, 2001).

To assess the uncertainty of the variance component estimates and reliability coefficients, the cluster bootstrap resampling method was employed with 1,000 iterations. This modern technique accounts for person-level clustering and is recommended for deriving more accurate confidence intervals in measurement models involving dependent observations (Efron & Tibshirani, 1994). Accordingly, 95% confidence intervals were computed for all estimated parameters.

Following the generalizability study, a decision study (D-study) was conducted to examine the impact of varying the number of items and raters on reliability. Generalizability (G) and absolute reliability (Φ) coefficients were estimated under different scenarios, with the number of items ranging from 8 to 15 and the number of raters from 1 to 5. The results of this simulation were visualized using a heat map to aid interpretation and application in practical assessment design.

Ethical approval for this study was obtained from the Erzincan Binali Yıldırım University Educational Sciences Ethics Committee (Protocol No: 08/08). All participants provided informed consent prior to data collection. Data were anonymized, securely stored, and used solely for research purposes in line with ethical guidelines.

FINDINGS

A fully crossed p × i × r (person × item × rater) generalizability design was employed, in which 80 students were evaluated by two independent raters on ten performance tasks. Table 1 presents the results of the generalizability study, including the estimated variance components and their percentage contributions to the total observed score variance.

Table 1

Variance Component Estimates Based on Student, Item, Rater, and Their Interactions (Fully Crossed p × i × r Design)

Source of Variance	df	Sum of Squares	Mean Square	Variance Component	% of Total Variance
Person (p)	79	6620,16	83,80	4,10	50,20
Item (i)	9	3040,80	337,87	2,09	25,60
Rater (r)	1	1089,00	1089,00	1,36	16,60
Person × Item	711	498,20	0,70	0,23	2,80
Person × Rater	79	106,20	1,34	0,11	1,40
Item × Rater	9	26,44	2,94	0,03	0,40
Person × Item × Rater (Residual)	711	172,36	0,24	0,24	3,00

As shown in Table 1, the largest proportion of total score variance was attributable to persons (p), accounting for 50.2% of the total variance. This indicates that the assessment tool was effective in capturing genuine differences in performance across individuals, suggesting strong discriminative power and sensitivity to individual ability or readiness levels.

The second largest component, item-related variance (i), accounted for 25.6% of the total variance. This relatively high proportion implies variability in item difficulty or discrimination,

indicating that improving item balance or expanding item coverage could help reduce measurement error and enhance test reliability.

Rater variance (r) contributed 16.6%, which is relatively substantial compared to findings in previous studies. This suggests a moderate level of inconsistency across raters, possibly due to differences in interpretation or scoring standards, and highlights the need for further rater training and rubric calibration.

The person-by-item interaction (p × i) accounted for 2.8%, indicating a minor tendency for individual students to perform inconsistently across specific items. Interaction effects involving raters, namely person-by-rater (p × r) at 1.4% and item-by-rater (i × r) at 0.4% were negligible, suggesting that raters generally applied the scoring criteria consistently across persons and items.

The three-way interaction (p × i × r), which reflects residual variance due to unmodeled or random factors, represented only 3.0% of the total variance. This relatively low proportion implies that the measurement design was adequate, though potential environmental or contextual factors could be explored in future studies to further reduce error.

From the perspective of Generalizability Theory, the high person variance combined with moderate rater variance indicates that the instrument possesses sufficient reliability for distinguishing among students. However, the contribution of raters to measurement error is non-negligible, reinforcing the importance of robust scoring protocols and ongoing rater calibration.

Table 2

Variance Components Relative to the Number of Items and Raters

d_i	d_r	p	i	r	pxi	pxr	ixr	pxixr
8	1	4,10	0,26	1,36	0,03	0,11	0,00	0,03
9	1	4,10	0,23	1,36	0,03	0,11	0,00	0,03
10	1	4,10	0,21	1,36	0,02	0,11	0,00	0,02
11	1	4,10	0,19	1,36	0,02	0,11	0,00	0,02
12	1	4,10	0,17	1,36	0,02	0,11	0,00	0,02
13	1	4,10	0,16	1,36	0,02	0,11	0,00	0,02
14	1	4,10	0,15	1,36	0,02	0,11	0,00	0,02
15	1	4,10	0,14	1,36	0,02	0,11	0,00	0,02
8	2	4,10	0,26	0,68	0,03	0,06	0,00	0,02
9	2	4,10	0,23	0,68	0,03	0,06	0,00	0,01
10	2	4,10	0,21	0,68	0,02	0,06	0,00	0,01
11	2	4,10	0,19	0,68	0,02	0,06	0,00	0,01
12	2	4,10	0,17	0,68	0,02	0,06	0,00	0,01
13	2	4,10	0,16	0,68	0,02	0,06	0,00	0,01
14	2	4,10	0,15	0,68	0,02	0,06	0,00	0,01
15	2	4,10	0,14	0,68	0,02	0,06	0,00	0,01
8	3	4,10	0,26	0,45	0,03	0,04	0,00	0,01
9	3	4,10	0,23	0,45	0,03	0,04	0,00	0,01
10	3	4,10	0,21	0,45	0,02	0,04	0,00	0,01
11	3	4,10	0,19	0,45	0,02	0,04	0,00	0,01
12	3	4,10	0,17	0,45	0,02	0,04	0,00	0,01
13	3	4,10	0,16	0,45	0,02	0,04	0,00	0,01
14	3	4,10	0,15	0,45	0,02	0,04	0,00	0,01
15	3	4,10	0,14	0,45	0,02	0,04	0,00	0,01
8	4	4,10	0,26	0,34	0,03	0,03	0,00	0,01
9	4	4,10	0,23	0,34	0,03	0,03	0,00	0,01
10	4	4,10	0,21	0,34	0,02	0,03	0,00	0,01

11	4	4,10	0,19	0,34	0,02	0,03	0,00	0,01
12	4	4,10	0,17	0,34	0,02	0,03	0,00	0,01
13	4	4,10	0,16	0,34	0,02	0,03	0,00	0,00
14	4	4,10	0,15	0,34	0,02	0,03	0,00	0,00
15	4	4,10	0,14	0,34	0,02	0,03	0,00	0,00
8	5	4,10	0,26	0,27	0,03	0,02	0,00	0,01
9	5	4,10	0,23	0,27	0,03	0,02	0,00	0,01
10	5	4,10	0,21	0,27	0,02	0,02	0,00	0,00
11	5	4,10	0,19	0,27	0,02	0,02	0,00	0,00
12	5	4,10	0,17	0,27	0,02	0,02	0,00	0,00
13	5	4,10	0,16	0,27	0,02	0,02	0,00	0,00
14	5	4,10	0,15	0,27	0,02	0,02	0,00	0,00
15	5	4,10	0,14	0,27	0,02	0,02	0,00	0,00

Note: *The abbreviations in this table represent the following: d_i: Number of items, d_r: Number of raters, p: Person (Student), i: Item, r: Rater, pxi: Person x Item interaction, pxr: Person x Rater interaction, ixr: Item x Rater interaction, pir: Person x Item x Rater interaction (and error variance).*

Findings presented in Table 2 indicate that increasing the number of items (d_i) from eight to fifteen, as well as the number of raters (d_r) from one to five, systematically reduced the error variance components in the measurement process. Specifically, the person variance component (p) remained stable at 4.10 across all conditions, signifying that true performance differences among individuals constitute a dominant and consistent source of variance in the measurement. In contrast, item variance (m) decreased from 0.26 at d_i = 8 to 0.14 at d_i = 15, and rater variance (r) declined from 1.36 at d_r = 1 to 0.27 at d_r = 5, confirming the mitigating effects of expanded test coverage and multiple raters on absolute error variance.

Regarding interaction components, the person × item interaction variance (pi) diminished from 0.03 to 0.02 as the number of items increased, while the person × rater interaction variance (pr) reduced from 0.11 to 0.02 with the addition of raters. These trends suggest that increasing the number of items compensates for extreme performance variations by specific individuals on particular items, whereas involving additional raters attenuates individual-related assessment errors. The item × rater interaction variance (ir) remained negligible (approximately 0.00) throughout, indicating consistent application of scoring criteria by raters and supporting the structural integrity of the rubric. Moreover, the residual or random error component (RCE), unexplained by the model, decreased from 0.03 to a range between 0.00 and 0.01 as both the number of items and raters increased, reflecting improved generalizability of the measurement design.

Interpreted within the Generalizability Theory framework, the persistent high person variance demonstrates that the instrument effectively discriminates among individuals. Concurrently, the gradual reduction in item and rater variances decreases both relative (σ^2_δ) and absolute (σ^2_φ) error variances. Consequently, it is recommended that test designs include at least eight items and a minimum of three raters to enhance reliability. These findings underscore the importance of employing multiple raters in conjunction with a sufficiently large item pool to minimize measurement error, thereby increasing accuracy and consistency. Furthermore, the results highlight the critical role of increasing item numbers and investing in rater training in future assessments to optimize measurement quality.

Table 3

Reliability Coefficients and Phi Values as a Function of Varying Numbers of Items and Raters

Number of Items	Number of Raters	Relative Error Variance ($\sigma^2(\delta)$)	Generalizability Coefficient ($E\rho^2$)	Absolute Error Variance ($\sigma2$ (Δ))	Phi Coefficient (Φ)
8,00	1,00	0,17	0,96	1,79	0,70

9,00	1,00	0,17	0,96	1,75	0,70
10,00	1,00	0,16	0,96	1,73	0,70
11,00	1,00	0,16	0,96	1,70	0,71
12,00	1,00	0,15	0,96	1,68	0,71
13,00	1,00	0,15	0,97	1,67	0,71
14,00	1,00	0,15	0,97	1,65	0,71
15,00	1,00	0,14	0,97	1,64	0,71
8,00	2,00	0,10	0,98	1,04	0,80
9,00	2,00	0,10	0,98	1,01	0,80
10,00	2,00	0,09	0,98	0,98	0,81
11,00	2,00	0,09	0,98	0,96	0,81
12,00	2,00	0,09	0,98	0,94	0,81
13,00	2,00	0,08	0,98	0,92	0,82
14,00	2,00	0,08	0,98	0,91	0,82
15,00	2,00	0,08	0,98	0,90	0,82
8,00	3,00	0,08	0,98	0,79	0,84
9,00	3,00	0,07	0,98	0,76	0,84
10,00	3,00	0,07	0,98	0,73	0,85
11,00	3,00	0,07	0,98	0,71	0,85
12,00	3,00	0,06	0,99	0,69	0,86
13,00	3,00	0,06	0,99	0,67	0,86
14,00	3,00	0,06	0,99	0,66	0,86
15,00	3,00	0,06	0,99	0,65	0,86
8,00	4,00	0,06	0,98	0,66	0,86
9,00	4,00	0,06	0,99	0,63	0,87
10,00	4,00	0,06	0,99	0,61	0,87
11,00	4,00	0,05	0,99	0,58	0,88
12,00	4,00	0,05	0,99	0,57	0,88
13,00	4,00	0,05	0,99	0,55	0,88
14,00	4,00	0,05	0,99	0,54	0,88
15,00	4,00	0,05	0,99	0,53	0,89
8,00	5,00	0,06	0,99	0,59	0,87
9,00	5,00	0,05	0,99	0,56	0,88
10,00	5,00	0,05	0,99	0,53	0,89
11,00	5,00	0,05	0,99	0,51	0,89
12,00	5,00	0,05	0,99	0,49	0,89
13,00	5,00	0,04	0,99	0,48	0,90
14,00	5,00	0,04	0,99	0,46	0,90
15,00	5,00	0,04	0,99	0,45	0,90

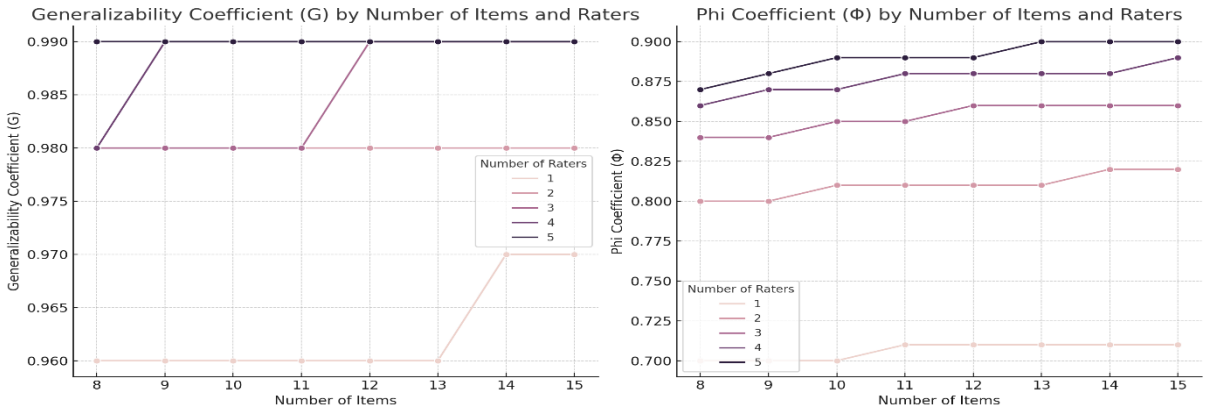
The results presented in Table 3 demonstrate that increasing both the number of items (d_i) and raters (d_r) exerts a systematic and favorable impact on reliability indicators. Specifically, the relative error variance ($\sigma^2(\delta)$) decreased from 0.17 with 8 items and a single rater to 0.04 with 15 items and five raters, evidencing a progressive reduction in relative error as test length and rater count increased. Similarly, the absolute error variance ($\sigma^2(\Delta)$) diminished from 1.79 to 0.45 across the same conditions, indicating that absolute error sources were effectively attenuated through both expanded item sampling and a multi-rater approach.

Generalizability coefficients ($E\rho^2$) were initially high at 0.96 and increased to 0.99 as either the number of items or raters rose, reflecting enhanced reliability in score generalization. Correspondingly, the phi coefficient (Φ), which pertains to decision consistency and absolute error control, increased from 0.70 (8 items, 1 rater) to 0.87 (5 raters) and ultimately reached 0.90 when the item count was increased to 15. This demonstrates that the reduction in absolute error variance is directly mirrored in improved decision reliability, surpassing the commonly accepted threshold of 0.80 for high-stakes decisions.

Moreover, the findings indicate that increases in the number of items contribute more steadily to reliability improvement compared to increases in the number of raters. For instance, under the single-rater condition, expanding items from 8 to 15 resulted in a modest but meaningful decrease in absolute error variance ($\sigma^2(\Delta)$: 1.79 → 1.64). In contrast, holding items constant while increasing raters from one to two yielded a substantial 0.10-point increase in the phi coefficient (0.70 → 0.80) and reduced absolute error variance by nearly 42% (1.79 → 1.04). These results underscore the critical role of both item quantity and rater number in enhancing measurement reliability, with particular emphasis on the efficacy of multiple raters in mitigating absolute measurement error.

Figure 1

Generalizability and Phi Coefficients Across Varying Numbers of Items and Raters



As shown in Figure 1, both the generalizability coefficient (G) and the phi coefficient (Φ) increase as the number of items and raters increases. Specifically, the generalizability coefficient reaches or exceeds the .98 threshold when two or more raters are included, indicating a high level of relative measurement reliability. Similarly, the phi coefficient, which reflects absolute decision consistency, shows a steady improvement across conditions, particularly with increased numbers of raters and items. These patterns demonstrate that enhancing the assessment design by increasing the number of measurement facets effectively reduces measurement error and improves the dependability of scores.

Figure 2

Heat Map of the Generalizability Coefficient ($E\rho^2$) According to the Number of Items and Raters

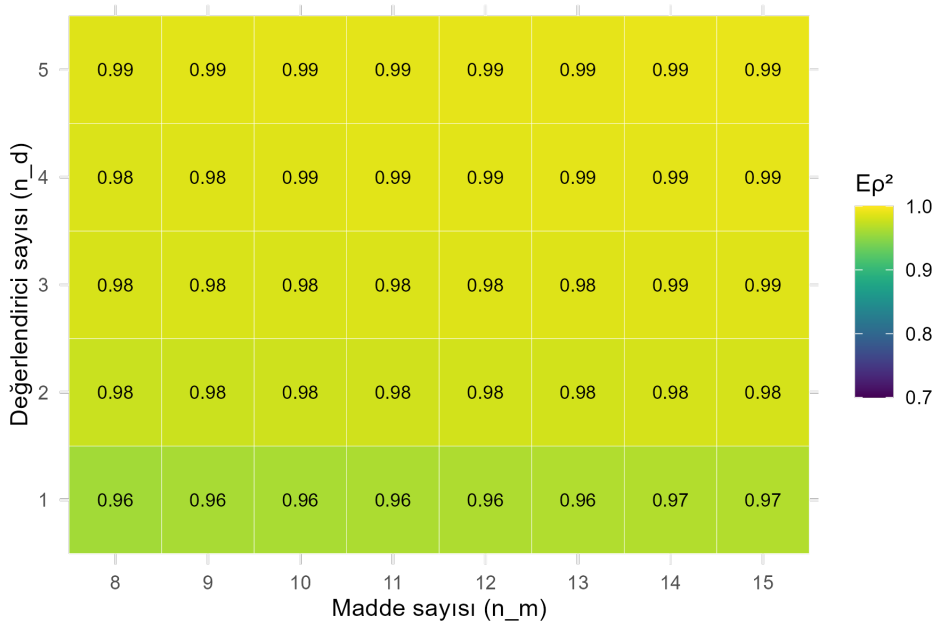


Figure 2 graphically depicts the impact of varying combinations of items (d_i) and raters (d_r) on the generalizability coefficient ($E\rho^2$). Under the condition of 8 items with a single rater, reliability was already notably high ($E\rho^2 \approx 0.96$). However, increasing the number of raters from one to two resulted in an immediate rise in $E\rho^2$ to approximately 0.98. This sharp increase suggests a substantial reduction in rater-related error variance, leading to more consistent capture of individual performance differences.

Further increases in raters from three to five yielded only marginal improvements, with $E\rho^2$ values ranging between 0.98 and 0.99. This pattern reflects the principle of diminishing returns, indicating that the incremental benefit of adding raters decreases as their number grows.

Regarding the number of items, its contribution to reliability was more modest but consistently stable. Specifically, when items increased from 8 to 15 in the single-rater context, $E\rho^2$ improved from 0.96 to 0.97. Although this improvement is less pronounced than that associated with increasing raters, it remains statistically meaningful, particularly when only one rater is involved. When the number of raters is two or greater, the effect of increasing items on $E\rho^2$ becomes negligible, and the generalizability coefficient stabilizes within the range of 0.98 to 0.99.

CONCLUSION AND DISCUSSION

The results of the generalizability theory (GT) analysis involving two raters demonstrate that the performance-based assessment of basic statistics skills exhibited high reliability. Analysis of variance components revealed that the largest proportion of the total variance (50.2%) was attributable to individual differences among students. This finding indicates that the assessment instrument effectively captured true performance variability, thus distinguishing individuals at an adequate level. The second largest source of variance was associated with items (tasks), accounting for 25.6%, suggesting that differences in item difficulty and discrimination significantly influence measurement outcomes. Consequently, employing a more balanced item pool or broadening the scope of assessment content may help mitigate this source of error. Inter-rater variance was identified as 16.6%, which implies that evaluation standards among raters are not completely aligned. Although the two-rater system enhanced reliability compared to single-rater conditions, rater effects were not fully eliminated. As a result, the computed generalizability coefficient (G) was approximately 0.98, reflecting excellent reliability. Regarding absolute decision reliability, the phi coefficient (Φ) was initially calculated as 0.81. The use of two raters combined with a sufficient number of tasks elevated Φ above the 0.80 threshold, reaching the desired level of reliability. These findings highlight the critical role of utilizing multiple raters and multiple tasks in improving reliability within performance-based assessments.

The results of this study are largely consistent with previous research on the reliability of performance-based measurement, while also offering noteworthy insights. It is commonly accepted and empirically observed that the largest variance component in GT analyses corresponds to individuals, here 50.2%. Prior studies in performance assessment have similarly reported that individual differences constitute the major source of score variance, thereby confirming the instrument's capacity to capture genuine ability or skill variations among students (Brennan, 2000; Shavelson & Webb, 1991). For instance, Shavelson et al. (1993) reported the highest variance component originating from individuals in a generalizability study of a science performance test.

The item/task variance, representing 25.6% of total variance, underscores a frequently noted challenge in performance assessments: task inconsistency. Substantial variation in task difficulty and discrimination contributes to variability in overall student performance scores and consequently introduces error variance. Dunbar et al. (1991) emphasized that if tasks within performance assessments are not carefully designed and balanced, reliability suffers. Increasing the number of tasks or broadening the content domain addressed by the tasks can enhance overall reliability by reducing task-related error

variance (Fitzpatrick & Morrison, 1971; Kutlu et al., 2009). Consistent with this, the decision study in the present research demonstrated that increasing items from 8 to 15 significantly improved the reliability coefficient, particularly under single-rater conditions ($E\rho^2$ increased from 0.96 to 0.97). This finding aligns with the well-established "diminishing returns" principle, which posits that beyond a certain point, adding more tasks yields only marginal gains in reliability (Brennan, 2001; Jonsson & Svingby, 2007).

The findings from this study reveal that inter-rater variance accounts for a substantial proportion of total variance, measured at 16.6%. This indicates that the independent evaluations conducted by the two raters did not fully align with identical standards, highlighting the issue of rater inconsistency widely documented in the literature. Cohen et al.(1996) emphasized that inconsistent scoring among different raters in performance-based assessments introduces unwanted error variance, thereby making rater reliability a critical component of measurement accuracy. This suggests that a complete consensus on the evaluation criteria was not achieved among raters in this study. Although a rubric was employed during the assessment, it is plausible that the rubric lacked sufficient clarity or that inadequacies in rater training contributed to discrepancies between raters. Jonsson and Svingby (2007) similarly noted that rubrics alone do not guarantee high reliability unless they are carefully designed and properly implemented; moreover, rater training and calibration are essential to establish common standards.

Conversely, extant research consistently demonstrates that increasing the number of raters enhances scoring reliability (Burry-Stock et al., 1996; Shrout & Fleiss, 1979; Sudweeks et al., 2004). It is widely accepted that employing a two-rater system yields fairer and more consistent scores by balancing the subjectivity inherent in a single rater's judgments. Even when one rater exhibits leniency and the other strictness, the combined scoring tends to neutralize such biases, promoting objectivity in assessment (Moskal & Leydens, 2000; Viera & Garrett, 2005).

The reliability coefficients obtained in this study are also noteworthy when compared to established benchmarks in the literature. The generalizability coefficient was calculated as 0.98, a value considered highly acceptable for reliability standards. The phi coefficient (Φ), reflecting decision reliability, was initially 0.81, exceeding the commonly recommended minimum threshold of 0.80 (Brennan, 2001; Shavelson & Webb, 1991). This threshold is generally recognized as the lower limit for high-stakes assessments (Linn et al.,1991; Nitko, 2001).

Moreover, the magnitude of these coefficients is consistent with recent findings that position Generalizability Theory as a preferred framework for establishing dependable scoring in performance-based assessments and for partitioning person-, task-, and rater-related error (Andersen et al., 2021). Rater effects particularly severity, central tendency, and misfit have been shown to meaningfully inflate total error, hence the recommendation that both G and Φ coefficients be reported in tandem (Wind, 2018). In high-stakes contexts, targeting $\Phi \geq .80$ as the index of decision reliability remains widely adopted in GT-based validation studies (Peeters et al., 2021). Complementarily, inter-rater consistency is benchmarked using ICC thresholds in line with current reporting guidance ($\geq .75$ "good," $\geq .90$ "excellent"), and the present findings align with these contemporary practices (Koo & Li, 2016).

For example, Kim et al. (2017), in their GT study on elementary writing skills, demonstrated that attaining a reliability level of 0.90 necessitates multiple tasks and multiple raters, whereas a single rater suffices only to reach 0.80 reliability, provided there are enough tasks. These findings caution against the "one test, one rater" practice common in many settings, identifying it as a reliability risk that aligns with the present study's results. Particularly in university classroom contexts, reliance on a single performance task and a single rater may undermine the reliability of scoring.

Therefore, it is imperative to employ multiple independent raters and/or increase the number of

tasks in performance assessments where consequential decisions will be made (Brennan, 2000; Kutlu et al., 2009).

Limitations and Future Directions

Although this study employed a robust methodology and offered detailed reliability analyses, certain limitations should be considered. The sample consisted of students from a single undergraduate program at one institution, which may limit the generalizability of the findings. The relatively homogeneous nature of the participants may have influenced the variance components. Moreover, the number of raters was limited to two; although sufficient for the current design, future studies may benefit from including more raters and a wider range of tasks. Additionally, the study focused on a single-session assessment of basic statistical skills. Future research should replicate this design across different institutions and academic disciplines, with more diverse samples and longitudinal data collection. Exploring rater cognition or integrating automated scoring technologies may also contribute to enhancing scoring consistency and validity in performance-based assessments.

Ethics Committee Approval

The ethics committee approval for this study/research was obtained from Erzincan Binali Yıldırım University Educational Sciences Ethics Committee (Protocol No: 08/08).

Author Contributions

Research Design (CRediT 1) Author

Data Collection (CRediT 2) Author

Research - Data analysis - Validation (CRediT 3-4-6-11) Author

Writing the Article (CRediT 12-13) Author

Revision and Improvement of the Text (CRediT 14) Author

Finance

There was no funding to report for this submission.

Conflict of Interests

The author has no conflicts of interest to disclose.

REFERENCES

Andersen, S. A. W., Nayahangan, L. J., Park, Y. S., & Konge, L. (2021). Use of Generalizability Theory for exploring reliability of and sources of variance in assessment of technical skills: A systematic review and meta-analysis. *Academic Medicine, 96*(11), 1609–1619. https://doi.org/10.1097/ACM.0000000000004150.

Bacon, D. R. (2003). Assessing Learning Outcomes: A Comparison of Multiple-Choice and Short-Answer Questions in a Marketing Context. *Journal of Marketing Education*, *25*(1), 31-36. https://doi.org/10.1177/0273475302250570

Baykul, Y. (2000). *Measurement in education and psychology: Classical test theory and its application.* Ankara: ÖSYM.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339–353. https://doi.org/10.1177/01466210022031796

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag Publishing. https://doi.org/10.1007/978-1-4757-3456-0

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

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Cohen, J., Swerdlik, M. E., & Phillips, S. M. (1996). *Psychological testing and assessment (4th ed.).* California: Mayfield

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Dunbar, S.B., Koretz, D., & Hoover, H.D. (1991). Quality Control in the Development and Use of Performance Assessments. *Applied Measurement in Education, 4*, 289-303.

Efron, B., & Tibshirani, R.J. (1994). *An Introduction to the Bootstrap* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593

Fitzpatrick, R. & Morrison, F. (1971). Performance-based testing. *Educational Technology, 11*(5), 63-64

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). https://doi.org/10.15390/ES.2009.840

Huebner, A., & Lucht, M. (2019). Generalizability theory in R. *Practical Assessment, Research & Evaluation, 24*(5), 1–12. https://doi.org/10.7275/5065-gc10.

Jiang, Z., Raymond, M., Shi, D., & DiStefano, C. (2020). Using a linear mixed-effect model framework to estimate multivariate generalizability theory parameters in R. *Behavior Research Methods, 52*, 2383–2393. https://doi.org/10.3758/s13428-020-01399-z.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144 https://doi.org/10.1016/j.edurev.2007.05.002

Karasar, N. (2020). *Scientific research method* (35th ed.). Ankara: Nobel Academic Publishing.

Kim, G. Y., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in grades 3 and 4. *Reading and Writing, 30*(6), 1287–1310. https://doi.org/10.1007/s11145-017-9724-6

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kutlu, Ö., Doğan, C. D. & Karakaya, İ. (2009). *Determining student achievement: Performance and portfoliobased assessment.* Ankara: Pegem Akademi. http://dx.doi.org/10.14527/9786053647003

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21. https://doi.org/10.3102/0013189X020008015

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate? *Teaching and Teacher Education, 60*, 366–376. https://doi.org/10.1016/j.tate.2016.07.011

Mertler, C. A., (2000). Designing scoring rubrics for your classroom, *Practical Assessment, Research, and Evaluation* 7(1): 25. doi: https://doi.org/10.7275/gcy8-0w24

Moskal, B. M. & Leydens, J. A., (2000) "Scoring Rubric Development: Validity and Reliability", *Practical Assessment, Research, and Evaluation* 7(1): 10. doi: https://doi.org/10.7275/q7rm-gg74

Nitko, A. J. (2001). Educational assessment of students (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

OECD. (2019). *OECD Learning Compass 2030: A series of concept notes*. Paris: OECD.

Panadero, E., Jonsson, A., Pinedo, L., & others. (2023). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic review. *Educational Psychology Review, 35*, 113. https://doi.org/10.1007/s10648-023-09823-4

Peeters, M. J., Cor, M. K., Petite, S. E., & Schroeder, M. N. (2021). Validation Evidence using Generalizability Theory for an Objective Structured Clinical Examination. *Innovations in pharmacy*, *12*(1), 10.24926/iip.v12i1.2110. https://doi.org/10.24926/iip.v12i1.2110

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

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215–232. https://doi.org/10.1111/j.1745-3984.1993.tb00424.x

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sturgis, P. W., Marchand, L., Miller, M. D., Xu, W., & Castiglioni, A. (2022). *Generalizability Theory and its application to institutional research* (AIR Professional File No. 156). Association for Institutional Research. https://doi.org/10.34315/apf1562022

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239–261. https://doi.org/10.1016/j.asw.2004.11.001

T.C. Millî Eğitim Bakanlığı. (2024). *Türkiye Yüzyılı Maarif Modeli: Okuryazarlık Becerileri*. Ankara: MEB.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med*, *37*(5), 360-363.

Villa, K. R., Sprunger, T. L., Walton, A. M., Costello, T. J., & Isaacs, A. N. (2020). Inter-rater reliability of a clinical documentation rubric within pharmacotherapy problem-based learning courses. *American Journal of Pharmaceutical Education, 84*(7), 7648. https://doi.org/10.5688/ajpe7648

Volpe, R. J., McConaughy, S. H., & Hintze, J. M. (2009). Generalizability of Classroom Behavior Problem and On-Task Scores From the Direct Observation Form. *School Psychology Review*, *38*(3), 382–401. https://doi.org/10.1080/02796015.2009.12087822

Wind, S. A. (2018). Examining the Impacts of Rater Effects in Performance Assessments. *Applied Psychological Measurement*, *43*(2), 159-171. https://doi.org/10.1177/0146621618789391

Yılmaz, F. N. (2024). Comparing the reliability of performance task scores obtained from rating scale and analytic rubric using the generalizability theory. *Studies in Educational Evaluation, 83*. https://doi.org/10.1016/j.stueduc.2024.101413