INUGSE
Inonu University Journal of the Graduate School of Education
Volume: 12, Issue: 24, October 2025

Assessing the Reliability of Open-Ended Exams: A Generalizability Theory Approach to Item and Rater Variance

Mustafa KÖROĞLU
Erzincan Binali Yıldırım University, Erzincan -Türkiye

Article History

Submitted: 12.05.2025
Accepted: 29.09.2025
Published Online: 24.10.2025

Keywords

Generalizability theory, measurement and evaluation, written test, reliability.

Abstract

Purpose: This study examines the reliability of open-ended university exams through the lens of Generalizability Theory (GT), aiming to identify key sources of measurement error.

Design & Methodology: Using a fully crossed person × item × rater (p × i × r) design, a five-item written exam administered to 76 students was scored by two raters. The Generalizability Study (G-Study) revealed that the largest portion of total score variance stemmed from individual student differences (62.2%) and the person × item interaction (30.7%), while item-related (3.9%) and rater-related (1.5%) variance components were relatively minor.

Findings: These results suggest that the exam effectively captures individual performance differences, and that increasing the breadth of item sampling may significantly reduce measurement error. Findings from the Decision Study (D-Study) indicated that expanding the number of items from 4 to 10 and raters from 1 to 5 led to substantial improvements in both relative ($\sigma^2\delta$) and absolute ($\sigma^2\Delta$) error variances. Correspondingly, generalizability and Phi coefficients increased from 0.81 to 0.95. The low rater variance implies that the use of detailed scoring rubrics and rater training contributed to consistent scoring. Moreover, residual error was minimal (1.6%), suggesting measurement model adequacy from a practical standpoint, results recommend increasing item count to at least eight and involving at least three raters to optimize reliability.

Implications & Suggestions: The study demonstrates the effectiveness of GT in dissecting multiple sources of error and offers guidance for improving assessment quality in higher education. Emphasizing item diversity, rater standardization, and data-informed decision-making can strengthen the validity and fairness of exam-based evaluations.

DOI: 10.29129/inujgse.1740879

İnönü Üniversitesi Eğitim Bilimleri Enstitüsü Dergisi
Cilt: 12, Sayı: 24, Ekim 2025

Açık Uçlu Sınavların Güvenirliğinin Değerlendirilmesi: Madde ve Puanlayıcı Varyansına Yönelik Genellenebilirlik Kuramı Yaklaşımı

Mustafa KÖROĞLU
Erzincan Binali Yıldırım Üniversitesi, Erzincan -Türkiye

Makale Geçmişi

Geliş: 12.05.2025
Kabul: 29.09.2025
Online Yayın: 24.10.2025

Anahtar Sözcükler

Genellenebilirlik kuramı, ölçme ve değerlendirme, yazılı sınav, güvenirlik.

Öz

Amaç: Bu çalışma, üniversitelerde uygulanan açık uçlu sınavların güvenirliğini Genellenebilirlik Kuramı (GK) çerçevesinde inceleyerek ölçme hatasının temel kaynaklarını belirlemeyi amaçlamaktadır.

Yöntem: Tam çapraz p × i × r (kişi × madde × puanlayıcı) desenine göre tasarlanan araştırmada, 76 öğrenciye uygulanan beş maddelik yazılı sınav iki puanlayıcı tarafından değerlendirilmiştir.

Bulgular: Genellenebilirlik çalışması (G-Study) sonucunda toplam varyansın en büyük kısmının bireysel öğrenci farklılıklarından (%62,2) ve kişi × madde etkileşiminden (%30,7) kaynaklandığı, buna karşın madde (%3,9) ve puanlayıcı (%1,5) kaynaklı varyans bileşenlerinin oldukça düşük olduğu görülmüştür. Bu bulgular, sınavın bireysel performans farklarını etkili biçimde yansıttığını ve madde kapsamının artırılmasının ölçme hatasını önemli ölçüde azaltabileceğini göstermektedir. Karar çalışması (D-Study) kapsamında madde sayısının 4'ten 10'a, puanlayıcı sayısının ise 1'den 5'e çıkarılması hem bağıl ($\sigma^2\delta$) hem de mutlak hata varyanslarını azaltmış; buna paralel olarak genellenebilirlik katsayısı ($E\rho^2$) ve Phi katsayısı (Φ) 0,81'den 0,95'e yükselmiştir. Puanlayıcı varyansının düşük olması, ayrıntılı puanlama anahtarları ve puanlayıcı eğitimlerinin tutarlı puanlamaya katkı sunduğunu göstermektedir. Ayrıca artık varyansın %1,6 gibi düşük bir düzeyde olması, modelin büyük ölçüde yeterli olduğunu göstermektedir. Uygulama açısından, güvenilirliği artırmak için öncelikle madde sayısının sekize çıkarılması ve ardından en az üç puanlayıcının görevlendirilmesi önerilmektedir.

Sonuçlar ve Öneriler: Çalışma, GK'nin çoklu hata kaynaklarını çözümlemedeki gücünü ortaya koymakta ve yükseköğretimde değerlendirme süreçlerinin kalitesini artırmaya yönelik somut öneriler sunmaktadır.

DOI: 10.29129/inujgse.1740879

INTRODUCTION

Measurement and evaluation in education are fundamental processes that enable the objective, valid, and reliable assessment of both learning processes and outcomes (Özçelik, 2016; Turgut & Baykul, 2021). While measurement refers to the quantification of a characteristic or behavior using specific units, evaluation involves interpreting these measurement results according to predetermined criteria or standards (Baykul, 2021). Within the educational system, these processes are critical for systematically analyzing students' learning levels and developmental progress. Moreover, measurement and evaluation play a central role in assessing the effectiveness of teaching practices and identifying individual differences among students (Güler, 2011). These practices guide educators in diagnosing learning deficiencies, revising instructional methods, and improving curriculum design (Turgut & Baykul, 2021).

A wide range of measurement tools have been developed to assess students' learning processes and uncover individual differences. These tools include multiple-choice tests, short-answer tests, true-false items, and extended-response formats (Doğan, 2024). Among them, written exams have long held a prominent place in educational settings. They are particularly favored for evaluating students' knowledge levels and higher order thinking skills such as problem-solving and critical analysis (Güler, 2011). Written exams provide opportunities for students to reflect, analyze, and articulate their ideas on a given topic. Therefore, ensuring content validity and construct validity in the design of exam questions is of great importance (Turgut & Baykul, 2021). In terms of reliability, it is essential to maintain the consistency of exam scores and minimize measurement errors. Both validity and reliability are indispensable for written exams to accurately reflect students' academic performance (Baykul, 2021).

Reliability, defined as the consistency and stability of measurement results, is a cornerstone of any effective assessment. An exam is considered reliable when it yields consistent results across different administrations, contexts, or raters (Baker, 2001). To enhance reliability, test instruments must be carefully constructed, and measurement errors must be reduced as much as possible (Erkuş et al., 2024). Measurement errors compromise result accuracy. They are generally categorized as systematic or random. Systematic errors consistently bias scores in one direction, whereas random errors cause unpredictable fluctuations in measurement outcomes (Koğar, 2023). Minimizing both types of errors is vital to obtaining valid and reliable exam results. Accordingly, addressing measurement error requires careful attention not only to the quality of test items but also to external factors that may influence student performance (Baker, 2001; Koğar & Koğar, 2022).

Classical Test Theory (CTT) provides a foundational framework for understanding the relationship between reliability and measurement error. According to CTT, an observed test score is composed of two components: a student's true score and a measurement error. This relationship is mathematically expressed as $X = T + E$, where X represents the observed score, T the true score, and E the error component (Baker, 2001). Reliability in CTT is quantified as a coefficient that reflects the internal consistency of the test, guiding efforts to enhance the accuracy of measurement outcomes (Erkuş et al., 2024; Koğar, 2023). The theory assumes that measurement errors are random, have a mean of zero, and do not follow any systematic pattern (Bloch & Norman, 2012).

However, this central assumption of random error presents a limitation in situations where multiple sources of error influence the measurement process. Although CTT acknowledges that measurement error may arise from various factors such as test administration conditions, the clarity and content of items, or the temporary psychological states of examinees it aggregates all such sources into a single undifferentiated error term. For instance, errors in test results may stem from fluctuations in a student's mood, inconsistencies in item difficulty, or external distractions like environmental noise. Under CTT, all these diverse sources are grouped into a single error component, which restricts the ability to conduct a more nuanced analysis of measurement quality (Vispoel et al., 2018).

CTT defines reliability as the proportion of true score variance within the total observed score variance. Accordingly, minimizing error variance leads to an increase in the reliability coefficient. However, since CTT does not decompose error into its distinct sources, it does not allow for an evaluation of how specific factors such as exam length or item content individually affect reliability (Baker, 2001; Koğar & Koğar, 2022). This limitation renders CTT inadequate for more complex measurement contexts where multiple interacting sources of error are present.

In response to this need, Generalizability Theory (GT) was developed as a more comprehensive approach to measurement. GT enables a deeper investigation into the measurement process by simultaneously modeling multiple sources of error that influence test results. It allows for the decomposition of error variance across various facets, such as item difficulty, rater consistency, and environmental conditions, offering a more detailed and realistic estimation of measurement reliability (Baker, 2001). By doing so, GT addresses the limitations of CTT and provides a more robust framework for evaluating the reliability of educational assessments in complex and multifaceted testing situations (VanLeeuwen et al., 1999)
Generalizability Theory (GT), developed by Cronbach and colleagues in 1963, was introduced as an extension of Classical Test Theory (CTT) to address its limitations (Brennan, 2001; Cronbach et al., 1963). While CTT reduces all sources of measurement error to a single undifferentiated component, GT allows for the identification and estimation of multiple sources of error. It decomposes an observed score into a universe score (analogous to the true score in CTT) and multiple error components (Shavelson & Webb, 1991). This enables a more nuanced understanding of how both systematic and random factors influence measurement outcomes. The primary aim of GT is to estimate the various components of variance in a measurement setting and assess their relative contributions. For example, GT allows for the separate analysis of the effects of items, raters, and environmental factors on student performance (Vispoel et al., 2017).

GT provides a comprehensive theoretical framework for examining the reliability of measurement instruments by adopting a multifaceted perspective. Unlike CTT, it enables the simultaneous analysis of multiple sources of error that can affect measurement precision, making it particularly valuable in fields such as education, health, and the social sciences (Shavelson & Webb, 1991). At the core of GT is the concept of a design, which refers to the specific configuration of facets (e.g., persons, items, raters, occasions) in a measurement setting. Understanding the structure and interaction of these facets is essential for accurately identifying the sources of measurement error and for developing robust instruments (Brennan, 2001). In this context, designs (also referred to as patterns or models) describe how facets are organized and how they relate to one another within the measurement process. Accurate specification of these designs facilitates the separation of variance components and enhances the interpretability of measurement results (Cardinet et al., 2009).
The structure of a design depends on the number of facets involved and how they are organized (i.e., crossed or nested). In a single-facet design, only one facet is analyzed. For example, if students are assessed solely based on test items, a one-facet design ($p \times i$) can be used to examine the variance attributable to persons and items (Shavelson & Webb, 1991). In contrast, multi-facet designs involve two or more facets. Whether the facets are crossed or nested affects how variance components are estimated. For instance, in a design where all students respond to all items and all items are evaluated by all raters, a fully crossed $p \times i \times r$ design is used. However, if each rater scores only a subset of items, a nested design is applied (Brennan, 2001). Such design structures are critical for analyzing measurement error in detail and for determining the extent to which error is attributable to specific facets such as persons, items, or raters (Huebner & Lucht, 2019).
GT employs two main types of studies: the Generalizability Study (G-Study) and the Decision Study (D-Study). The G-Study estimates the variance components associated with each facet, thus identifying

potential sources of measurement error. In contrast, the D-Study utilizes these variance estimates to simulate and recommend optimal measurement conditions for specific decision-making contexts (Brennan, 2001; Shavelson & Webb, 1991;). The G-Study typically relies on analysis of variance (ANOVA) techniques to quantify the contributions of each facet to total score variance. These results are then used in the D-Study to compute reliability indices such as the generalizability coefficient ($E\rho^2$) and the phi coefficient (Φ), and to optimize measurement designs. For instance, D-Studies can inform decisions about the ideal number of items, raters, or test occasions needed to improve reliability (Brennan, 2001). Overall, GT offers a powerful methodological approach for identifying, quantifying, and interpreting the various sources of error in measurement processes. Its applications span diverse fields, from educational testing to clinical assessment. Numerous studies in the literature have demonstrated how generalizability coefficients and phi coefficients derived from different design structures can inform the development and refinement of measurement tools (Brennan, 2001; Shavelson & Webb, 1991). GT's ability to disentangle multiple sources of variance makes it an essential framework for enhancing the validity and reliability of assessment practices in complex measurement contexts.

In the field of health sciences, Generalizability Theory (GT) has been widely employed to evaluate clinical competencies and analyze the performance of healthcare professionals. For example, Monteiro et al. (2019) utilized GT to enhance the reliability of Objective Structured Clinical Examinations (OSCEs). Their study examined rater consistency and evaluated the validity of the overall assessment process. Similarly, Smith and Paige (2019) used GT to analyze inter-rater consistency in clinical assessment scenarios and provided practical recommendations to optimize evaluation procedures. In the educational domain, Kara and Kelecioğlu (2015) applied GT to analyze the influence of different panels of teachers and experts on cut scores, thereby demonstrating the theory's utility in standard-setting procedures.

To promote the broader use of GT and increase its visibility in the literature, researchers have conducted studies not only using real-world data but also simulated (artificial) datasets. Güler (2011), for instance, compared the reliability coefficients derived from GT and Classical Test Theory (CTT) using randomly generated data. By applying both crossed and nested designs to a sample of 125 simulated individuals, the study evaluated how each theory contributed to reliability analysis. The results showed that GT was more effective in distinguishing between different sources of error. In crossed designs, the true score and error variances were analyzed separately, offering a more nuanced understanding than CTT. In nested designs, the effects of items and rater evaluations were examined, confirming GT's versatility in handling complex measurement structures. The study emphasized that GT provides a more flexible and powerful alternative, especially in measurement settings involving multiple error sources.

Similarly, Atılgan (2005) investigated inter-rater reliability by decomposing error variance using both crossed and nested designs. GT enabled a detailed analysis of how rater variability influenced the assessment process. In particular, the study demonstrated the impact of rater inconsistency on measurement results, especially in settings where multiple raters evaluated the same individuals or items.

When the literature is examined, it becomes evident that subjective scoring particularly in the evaluation of extended-response (long-answer) exams poses significant challenges to measurement reliability. Therefore, analyzing such exams through the lens of Generalizability Theory is crucial. This approach enables the decomposition of error variance, the identification of reliability threats, and the development of strategies to enhance the accuracy of assessment outcomes.

Purpose of the Study

The primary aim of this study is to examine the reliability of written exams administered in educational measurement and evaluation courses at the university level through the lens of Generalizability Theory.

Specifically, the study seeks to evaluate how items and scorers contribute to measurement error and to determine how these factors influence the reliability of the assessments. By decomposing sources of error, the study aims to identify factors that positively or negatively affect reliability and to offer data-driven recommendations for improving the assessment process.

Sub-Purposes of the Study

- To estimate the generalizability coefficient ($E\rho^2$) of written exams administered in universities.
- To estimate the Phi coefficient (Φ), which reflects absolute reliability, of written exams.
- To conduct decision studies (D-Studies) based on the variance components identified in the G-Study to determine optimal testing conditions.

METHODOLOGY

This study adopts a quantitative research approach grounded in the framework of Generalizability Theory (GT) to examine the reliability of written exams administered at the university level. The primary goal is to estimate and interpret variance components associated with persons, items, and raters in a fully crossed measurement design.

Research Design

This study employed a quantitative non-experimental design within the framework of Generalizability Theory (GT). Specifically, a fully crossed p × i × r (person × item × rater) design was implemented to estimate and interpret variance components associated with students, items, and raters. While descriptive in nature, this design enables a systematic examination of multiple error sources without manipulating variables, thereby providing robust evidence on the reliability of open-ended exam scores (Brennan, 2001; Karasar, 2020).

Study Group

The research group consists of 76 students enrolled in the Guidance and Psychological Counseling Program at a state university during the 2024-2025 academic year. All students in the course (n=76) were included as an accessible census sample. Although the sample is limited, the variance decomposition yielded stable estimates, suggesting sufficient power for the design. The raters determined within the scope of the study had at least one year of experience in their fields and served as raters in this study. Furthermore, the assessment data were analyzed using a fully crossed p × i × r (person × item × rater) design (Brennan, 2001).

Table 1.

Descriptive Statistics of the Study Group

Categorical Variables	Categories	Frequency	Percentage
Gender	Female	60	.79
	Male	16	.21
	Total	76	.100

Table 1 shows the distribution of the students participating in the study according to their gender. Of the 76 participants, 79% (n = 60) were females and 21% (n = 16) were males.

Data Collection

The data used in the study were obtained through the written exam of measurement and evaluation in an education course consisting of 5 items. The exam was prepared to measure the participants' academic knowledge levels and analytical thinking skills. The raters were allowed to independently evaluate each

item in line with predetermined rubrics. During the evaluations, scores between 0-20 were given for each item and the data obtained were subjected to reliability analysis.

Data Analysis

The data collected were analyzed within the framework of Generalizability Theory (GT). Initially, variance components were estimated, followed by the calculation of the generalizability coefficient ($E\rho^2$) and the Phi coefficient (Φ). All analyses were conducted using R (gtheory) statistical software. The findings were evaluated to formulate recommendations for enhancing the reliability of the measurement instruments. A fully crossed design (person × item × rater) was employed, allowing for the separate analysis of variance components attributable to persons, items, and raters.

Ethical approval for this study was obtained from the Erzincan Binali Yıldırım University Educational Sciences Ethics Committee (Protocol No: E-88012460-050.04-460473, 08/07). 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

Based on Generalizability Theory, a fully crossed p × i × r (person × item × rater) design was employed, in which 76 students were evaluated by two raters across five items. The analysis results, including estimated variance components and their corresponding percentage contributions, are presented in Table 2.

Table 2.

Analysis of Variance Results Based on Student, Item, Rater and Their Interaction

Variables	df	Sum of Squares	Mean Squares	Variance	Percentage
Person (p)	75	15379,22	205,06	18,60	62,20
Item (i)	4	779,47	194,87	1,16	3,90
Rater (r)	1	175,30	175,30	0,46	1,50
Person × Item	300	5652,53	18,84	9,19	30,70
Person × Rater	75	52,60	0,70	0,05	0,20
Item × Rater	4	1,61	0,40	0,00	0,00
Person × Item × Rater	300	139,99	0,47	0,47	1,60

According to Table 2, the largest source of variance in the measurement is the person (p) component, accounting for 62.2% of the total variance. This indicates that the assessment tool is effective in capturing true performance differences among individuals and that the variability in ability or readiness across students is substantially reflected in their test scores. The second largest variance component is the person × item interaction (p × i), contributing 30.7%, which suggests that certain individuals perform differently on specific items relative to others. This underscores the critical role of item sampling in measurement reliability and implies that expanding item coverage in future assessments could help reduce measurement error. Item-related variance (i) accounted for only 3.9% of the total variance, indicating that differences in item difficulty or discrimination were relatively minor and that the items functioned as a structurally homogeneous set. Therefore, it can be concluded that item characteristics had a limited influence on the construct being measured.

The rater (r) variance was found to be notably low at 1.5%, indicating a high level of inter-rater consistency and minimal variance attributable to rater-related error. Similarly, the person × rater (p × r; 0.2%) and item × rater (i × r; 0%) interaction components were negligible, suggesting that rater effects were stable across both persons and items. Additionally, 1.6% of the total variance was attributed to the residual component (p × i × r), which reflects random error and unexplained variance not accounted for by the specified facets. The relatively small magnitude of this residual variance implies that the

measurement model was largely sufficient; however, further investigation of uncontrolled sources such as testing conditions or environmental noise may help reduce residual error in future studies.

Within the framework of Generalizability Theory, the combination of a high person variance (62.2%) and minimal rater variance (1.5%) suggests that the instrument possesses strong discriminative power across individuals, while the influence of raters is negligible. However, the considerable person × item interaction variance (30.7%) highlights the importance of expanding the item pool or ensuring a more balanced item set to enhance measurement reliability. Such enhancements are likely to reduce measurement error and, in turn, contribute to an increase in the Φ coefficient, thereby improving decision reliability.

Table 3.

Variance Components and Error Terms across D-study Designs

d_i	d_r	p	i	r	pxi	pxr	ixr	pxixr
4	1	18,60	0,29	0,46	2,30	0,05	0,00	0,12
5	1	18,60	0,23	0,46	1,84	0,05	0,00	0,09
6	1	18,60	0,19	0,46	1,53	0,05	0,00	0,08
7	1	18,60	0,17	0,46	1,31	0,05	0,00	0,07
8	1	18,60	0,14	0,46	1,15	0,05	0,00	0,06
9	1	18,60	0,13	0,46	1,02	0,05	0,00	0,05
10	1	18,60	0,12	0,46	0,92	0,05	0,00	0,05
4	2	18,60	0,29	0,23	2,30	0,02	0,00	0,06
5	2	18,60	0,23	0,23	1,84	0,02	0,00	0,05
6	2	18,60	0,19	0,23	1,53	0,02	0,00	0,04
7	2	18,60	0,17	0,23	1,31	0,02	0,00	0,03
8	2	18,60	0,14	0,23	1,15	0,02	0,00	0,03
9	2	18,60	0,13	0,23	1,02	0,02	0,00	0,03
10	2	18,60	0,12	0,23	0,92	0,02	0,00	0,02
4	3	18,60	0,29	0,15	2,30	0,02	0,00	0,04
5	3	18,60	0,23	0,15	1,84	0,02	0,00	0,03
6	3	18,60	0,19	0,15	1,53	0,02	0,00	0,03
7	3	18,60	0,17	0,15	1,31	0,02	0,00	0,02
8	3	18,60	0,14	0,15	1,15	0,02	0,00	0,02
9	3	18,60	0,13	0,15	1,02	0,02	0,00	0,02
10	3	18,60	0,12	0,15	0,92	0,02	0,00	0,02
4	4	18,60	0,29	0,11	2,30	0,01	0,00	0,03
5	4	18,60	0,23	0,11	1,84	0,01	0,00	0,02
6	4	18,60	0,19	0,11	1,53	0,01	0,00	0,02
7	4	18,60	0,17	0,11	1,31	0,01	0,00	0,02
8	4	18,60	0,14	0,11	1,15	0,01	0,00	0,01
9	4	18,60	0,13	0,11	1,02	0,01	0,00	0,01
10	4	18,60	0,12	0,11	0,92	0,01	0,00	0,01
4	5	18,60	0,29	0,09	2,30	0,01	0,00	0,02
5	5	18,60	0,23	0,09	1,84	0,01	0,00	0,02
6	5	18,60	0,19	0,09	1,53	0,01	0,00	0,02
7	5	18,60	0,17	0,09	1,31	0,01	0,00	0,01
8	5	18,60	0,14	0,09	1,15	0,01	0,00	0,01
9	5	18,60	0,13	0,09	1,02	0,01	0,00	0,01
10	5	18,60	0,12	0,09	0,92	0,01	0,00	0,01

Note1: *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).*

Note2. *Relative error ($\sigma^2\delta$) and absolute error ($\sigma^2\Delta$) values are scaled according to the number of items (i) and raters (r): $\sigma^2\delta = \sigma^2(pi)/i + \sigma^2(pr)/r + \sigma^2(pir)/ir$.*

Table 3 offers a detailed examination of measurement reliability by evaluating the impact of varying the number of items (d_i) and raters (d_r) on the distribution of variance components. The findings suggest the sources of error and point to design strategies that may help minimize their impact. Specifically, both item (i) and rater (r) variances systematically decreased as d_iand d_r increased. For instance, item variance declined from 0.29 at d_i= 4 to 0.12 at d_i= 10, while rater variance dropped from 0.46 at d_r = 1 to 0.09 at d_r = 5. This trend indicates that employing a broader item pool and involving multiple raters enhances the absolute reliability of the measurement by reducing item- and rater-related error.

In terms of interaction effects, the person × item (p × i) variance also exhibited a notable decline with increasing item count (e.g., 2.30 at d_i= 4 to 0.92 at d_i= 10). This suggests that a greater number of items can mitigate relative error by balancing individual performance across different test items. Meanwhile, person × rater (p × r) and item × rater (i × r) interaction variances remained low and stable (e.g., p × r declined from 0.05 to 0.01, and i × r remained at or near 0.00), reinforcing the minimal influence of rater inconsistencies.

The residual variance component (p × i × r), which captures random error not explained by the modeled facets, also decreased with an increase in both item and rater counts for example, from 0.12 at d_i= 4 and d_r = 1 to 0.01 at d_i= 10 and d_r = 5. This reduction implies that as the design becomes more comprehensive, the influence of unmodeled random error diminishes, thereby improving the generalizability ($E\rho^2$) and Phi (Φ) coefficients.

Overall, increasing the number of items (to ≥ 8) and raters (to ≥ 3) substantially reduces both relative (σ^2_δ) and absolute (σ^2_ϕ) error variances, optimizing decision reliability (Φ). Within the framework of Generalizability Theory, these results demonstrate the significant contributions of expanded item coverage and the use of multiple raters to measurement accuracy, offering practical guidance for optimizing future assessment designs.

Table 4.

Reliability and Phi Values According to Different Item and Rater Numbers

Number of Items	Number of Raters	Relative Error Variance ($\sigma^2(\delta)$)	Generalizability Coefficient ($E\rho^2$)	Absolute Error Variance ($\sigma2\ (\Delta)$)	Phi Coefficient (Φ)
4	1	0,45	0,81	2,43	0,81
4	**2**	**0,31**	**0,87**	**2,29**	**0,86**
4	3	0,26	0,88	2,24	0,88
4	4	0,24	0,89	2,22	0,89
4	5	0,22	0,9	2,2	0,9
5	1	0,38	0,84	2,36	0,84
5	2	0,26	0,89	2,24	0,88
5	3	0,21	0,9	2,19	0,9
5	4	0,19	0,91	2,17	0,91
5	5	0,18	0,92	2,16	0,92
6	1	0,33	0,86	2,31	0,86
6	2	0,22	0,9	2,2	0,9
6	3	0,18	0,92	2,16	0,92
6	4	0,16	0,92	2,14	0,93
6	5	0,15	0,93	2,13	0,93
7	1	0,29	0,87	2,27	0,87
7	2	0,19	0,91	2,17	0,91

7	3	0,16	0,93	2,14	0,93
7	4	0,14	0,93	2,12	0,93
7	5	0,13	0,94	2,11	0,94
8	1	0,27	0,88	2,25	0,88
8	2	0,17	0,92	2,15	0,92
8	3	0,14	0,93	2,12	0,93
8	4	0,13	0,94	2,11	0,94
8	5	0,12	0,94	2,1	0,94
9	1	0,25	0,89	2,23	0,89
9	2	0,16	0,93	2,14	0,93
9	3	0,13	0,94	2,11	0,94
9	4	0,12	0,94	2,1	0,94
9	**5**	**0,11**	**0,95**	**2,09**	**0,95**
10	1	0,23	0,9	2,21	0,9
10	2	0,15	0,93	2,13	0,93
10	3	0,12	0,94	2,1	0,94
10	**4**	**0,11**	**0,95**	**2,09**	**0,95**
10	5	0,10	0,95	2,08	0,95

The findings in Table 4 demonstrate notable changes in both relative and absolute error variances, as well as reliability coefficients, as the number of items and raters increases. For instance, in the condition with four items and one rater, the relative error variance ($\sigma^2(\delta)$) was calculated as 0.45, and the absolute error variance ($\sigma^2(\Delta)$) as 2.43. These values gradually decreased to 0.10 and 2.08, respectively, when ten items and five raters were used. This reduction indicates that increasing the number of items and involving multiple raters enhances the precision of measurement by reducing non-person sources of error. Parallel to this trend, both the generalizability coefficient ($E\rho^2$) and the Phi coefficient (Φ) showed consistent increases across design conditions. While the observed design was 5 items and 2 raters, the D-study explored 4–10 items and 1–5 raters: for the observed design, $E\rho^2 \approx 0.89$ and $\Phi \approx 0.88$. These values surpass the commonly accepted threshold of 0.80, indicating that the open-ended assessment tool used in this study yields reliable measurements for both relative and absolute decisions.

A closer look at the table suggests that increasing the number of items has a more pronounced effect on reliability than increasing the number of raters. For example, increasing the item count from 4 to 10 with one rater raised the generalizability coefficient from 0.81 to 0.90. A similar increase was observed when the number of raters was raised from 1 to 5 while keeping the item count constant at four. This finding highlights that item sampling plays a more significant role in improving measurement reliability compared to rater sampling. Nonetheless, moderate increases in the number of raters also contribute to reliability, particularly in reducing both relative and absolute error variances. Adding a second or third rater resulted in substantial reductions in error components, suggesting that even a small increase in rater numbers can enhance score dependability. However, the marginal gains in reliability beyond three raters appear minimal, indicating diminishing returns and emphasizing the need for efficient resource allocation in test design.

In conclusion, these results support the notion that increasing the number of items should be prioritized to enhance the reliability of open-ended assessments. Incorporating at least two independent raters is also recommended, especially in high-stakes evaluation contexts, to minimize subjectivity and scoring bias. This dual strategy offers a practical balance for achieving robust reliability under resource constraints.

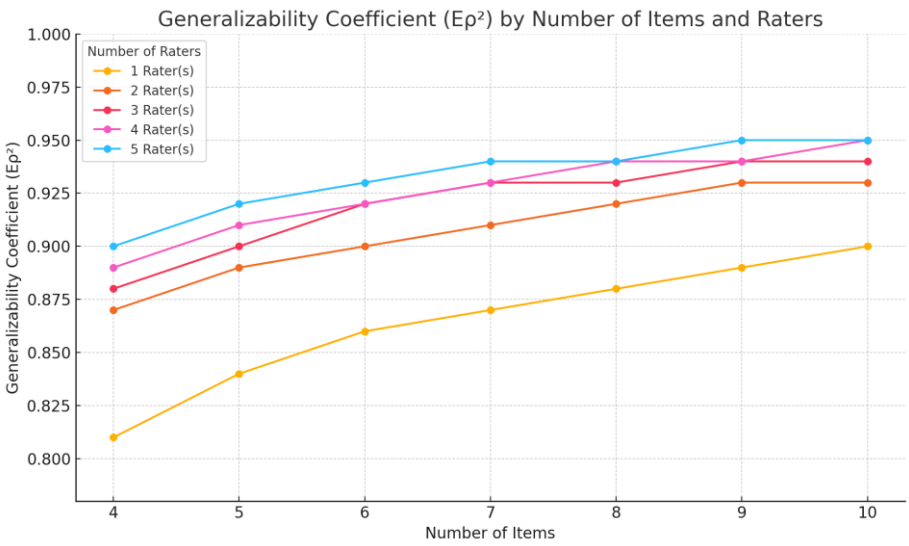

Figure 1. *Changes in Generalizability Coefficient (Eρ²) across different item and rater numbers.*

Figure 1 illustrates how the generalizability coefficient (Eρ²) appears to change with varying numbers of items and raters. The trend suggests that increasing the number of items may result in a sharper improvement. Notably, the coefficient approaches or exceeds .95 when item count is increased to 8 or more and at least three raters are used. These results reinforce the earlier findings from Table 4 and provide practical insight into the most effective ways to enhance exam reliability through item and rater design.

In conclusion, within the framework of Generalizability Theory, Table 4 demonstrates that the instrument achieves acceptable reliability of 0.81 with just four items and one rater, but increasing items and raters can boost both relative and absolute reliability up to 0.95. These findings highlight the importance of flexibly optimizing the number of items and assessors. Such optimization should depend on the intended purpose, whether ranking or decision-making.

CONCLUSION AND DISCUSSION

This study examined the reliability of written exams administered at universities using Generalizability Theory analysis based on a p × i × r design, where 76 students completed a five-item exam scored by two raters. The main findings indicate that the largest proportion of variance components is attributed to students (62.2%) and the student × item interaction (30.7%), while errors related to items (3.9%) and raters (1.5%) are relatively limited. The substantial variance among students suggests that the instrument effectively discriminates individual performance differences. However, the relatively large student × item interaction highlights the need to expand both the number and content of items to enhance content validity (Webb et al., 2006). Additionally, a brief content mapping matrix linking items to targeted learning outcomes was developed to support content validity claims. Since content validity and item content directly influence exam reliability, these aspects should be handled with meticulous attention (Güler, 2011; Monteiro et al., 2019).

This study is one of the few studies applying Generalizability Theory at the university level in Turkey and therefore provides an original contribution to the literature. Through the results of the G-study and D-study, evidence is presented regarding decision-making related to the structure of in-class written exams,

including the number of items, the number of evaluators, and the scoring processes. This establishes applicable reliability and validity principles to enhance measurement reliability.

The low rater variance is consistent with the use of detailed rubrics; however, in the absence of a formal training protocol reported here, this should be interpreted cautiously (Monteiro et al., 2019). Monteiro et al. (2019) demonstrated that the use of standardized rubrics in assessment processes significantly contributes to increasing inter-rater agreement.

The residual variance, which accounts for 1.6%, indicates that random errors have a minimal effect on the measurement results. In contrast, Smith and Paige (2019) reported a considerably higher error variance of 14.5 and emphasized the need for strategies to reduce the influence of uncontrollable factors on measurement reliability during exam processes. Similarly, Güler (2011), using simulated data with 125 students, 18 items, and 4 raters, underscored the importance of analyzing variance components with larger samples to obtain more robust reliability estimates.

Furthermore, when examining results across different items and raters, increasing the number of items and raters significantly reduced both the relative error variance ($\sigma^2(\delta)$) and absolute error variance ($\sigma^2(\Delta)$), while simultaneously boosting the generalizability coefficient ($E\rho^2$) and Phi coefficient (Φ) to 0.95. These reliability values exceed the commonly accepted threshold of 0.80 suggested by Brennan (2001), indicating that the instrument is reliable for both ranking and decision-making purposes. Nevertheless, the substantial student × item interaction highlights the importance of diversifying the question pool to further enhance reliability.

Based on these findings, under typical course decisions, we recommend ≥8 items and at least two independent raters; for higher-stakes decisions, consider three raters. This approach promotes an efficient use of limited resources while minimizing error variances. Additionally, controlling external factors such as session conditions and environmental variables in future studies may further reduce residual variance (Smith & Paige, 2019). The multi-faceted error decomposition enabled by Generalizability Theory allows university assessment units to review and improve exam designs in an evidence-based manner, thus facilitating a more valid and fair evaluation of students' actual performance (Güler, 2011). Beyond these practical implications, this study is among the few applications of Generalizability Theory conducted at the university level in Türkiye. By demonstrating how GT can inform item sampling, rater training, and exam design, the study not only contributes methodologically but also highlights the applied value of GT for improving fairness and reliability in higher education assessment practices.

LIMITATIONS AND FUTURE RESEARCH

Despite the strong design and detailed variance decomposition presented in this study, several limitations should be acknowledged:

Limited Number of Items and Raters: The exam consisted of only five items and was evaluated by two raters. While the fully crossed p × i × r design enables robust variance estimation, the small number of items may restrict the generalizability of findings. Future studies should replicate this design with larger item pools and more raters to validate the consistency of results.

Sample Size and Context: The study was conducted with 76 students enrolled in a single program at one public university. Broader samples across different disciplines, institutions, and academic levels would strengthen generalizability.

Focus on Written Exams Only: This study concentrated exclusively on open-ended written exams. Other types of performance-based assessments (e.g., oral exams, portfolios, or presentations) may exhibit different variance structures. Expanding the scope of future research to include diverse assessment formats would yield more comprehensive insights.

Environmental and Contextual Factors Not Controlled: Although residual variance was low (1.6%), uncontrolled situational factors such as exam room conditions, stress levels, or distractions may have affected student performance. Future studies could include these environmental variables as additional facets in the design.

Crossed Design Constraints: The fully crossed design used here, while analytically rigorous, is often difficult to implement in real educational settings due to time and logistical constraints. Nested or partially nested designs could be explored in future research to reflect more practical scoring conditions.

REFERENCES

Atılgan, H. (2005). A sample application for the theory of generalizability and inter-rater reliability. Educational Sciences and Practice, 4(7), 95–108.

Baker, F. B. (2001). The basics of item response theory (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Baykul, Y. (2021). Measurement in Education and Psychology: Classical Test Theory and Practice. Pegem Akademi Publishing.

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. Medical Teacher, 34(11), 960–992. https://doi.org/10.3109/0142159X.2012.703791.

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

Cardinet, J., Johnson, S. & Pini, G. (2009). Applying generalizability theory using EduG. Routledge. https://doi.org/10.4324/9780203866948.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Doğan, N. (Ed.). (2024). Measurement and evaluation in education (5th ed.). Pegem Academy Publishing.

Erkuş, A., Sünbül, Ö., Sünbül, S. Ö., Yormaz, S., & Aşiret, S. (2024). Measurement and scale development in psychology 2: Psychometric properties of measurement tools and measurement theories (3rd ed.). Pegem Academy Publishing.

Güler, N. (2011). Comparison of reliability according to the theory of generalizability on random data and classical test theory. Education and Science, 36(162), 225–234.

Huebner, A. & Lucht, M., (2019) “Generalizability Theory in R”, Practical Assessment, Research, and Evaluation 24(1): 5. https://doi.org/10.7275/5065-gc10.

Kara, Y. & Kelecioğlu, H. (2015). Examining the effect of raters' characteristics on determining cut-off scores using Generalizability Theory. Journal of Measurement and Evaluation in Education and Psychology, 6(1), 58–71.

Karasar, N. (2020). Scientific Research Method: Concepts, Principles, Techniques (36th Edition). Nobel Publishing.

Koğar, E. Y., & Koğar, H. (2022). Advanced psychometric applications with Mplus and R (1st ed.). Pegem Academy Publishing.

Koğar, H. (2023). Validity and reliability analyses with R: Applications of classical test theory, factor analysis approach, and item response theory (5th ed.). Pegem Academy Publishing.

Monteiro, S. Sullivan, G. M. & Chan, T. M. (2019). Generalizability theory made simple(r): An introductory primer to G-studies. Journal of Graduate Medical Education, 11(4), 365-370. https://doi.org/10.4300/JGME-D-19-00464.1.

Özçelik, D. A. (2016). Measurement and evaluation (5th edition). Pegem Academy Publishing.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Sage Publications. https://doi.org/10.1002/9781118445112.stat00068.

Smith, G. S. & Paige, D. D. (2019). A study of reliability across multiple raters when using the NAEP and MDFS rubrics to measure oral reading fluency. Reading Psychology, 40(1), 34-69. https://doi.org/10.1080/02702711.2018.1555361.

Turgut, M. F., & Baykul, Y. (2021). Eğitimde ölçme ve değerlendirme (9. baskı). Pegem Akademi.

VanLeeuwen, D. M., Dormody, T. J., & Seevers, B. S. (1999). Assessing The Reliability Of Student Evaluations Of Teaching (Sets) With Generalizability Theory. Journal of Agricultural Education, 40(4), 1–9. https://doi.org/10.5032/jae.1999.04001.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2017). Practical Applications of Generalizability Theory for Designing, Evaluating, and Improving Psychological Assessments. Journal of Personality Assessment, 100(1), 53–67. https://doi.org/10.1080/00223891.2017.1296455.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Using Generalizability Theory to Disattenuate Correlation Coefficients for Multiple Sources of Measurement Error. Multivariate Behavioral Research, 53(4), 481–501. https://doi.org/10.1080/00273171.2018.1457938

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. Handbook of Statistics, 26, 1-45. https://doi.org/10.1016/S0169-7161(06)26004-8.