

The Effect of Test Length, Sample Size, and Item Response Theory Models on Generalizability Theory

Priati Megawanti¹, Siti Alifah², Sulastry Pardede³, Erna Megawati⁴, Wardani Rahayu⁵

¹Department of Mathematic Education, Universitas Indraprasta PGRI, Jakarta, Indonesia

²Department of Industry Engineering, Universitas Indraprasta PGRI, Jakarta, Indonesia

³Department of Guidance and Counseling, Universitas Indraprasta PGRI, Jakarta, Indonesia

⁴Department of Indonesian Language and Literature Education, Universitas Indraprasta PGRI, Jakarta, Indonesia

⁵Department of Evaluation and Research in Education, Universitas Negeri Jakarta, Jakarta, Indonesia

*Email: priartimegawanti@gmail.com

ARTICLE INFO

Keywords:

Generalizability Theory

Item Response Theory

Test Length

Sample Size

EduG

ABSTRACT

Purpose - This study examined the effects of test length, sample size, and Item Response Theory (IRT) models on the reliability of Generalizability Theory (GT) coefficients.

Methodology - Simulated dichotomous and polytomous data were generated using WinGen and analyzed using EduG. The study employed several combinations of sample sizes ($N = 200, 300, 500, 1000, \text{ and } 5000$), test lengths ($n = 5, 10, 30, 50, \text{ and } 100$), dichotomous models (1PL, 2PL, and 3PL), and polytomous models (GRM and GPCM) with different response categories.

Findings - The findings indicate that test length had a stronger influence on G-coefficient reliability than sample size. In dichotomous data, the 1PL and 2PL models generally produced higher G-coefficients than the 3PL model, especially with a test length of 50 items. In polytomous data, the GPCM model produced higher and more stable G-coefficients than the GRM model, particularly when the number of response categories and test length increased. The results indicate that optimal combinations of test length, sample size, parametric model, and response categories are important for achieving high measurement reliability. This study also demonstrates the effectiveness of Generalizability Theory and EduG software in evaluating reliability for both dichotomous and polytomous data.

Contribution - This research helps researchers determine appropriate model parameters, sample size, and test length. Furthermore, it can strengthen theory in empirical research.

Received 11 June 2026; Received in revised form 31 July 2026; Accepted 03 September 2026

Jurnal Eduscience (JES) Volume 13 No. 5 (2026)

Available online 30 October 2026

©2026 The Author(s). Published by LPPM Universitas Labuhanbatu. This is an open-access article under the [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License \(CC BY - NC - SA 4.0\)](https://creativecommons.org/licenses/by-nc-sa/4.0/)

INTRODUCTION

Generalizability Theory (GT) was introduced by Cronbach and his colleagues (Brennan & Liao, 2020; Cronbach et al., 1972). GT began to be recognized in psychometrics as a refinement of Classical Test Theory (CTT). As is well known, in CTT, observed scores are a combination of true scores and errors (Cronbach et al., 1972; Eason, 1989; Irianty et al., 2025; Shavelson & Webb, 1991). CTT combines all measurement errors into a single, undifferentiated error term, making it impossible to examine the specific causes of variance (Parriott, 2017). As a result, researchers were never able to determine the true scores until Cronbach and his colleagues proposed GT as a "grand theory" of the reliability of behavioral measurements (Cardinet et al., 2010). Although GT and Item Response Theory (IRT) are both modern measurement approaches, they can help researchers identify error variances that are not distinguishable when using CTT calculations (Eason, 1989; Lakes & Hoyt, 2013). Unlike CTT, which is based on correlation, GT is based on ANOVA (Karami, 2012).

GT enables researchers to estimate sources of error variance arising from interactions among measurement facets. GT can identify both systematic and random errors (Susongko, 2015). GT can also differentiate between absolute and relative coefficient values in measurement results (Shavelson & Webb, 1991). Generalizability theory basically consists of G-Theory and D-Theory. G-Theory is used to estimate the magnitude of the inter-rater reliability coefficient in specific circumstances. The results of the G-Theory analysis are then used to analyze D-Theory (Fiangga & Sari, 2017; Irianty et al., 2025).

GT contains a key term, namely facets. These facets are crucial when measuring the G-Coefficient. Determining whether a facet is fixed or differentiated is crucial if researchers intend to use EduG software. There are four types of facets: differentiation, fixed, random, and stratification (Brennan, 2001). A facet is simply a collection of comparable measuring circumstances (Brennan, 2011). Any aspect of the measurement situation that might influence observed scores on the measure of interest and across which generalization of the observed score is required is referred to as a facet (White, 2017).

Matters related to generalizability analysis include the G-coefficient and D-study. Obtaining the G-coefficient and D-study results no longer requires manual calculations. Manual calculations can lead to errors in calculations and placement. Software is now available to assist researchers, particularly with calculations related to generalizability analysis. EduG is a tool created for generalizability analysis based on the Analysis of Variance (ANOVA) and Generalizability Theory (G theory) (Cardinet et al., 2010). EduG makes it easy for users to complete generalizability analyses of observed data. The entire software is built around user-friendly, typically self-explanatory displays that let the user specify the design's features (Clauser, 2008). As the creator of EduG, Cardinet stated in 2006 that EduG is a versatile software (Teket et al., 2015). One example of EduG's use in education is in essay-based exam assessments (Handayani et al., 2024). Subjective bias can occur during the scoring process. Even though raters have read the rubric and understood the scoring guidelines, in practice, one rater's assessment may differ from another's. Consequently, EduG can provide a G-coefficient value as a measure of reliability. However, researchers can also use EduG to further compare parameter stability or reliability coefficients across combinations of sample size, test length, and item-modeling choices. This represents a research gap left by previous studies that the current study aims to address.

IRT allows for dichotomous or polytomous test or exam results. Dichotomous, as the name suggests, means the resulting data contains only two values: 1 for correct and 0 for incorrect. Polytomous, on the other hand, allows for more than two values. In dichotomous models, the parameter models include 1 Parameter Logistic (PL), 2-Parameter Logistic (2PL), and 3-Parameter Logistic (3PL) (Hambleton et al., 1991). The 1PL (Rasch) model uses only the item difficulty (b) parameter (Boone & Noltemeyer, 2017; Susongko, 2016). The 2PL model uses the discrimination parameter (a) and difficulty (b) parameters. The 3PL model uses the discrimination parameter (a), difficulty (b), and pseudo-guessing (c) parameters. Item difficulty parameters ranged from -2 to +2, representing items from easy to difficult. The discrimination parameter is generated in the range of 0.5 to 2.0, while the guessing parameter in the 3PL model is generated in the range of 0.1 to 0.25.

Polytomous item response theory models can be applied to questions that assess cognitive, affective, and

psychomotor domains. There are also five types of scoring models in polytomous models: the Graded Response Model (GRM), Nominal Response Model (NRM), Partial Credit Model (PCM), Rating Scale Model (RSM), and Graded Partial Credit Model (GPCM) (Hambleton et al., 1991; Kartono, 2008; Ostini & Nering, 2006). In polytomous IRT models, the Partial Credit Model (PCM) is a special case of the Generalized Partial Credit Model (GPCM). In PCM, the discrimination parameter (a) is constrained to be constant or equal to 1, thereby extending the Rasch model to polytomous data. On the other hand, GPCM allows discrimination values to vary across items, extending the 2PL model to polytomous scoring (Kartono, 2008; Masters, 1982). In other words, the PCM model is a Rasch model, and 2PL is a special case of the GPCM model (Istiyono, 2016; Kartono, 2008; Ostini & Nering, 2006). The item discrimination parameter in PCM is constant at 1, whereas in GPCM, it varies (Istiyono, 2016; Ostini & Nering, 2006).

In development, dichotomous and polytomous models are no longer measured separately; they can now be assessed together in a single test. One example is the Minimum Competency Assessment (AKM), implemented since 2022, and the Academic Competency Test (TKA), implemented since the 2026 academic year. The AKM includes five types of tests, not all of which yield dichotomous scores (Fitriyah, 2023; Suliyanah et al., 2021). Meanwhile, in TKA, complex multiple-choice questions produce polytomous data (Nababan et al., 2025). Traditionally, assessment scoring schemes distinguished between item types that yielded dichotomous scores and those that yielded polytomous scores; dichotomous scoring was typically applied to multiple-choice and short-answer items, whereas polytomous scoring was used for essay questions. However, the AKM and TKA assessments have introduced new challenges by combining dichotomous and polytomous scoring within a single section of the test. Students encounter not only conventional multiple-choice items but also complex multiple-choice items—the former yielding dichotomous scores and the latter polytomous scores. This mixed-scoring format makes the reliability behavior of the GT coefficient—under conditions involving a mix of dichotomous and polytomous items—a practical issue that warrants urgent investigation. Consequently, this study aims to analyze the differences in G and D coefficients using simulated datasets containing both dichotomous and polytomous data. These datasets were processed using EduG and modified based on sample size (N), test length (n), and various other model parameters.

In measurement theory, ensuring the reliability of assessment instruments is a crucial yet complex task, particularly concerning error variance. As previously discussed, CTT provides a simple framework but lumps all measurement errors into a single, undifferentiated term. This makes it difficult for researchers to identify the precise sources of variance. GT addresses this limitation by using ANOVA to identify multiple error sources simultaneously. However, in empirical studies, researchers often face practical dilemmas regarding the required combination of model parameters, sample size, and test length. The appropriate combination needs to be explored to obtain stable and reliable G-coefficients. However, combinations such as integrating dichotomous (1PL, 2PL, 3PL) and polytomous (GRM and GPCM) parameter models remain underexplored under various simulation constraints. The primary research gap identified here is that existing Generalizability Theory (GT) studies tend to examine dichotomous and polytomous data separately. There is a lack of research systematically mapping how G and D coefficients behave when sample size and test length are adjusted to meet target reliability levels. Without a clear understanding of the outcomes resulting from various combinations of these factors, researchers risk collecting excessively large samples or designing assessment instruments that are unnecessarily long and inefficient. This issue is particularly pressing given that the AKM and TKA assessments combine the two scoring formats into a single test. Consequently, instrument developers require concrete, evidence-based guidance regarding the appropriate number of test items, sample size, and model parameters.

To address the above issues, this simulation study systematically examines the simultaneous effects of test length, sample size, and various IRT model parameters on the reliability of GT coefficients using a simulated dataset analyzed through EduG software. The novelty of this work lies in its comprehensive framework that compares dichotomous and polytomous parameters across a wide gradient of sample sizes (up to $N = 5000$) and test lengths (up to $n = 100$). By placing this research within a proper psychometric context,

this study builds on recent baseline studies that employ modern measurement approaches. For example, the generalizability analysis framework relies heavily on the versatility of the EduG software, as evaluated by Clauser (2008) and Teker et al. (2015). Furthermore, this study's methodological approach aligns with recent GT implementations in educational testing instruments, as demonstrated by the multifaceted validation frameworks reported by Handayani et al. (2024) and Irianty et al. (2025). To optimize instrument design, this study also draws on insights from Ngozika (2026), who emphasized the critical need to determine a minimum test length to ensure a reliable and efficient multiple-choice economics assessment. Through this structural comparison, the study provides a clear empirical benchmark to help researchers determine the most efficient sample size and item configuration. Its theoretical contribution lies in extending Generalizability Theory (GT) literature by explaining how coefficient stability improves with specific combinations. From a practical standpoint, the study offers a reference for test developers and policymakers to design assessments using appropriate item counts, sample sizes, and model parameters.

METHODOLOGY

Research Design

This research is a simulation study using data generated by WinGen and processed with EduG software. WinGen is software developed by Kyung T. Han and Ronald K. Hambleton (Han & Amherst, 2007). EduG is software programmed by Cardinet and colleagues to obtain G and D coefficient values for the study (Clauser, 2008). The simulated data in this study were generated using a normal distribution with a mean (μ) of 0 and a standard deviation (σ) of 1. The normal distribution was chosen because, in item response theory (IRT), examinees' ability (ability parameter/ θ) is generally assumed to follow a standard normal distribution. With this assumption, the simulated data can proportionally represent the variation in examinees' ability, from low to high. The standard normal distribution was used to generate participant ability parameters (θ) to approximate empirical conditions in educational and psychological measurement.

Data Collection

The generated data consisted of several data combinations, both dichotomous and polytomous. For the dichotomous data, all 1PL, 2PL, and 3PL models were selected. For the polytomous data, only two models were selected: GRM and GPCM. The GRM and GPCM were selected because they are widely applied in educational and psychological assessment. The GRM is used for graded data such as attitude scales or graded assessments. This model usually takes the form of a Likert scale. Meanwhile, the GPCM is used for partial responses to essay questions or step-by-step scoring. This model is typically used for essay questions or practical tests in schools. The number of response categories used in simulations is 5, 7, or 10.

Data Analysis

This study used simulated data generated with WinGen and then processed with EduG. Several combinations were used to determine the G-coefficient value and good D-study results. The first combination is a fixed sample size (N) and test length (n), but the model parameters are modified. The second combination is a fixed sample size (N) and model parameters, while the test length (n) is varied. The third combination is a fixed test length (n) and model parameters, with the sample size (N) modified. Table 1 shows the aspects and combinations used in this study.

Table 1. Data Combination Design

Aspects	Combinations
Sample Size (N)	200, 300, 500, 1000, 5000
Test Length (n)	5, 10, 30, 50, 100
Dichotomous Model	1PL, 2PL, 3PL
Polytomous Model	GRM, GPCM
Number of Categories	5, 7, 10

FINDINGS

First Combination (N and n Fixed, Model Parameters Modified)

This combination consists of a sample size (N) of 200 and a test length (n) of 10, which are fixed, while the model parameters are varied across the 1PL, 2PL, and 3PL models. Table 2 shows that the highest relative G-coefficient is in the combination of N=200, n=10, and 1PL. Meanwhile, the highest absolute G-coefficient is obtained with N=200, n=10, and 2PL.

Table 2. Result of G Study Dichotomous Model for N=200 and n=10

	Relative G-coefficient	Absolute G-coefficient
N=200; n=10; 1PL	0.63	0.54
N=200; n=10; 2PL	0.62	0.56
N=200; n=10; 3PL	0.53	0.48

Table 2 shows that the relative and absolute G-coefficient values for N=200 and n=10 did not reach 0.7, even though the model parameters were changed from 1PL to 2PL and then to 3PL. This indicates that a sample size of 200 and a test length of 10 items are insufficient to yield a G-coefficient of 0.7 or higher.

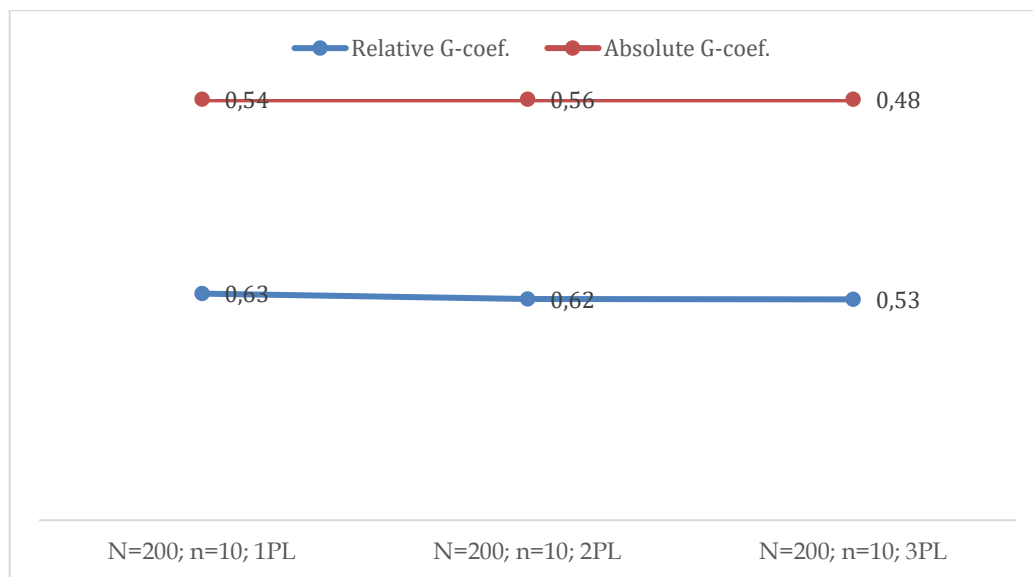


Figure 1. Graph of G-relative and G-absolute Coefficient in Dichotomous Model Modification

Meanwhile, for the results of the D-study with a sample size of 200, with n' consisting of 5, 8, 12, 15, and 20. The test length is 10, and the parameter model is 1PL, where n' is the number of measurement conditions in the desired design (decision study/D-study). The specific test lengths for the D-study (n' = 5, 8, 12, 15, and 20) were selected to evaluate the impact of both shortening and lengthening the baseline instrument (n = 10). This incremental range allows precise identification of the optimal point – minimizing measurement error and ensuring practical test efficiency while identifying the threshold at which the visual G-coefficient meets the acceptable psychometric benchmark (≥ 0.70). For more details, see Figure 2 below Table 3.

Table 3. Result D-study Dichotomous Model 1PL for N=200 and n=10

Result D-study N=200 n 10 1PL					
n'	5	8	12	15	20
Relative G-coefficient	0.4577	0.57454	0.66948	0.71687	0.77148
Absolute G-coefficient	0.37428	0.48902	0.58942	0.64215	0.70524
Relative D-coefficient	0.03252	0.02033	0.01355	0.01084	0.00813

Figure 2 compares four coefficients: the relative G-coefficient, the absolute G-coefficient, the relative D-coefficient, and the absolute D-coefficient. The top two lines—representing the relative and absolute G-coefficients—show a continuous upward trend as values on the horizontal axis increase. In contrast, the lines for the relative and absolute D-coefficients remain much lower, staying relatively stable near zero. This striking difference indicates that G-coefficients are far more sensitive to variable changes than D-coefficients. The G-coefficient values in the upper section represent measurement reliability, whereas the lower section of the graph shows the magnitude of error variance (both relative and absolute). This distinction accounts for the observed differences between the G-coefficient and D-coefficient plots.

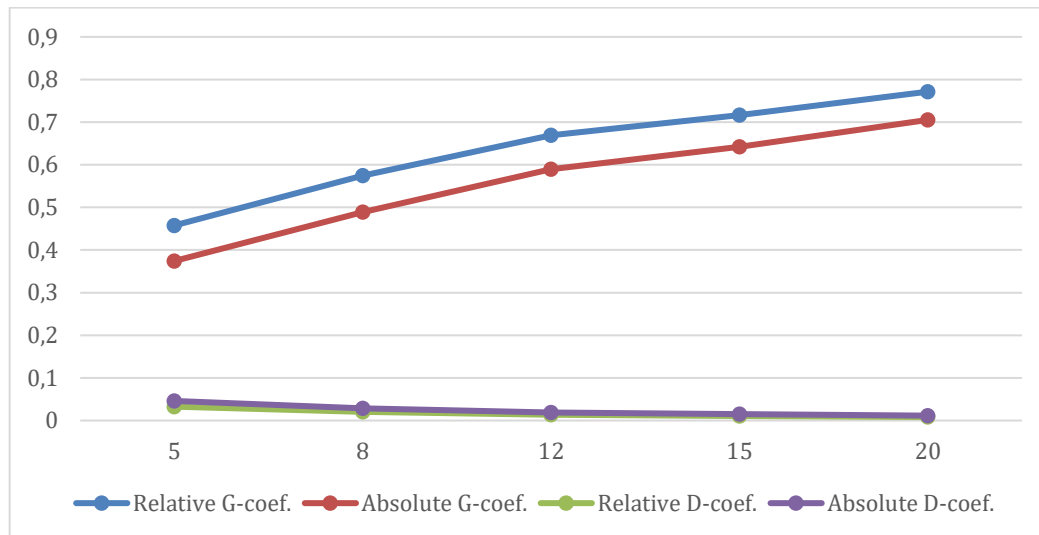


Figure 2. Graph of D-study Dichotomous Model 1PL for N=200 and n=10

For a sample size of 200, a test length of 10, and a 2PL parameter model, the results of the D-study are shown in Table 4 and Figure 3.

Table 4. Result of D-study Dichotomous Model 2PL for N=200 and n=10

Result D-study N=200 n 10 2PL					
n'	5	8	12	15	20
Relative G-coefficient	0.4537	0.57059	0.66591	0.71359	0.76862
Absolute G-coefficient	0.38942	0.50507	0.60486	0.65676	0.7184
Relative D-coefficient	0.03474	0.02171	0.01447	0.01158	0.00868
Absolute D-coefficient	0.04523	0.02827	0.01885	0.01508	0.01131

Table 4 and Figure 3 show the results regarding the impact of increasing the number of items (n') on the instrument's coefficient values. As n' increases from 5 to 20, both the Relative G-coefficient and the Absolute G-coefficient show a consistent rise, moving from the 0.38–0.45 range to 0.71–0.76. Conversely, the Relative D-coefficient and Absolute D-coefficient values steadily decline, approaching the 0.008–0.011 range. A stark contrast is evident in the scale of these values: the G-coefficients rise into the moderate-to-high range, whereas the D-coefficients—representing the instrument's error variance—remain very low and continue to decrease toward zero. This demonstrates that increasing the number of items effectively reduces error drastically while simultaneously enhancing the reliability coefficient.

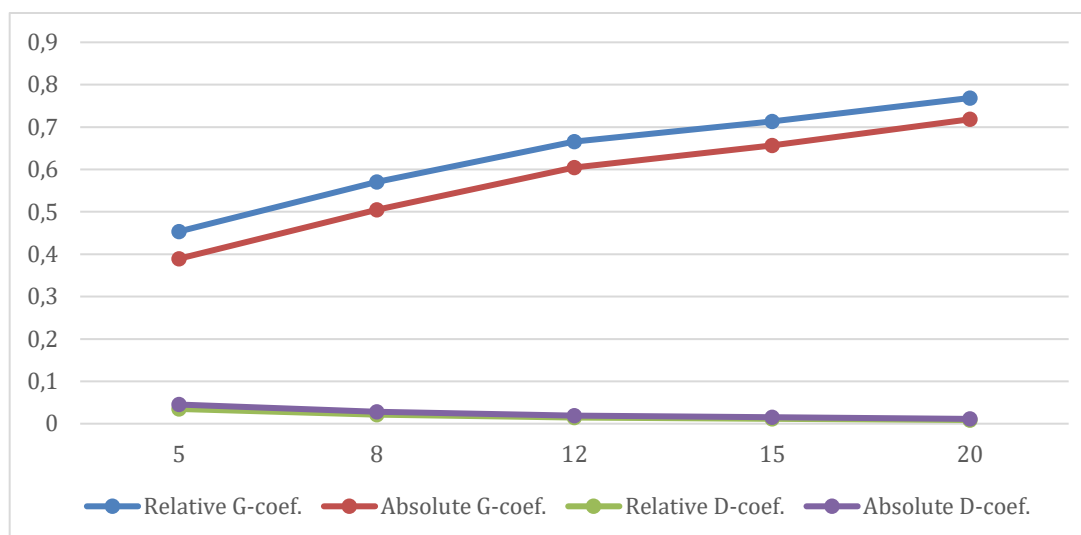


Figure 3. Graph of D-study Dichotomous Model 2PL for N=200 and n=10

Table 5 and Figure 4 show the results of the D-study for a sample size of 200, a test length of 10, and a 3PL parameter model, with $n' = 5, 8, 12, 15$, and 20.

Table 5. Result of D-study Dichotomous Model 3PL for N=200 and n=10

Result D-study N=200 n 10 3PL					
n'	5	8	12	15	20
Relative G-coefficient	0.36106	0.47483	0.57559	0.62898	0.69328
Absolute G-coefficient	0.3143	0.42309	0.52383	0.57896	0.64707
Relative D-coefficient	0.03778	0.02361	0.01574	0.01259	0.00945
Absolute D-coefficient	0.04658	0.02911	0.01941	0.01553	0.01164

Table 5 and Figure 4 demonstrate that increasing the number of items (n') raises the reliability coefficient while simultaneously reducing error. As n' increases from 5 to 20, the Relative G-coefficient rises from 0.36106 to 0.69328, and the Absolute G-coefficient climbs sharply from 0.3143 to 0.64707. Conversely, the error variance values (Relative and Absolute D-coefficients) steadily decline, reaching 0.00945 and 0.01164, respectively. This confirms that increasing the number of items in the 3PL model effectively minimizes measurement error and enhances the reliability of test results.

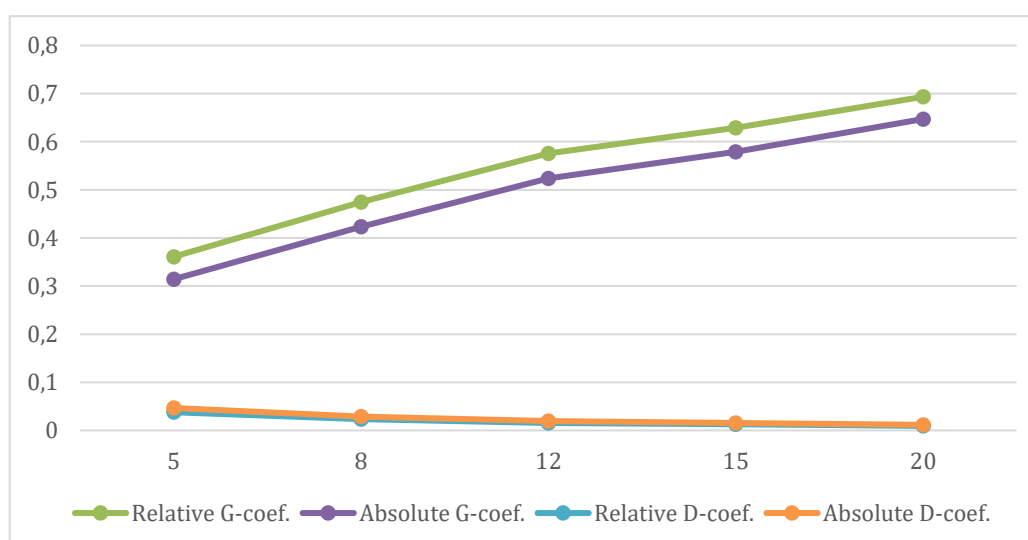


Figure 4. Graph D-study Dichotomous Model 3PL for N=200 and n=10

Second Combination (N and Model Parameters are Fixed, n Modified)

The next dichotomous modification combines a fixed sample size and model parameters with a modified test length. In Table 6, the sample size remains at $N=200$, and the model parameters remain at 3PL. However, the test lengths are modified to 10, 30, and 50. Based on the EduG processing results, absolute and relative G-coefficient values are obtained. The table shows that the highest relative and absolute G-coefficient values are at $N=200$, $n=50$, and 3PL. This indicates that test length significantly influences the G-coefficient value. The larger the test length, the larger the G-coefficient values, both relative and absolute. If the sample size remains at 200 and the test length (n) is 30 or 50, with the model parameters set to 1PL, 2PL, and 3PL, respectively, the relative and absolute G-coefficient values are shown in Table 6 and Figure 5.

Table 6. Result of G-Study Dichotomous Model for $N=200$, $n=10$; 30; 50

	Relative G-coefficient	Absolute G-coefficient
$N=200$; $n=10$; 1PL	0.63	0.54
$N=200$; $n=10$; 2PL	0.62	0.56
$N=200$; $n=10$; 3PL	0.53	0.48
	Relative G-coefficient	Absolute G-coefficient
$N=200$; $n=30$; 1PL	0.84	0.79
$N=200$; $n=30$; 2PL	0.84	0.80
$N=200$; $n=30$; 3PL	0.80	0.76
	Relative G-coefficient	Absolute G-coefficient
$N=200$; $n=50$; 1PL	0.90	0.88
$N=200$; $n=50$; 2PL	0.90	0.88
$N=200$; $n=50$; 3PL	0.88	0.85

Based on Table 6 and Figure 5, the highest relative and absolute G-coefficient values are at a test length of 50. The 1PL and 2PL parameter models are not too different, although in the EduG output results, the 1PL model is slightly larger than the 2PL model. Thus, it can be temporarily concluded that for $N = 200$, the most suitable test length for questions with Dichotomous scores is 50. The findings suggest that the 1PL or 2PL parameter model yields the highest G-coefficient values, both relative and absolute.

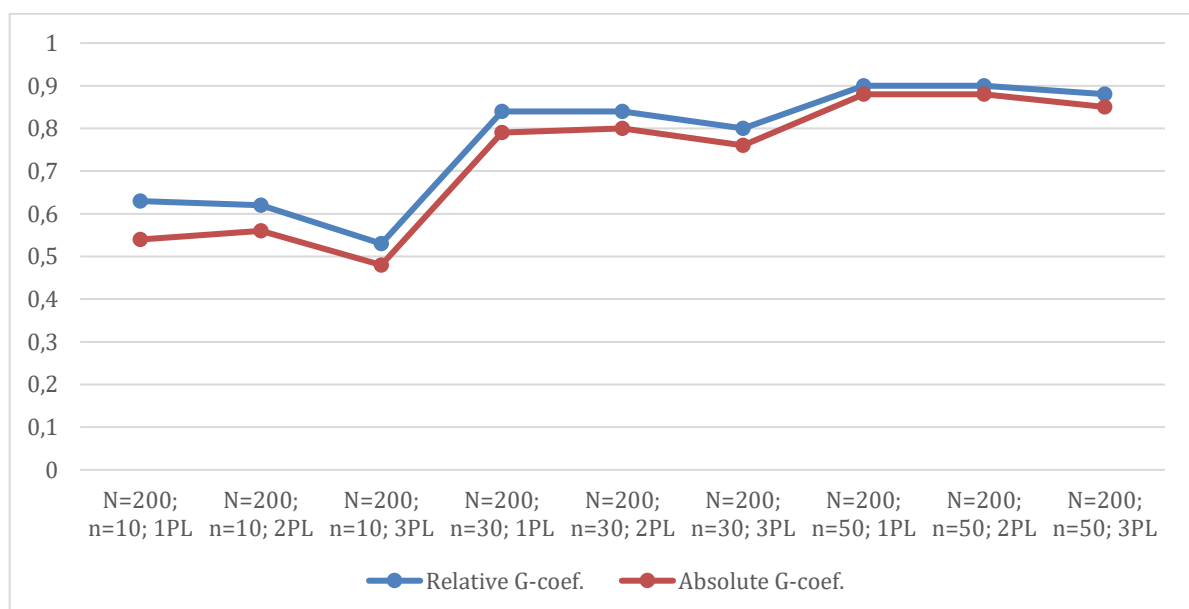


Figure 5. Graph of Relative G-coefficient and Absolute G-coefficient in Test Length Modification

Third Combination (n and Model Parameters are Fixed, N Modified)

Meanwhile, in the third experiment, the sample size was varied while the test length and model parameters were held fixed. The test length was chosen as 50 because the second experiment obtained the largest relative and absolute G-coefficient values at this test length. Then, the modified sample size was tested with $n=50$ and the 1PL, 2PL, and 3PL model parameters to determine which combinations yielded the highest relative and absolute G-coefficient values (Table 7).

Table 7. Result of G-Study Dichotomous Model for $n=50$ with Sample Size and Model Parameter Modification

	Relative G-coefficient	Absolute G-coefficient
N=200; $n=50$; 1PL	0.90	0.88
N=300; $n=50$; 1PL	0.90	0.88
N=500; $n=50$; 1PL	0.90	0.88
N=1000; $n=50$; 1PL	0.89	0.86
N=5000; $n=50$; 1PL	0.90	0.88
	Relative G-coefficient	Absolute G-coefficient
N=200; $n=50$; 2PL	0.90	0.88
N=300; $n=50$; 2PL	0.90	0.88
N=500; $n=50$; 2PL	0.91	0.88
N=1000; $n=50$; 2PL	0.92	0.90
N=5000; $n=50$; 2PL	0.91	0.88
	Relative G-coefficient	Absolute G-coefficient
N=200; $n=50$; 3PL	0.88	0.85
N=300; $n=50$; 3PL	0.88	0.85
N=500; $n=50$; 3PL	0.88	0.85
N=1000; $n=50$; 3PL	0.87	0.85
N=5000; $n=50$; 3PL	0.87	0.84

Table 7 above shows that the relative and absolute G-coefficient values for each combination of parameter models (1PL, 2PL, and 3PL) are not significantly different. For the 1PL parameter model, the highest G-coefficient values were 0.90 and 0.88 for sample sizes of 200, 300, 500, and 5000, respectively. This means that if a researcher were to use a sample size of 200 for a test length of 50, the results would not differ significantly from those of a sample size of 5000. Meanwhile, for the 2PL parameter model, the highest G-coefficient was achieved with a sample size of 1000 and a test length of 50 items. Compared with sample sizes of 200, 300, 500, and 5000, the sample size of 1000 yielded relative and absolute G-coefficient values of 0.92 and 0.90, respectively.

Finally, in the 3PL parameter model, sample sizes of 200, 300, and 500 have the highest relative and absolute G-coefficient values compared to sample sizes of 1000 and 5000. The resulting relative and absolute G-coefficient values were 0.88 and 0.85, respectively. Therefore, it can be concluded that, in dichotomous data, the best combination for obtaining the highest relative and absolute G-coefficient values in the 1PL for a test length of 50 is sample sizes of 200, 300, 500, and 5000. For the 2PL parameter model with a test length of 50, the sample size that produced the highest relative and absolute G-coefficient values was 1000. Furthermore, for the 3PL parameter model and a test length of 50, the highest relative and absolute G-coefficient values were observed at sample sizes of 200, 300, and 500.

Figure 6 shows that despite varying sample sizes, the absolute and relative G-coefficients tend to be stable. Empirically, there were no significant changes when the sample size was varied from 200, 300, 500, 1000, and 5000. The relative and absolute G-coefficient values remained stable at sample sizes of 200, 300, and 500. Meanwhile, the relative and absolute G-coefficient values decreased by 0.01 at sample sizes of 1000 and 5000. This indicates that there is no significant difference between sample sizes $N=200$, 300, and 500. In practice, researchers can choose $N=200$ for a test length of 50 and a 3PL parameter model to yield the same G-coefficient

as $N=500$. Next, we will discuss polytomous data: the GRM and GPCM. The distribution used is the normal distribution.

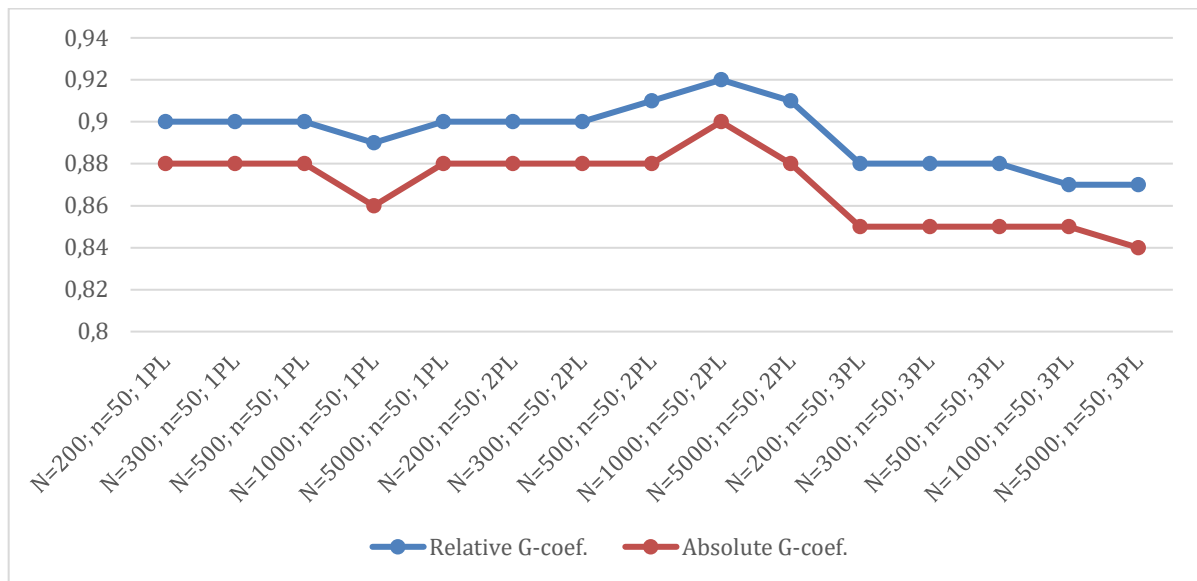


Figure 6. Graph of Relative G-coefficient and Absolute G-coefficient in Sample Size Modification

Polytomous with Normal Distribution

In polytomous models, simulated data uses several combinations. These combinations are used to determine the high G-coefficient characteristic of polytomous response models (GRM and GPCM). The first part uses the GRM model. The first combination holds the sample size (N) and test length (n) constant, while varying the number of categories. The second combination holds the sample size (N) and model parameters constant, while varying the test length (n). The third combination holds the test length (n) and the model parameters constant, while varying the sample size (N). The detailed G study results for the simulated GRM, along with their modifications and combinations, are as follows.

First Combination of the GRM Model (N and n are Fixed, Number of Categories is Modified)

In combination with the modified number of categories, the best G-coefficient value was obtained at $N=200$ and $n=5$ with 10 categories. Meanwhile, the lowest coefficient value was obtained at 7 categories. Thus, it can be concluded that if researchers wish to use a GRM, the appropriate combination to obtain a coefficient value above 0.7 is $N=200$, $n=5$, and a 10-category.

Table 8. Result of G-Study Polytomous with Sample Size (N)=200, Test Length (n)=5, and Parameter Model GRM, with Categories Modification

	Relative G-coefficient	Absolute G-coefficient
$N=200$; $n=5$; 5-category, GRM	0.51	0.51
$N=200$; $n=5$; 7-category, GRM	0.5	0.49
$N=200$; $n=5$; 10-category, GRM	0.74	0.72

Meanwhile, for $N=200$ and $n=5$, the 10-category has the highest G-coefficient compared to the 5- and 7-category responses. The graph is shown in Figure 8 below. However, the highest value is only around 0.7.

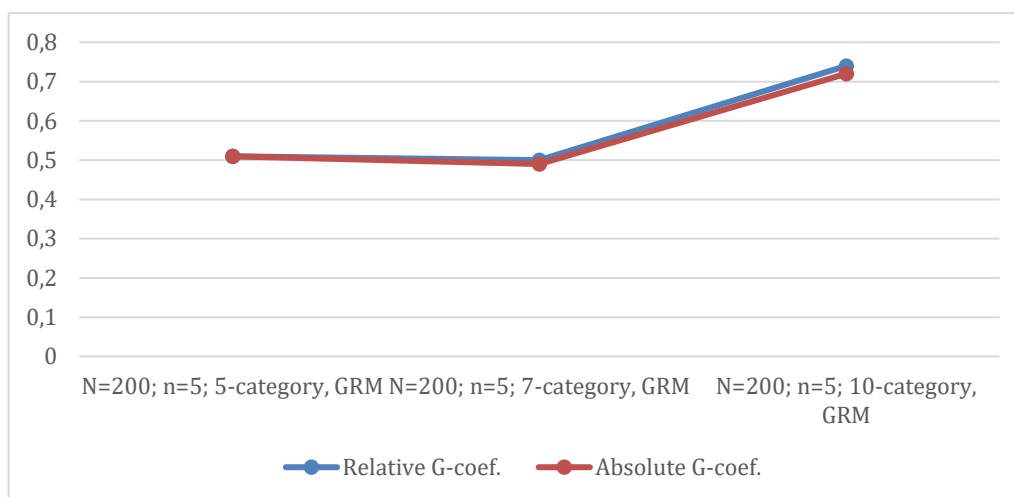


Figure 7. Graph of Relative G-coefficient and Absolute G-coefficient with Response Categories modification for Model GRM

Second Combination in the GRM Model (N and Categories are Fixed, n Modified)

Below are several combinations that modify the test length. Based on Table 9, the relative and absolute G-coefficient values for almost all combinations are above 0.7. Only the combination of N=200, n=5, and 5 categories has a relative and absolute G-coefficient below 0.7, namely 0.51.

Table 9. Result of G-Study Polytomous GRM with Sample Size (N)=200, Categories, and Test Length (n) Modification

	Relative G-coefficient	Absolute G-coefficient
N=200; n=5; 5-category, GRM	0.51	0.51
N=200; n=10; 5-category, GRM	0.73	0.71
N=200; n=30; 5-category, GRM	0.89	0.89
N=200; n=50; 5-category, GRM	0.95	0.94
N=200; n=100; 5-category, GRM	0.96	0.96
	Relative G-coefficient	Absolute G-coefficient
N=200; n=5; 10-category, GRM	0.74	0.72
N=200; n=10; 10-category, GRM	0.80	0.78
N=200; n=30; 10-category, GRM	0.88	0.87
N=200; n=50; 10-category, GRM	0.94	0.94
N=200; n=100; 10-category, GRM	0.97	0.97

Based on Table 9 above, the largest G-coefficient is at N = 200, n = 100, and 5 categories. Meanwhile, for the 10-category case, the highest G-coefficient occurs when N = 200 and the test length is 100. In field data, the Polytomous score, which has a test length of 100 items, is a psychological test in which examinees only need to respond according to their perceived circumstances. However, if the cognitive test items require examinees to think about the questions in a complex manner, it is not recommended to use items with a test length of 100 – the graph of the stability of the relative and absolute G-coefficient values in Figure 8.

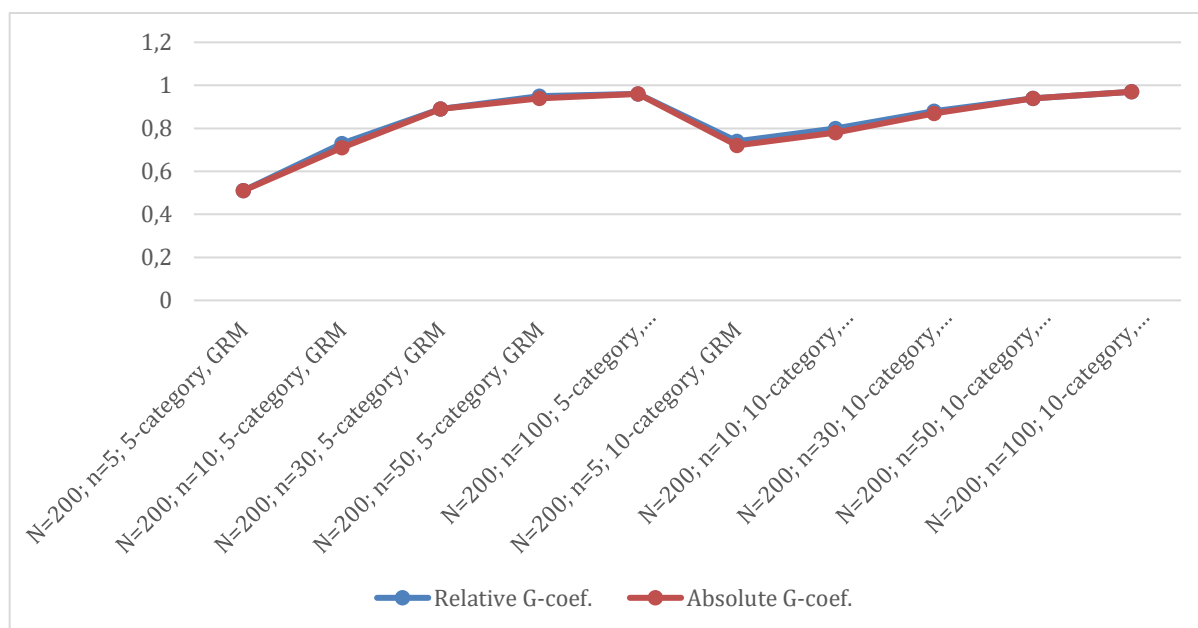


Figure 8. Graph of Relative G-coefficient and Absolute G-coefficient in Test Length Modification

Third Combination of the GRM Model (Number of Categories and n are Fixed, N Modified)

Table 10 and Figure 10 show that relative and absolute G-coefficient values exceeded 0.7 when the sample size was increased to 300. Thus, the number of examinees for whom the relative and absolute G-coefficient values of 0.7 was 300. Increasing the sample size to 500 actually reduced the relative and absolute G-coefficient values to 0.37 in the 5-category. Similarly, when the number of examinees was reduced to 200, the relative and absolute G-coefficient values dropped even further, to 0.51. The worst G-coefficient value was obtained when N, or the sample size, was 500.

Table 10. Result of G-study Polytomous GRM for Test Length =5, and 5; 10 Categories with Sample Size Modification

	Relative G-coefficient	Absolute G-coefficient
N=200; n=5; 5-category, GRM	0.51	0.51
N=300; n=5; 5-category, GRM	0.73	0.70
N=500; n=5; 5-category, GRM	0.37	0.37
N=1000; n=5; 5-category, GRM	0.63	0.59
N=5000; n=5; 5-category, GRM	0.59	0.54
	Relative G-coefficient	Absolute G-coefficient
N=200; n=5; 10-category, GRM	0.74	0.72
N=300; n=5; 10-category, GRM	0.71	0.70
N=500; n=5; 10-category, GRM	0.62	0.61
N=1000; n=5; 10-category, GRM	0.53	0.53
N=5000; n=5; 10-category, GRM	0.42	0.42

Based on Table 10, if the test length is 5 and the categories are 5 and 10, then the highest G-coefficient is at the sample size of 300. This indicates that the researcher can develop an instrument with GRM for test lengths of 5 or 10 and a sample size of 300 people to obtain a sufficient G-coefficient (Figure 9).

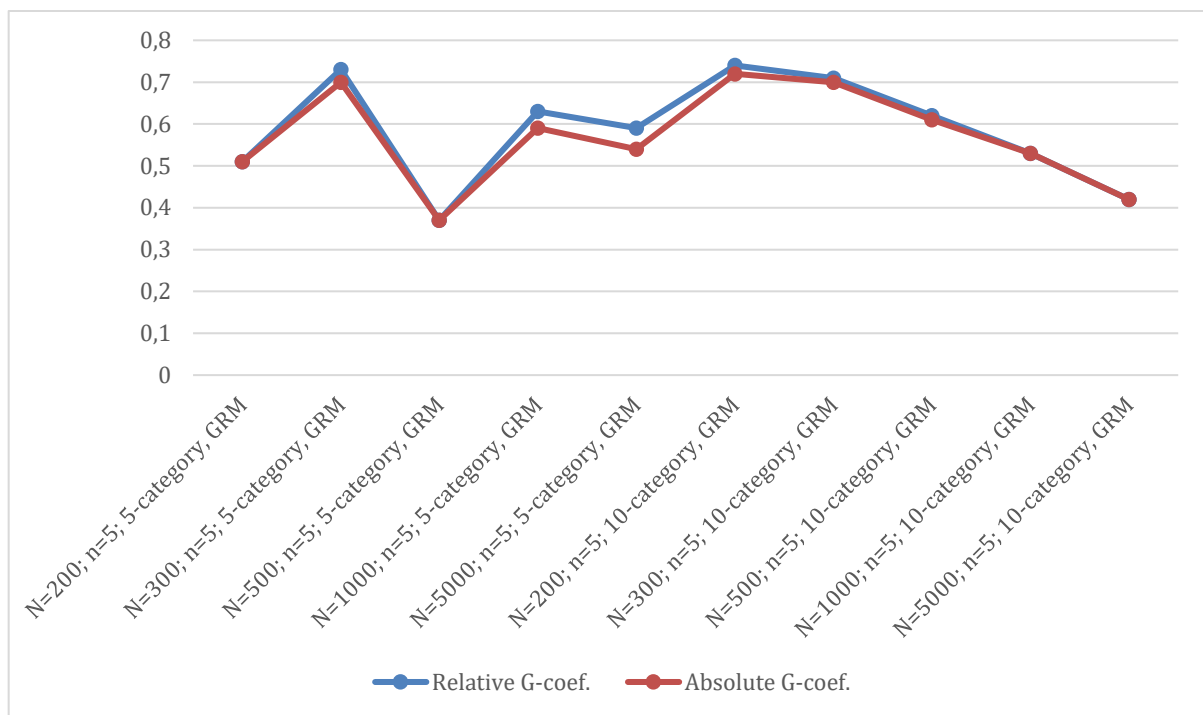


Figure 9. Graph of Relative and Absolute G-coefficient in Sample Size Modification

The next combination uses the GPCM model with the same combination as the previous GRM. For simplicity, the subpoint numbers will continue. The results of the G-study of the GPCM model, along with its combinations and modifications, are as follows.

First Combination in the GPCM Model (Fixed N and n, Modified Number of Categories)

Similar to the GRM model, Table 11 and Figure 10 show that when the sample size (N) and test length (n) are fixed, the 10-category model provides relative and absolute G-coefficient values above 0.8, namely 0.9 and 0.85. While the GRM with 7 categories yields the lowest G-coefficient, the GPCM 10-category model yields a higher G-coefficient than the 5-category model.

Table 11. Result of G-study Polytomous GPCM for Sample Size (N)=200 and Test Length (n)=5 with Categories Modification

	Relative G-coefficient	Absolute G-coefficient
N=200; n=5; 5-category, GPCM	0.73	0.72
N=200; n=5; 7-category, GPCM	0.86	0.81
N=200; n=5; 10-category, GPCM	0.90	0.85

Table 11 and Figure 10 demonstrate the impact of increasing the number of response categories on instrument reliability, while keeping sample size (N=200) and the number of items (n=5) constant. Both the table and the graph reveal that the Relative G-coefficient rises from 0.73 (5 categories) to 0.86 (7 categories), then reaches 0.90 (10 categories). Meanwhile, the absolute G-coefficient increases from 0.72 to 0.81, then to 0.85. The line visualizations in Figure 10 highlight that the relative coefficient curve consistently remains slightly above the absolute coefficient curve across all category levels. Overall, the combined data from the table and graph confirm that increasing the number of response categories in the GPCM model consistently and significantly enhances the measurement instrument's reliability.

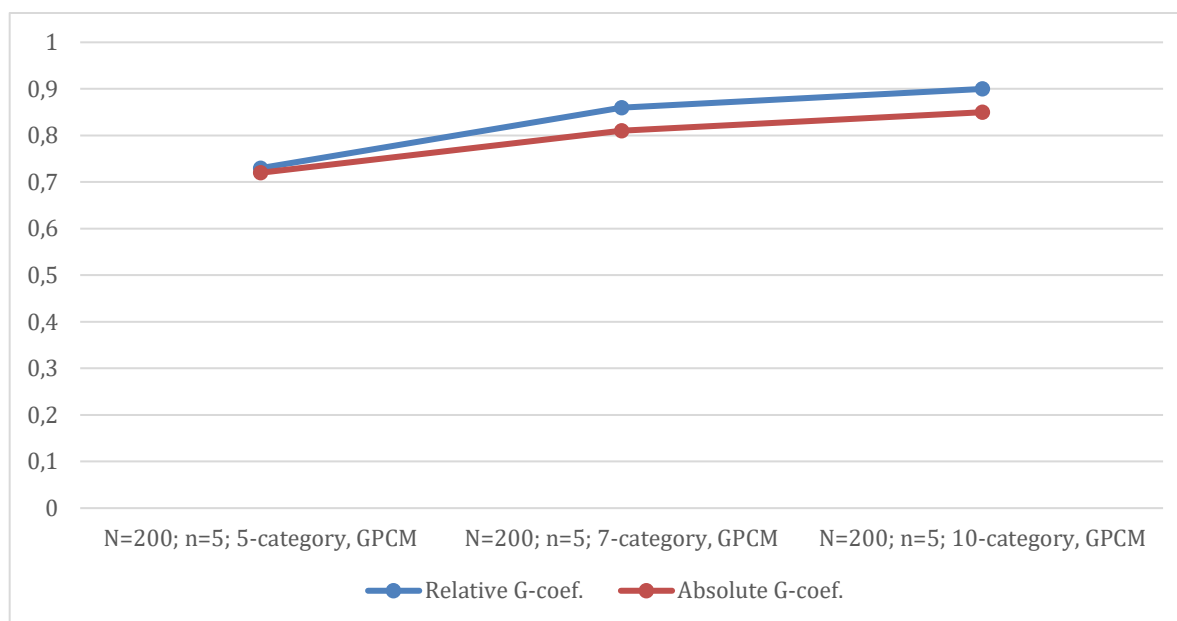


Figure 10. Graph of Relative and Absolute G-Coefficients for Modified Number of Categories

Second Combination in the GPCM Model (*N* and number of categories fixed, *n* modified)

In the GPCM model, the highest G-coefficient was obtained when the sample size was 200, the test length was 50, and the 5-category. Table 12 shows that a sample size of 200 and a modified test length produced relative and absolute G-coefficient values above 0.7.

Table 12. Results of the GPCM Polytomous G-Study with Sample Size (*N*) = 200, and Categories of 5 and 10, with Modified Test Length (*n*)

	Relative G-coefficient	Absolute G-coefficient
N=200; n=5; 5-category, GPCM	0.73	0.72
N=200; n=10; 5-category, GPCM	0.88	0.84
N=200; n=30; 5-category, GPCM	0.95	0.94
N=200; n=50; 5-category, GPCM	0.98	0.97
	Relative G-coefficient	Absolute G-coefficient
N=200; n=5; 10-category, GPCM	0.90	0.85
N=200; n=10; 10-category, GPCM	0.94	0.91
N=200; n=30; 10-category, GPCM	0.98	0.97
N=200; n=50; 10-category, GPCM	0.99	0.98

Table 12 above shows that when *N* = 200 and the number of categories is 5 or 10, the highest G-coefficient values in the GPCM model are obtained for a test length of 50. This can serve as a reference for researchers who want to create questions for the GPCM: a 10-category, 50-item test with a sample size of 200 yields the highest G-coefficient among other combinations. However, a test length of 30, a sample size of 200, and response categories of 5 or 10 can provide relative and absolute G-coefficient values that are also worth considering (Figure 11).

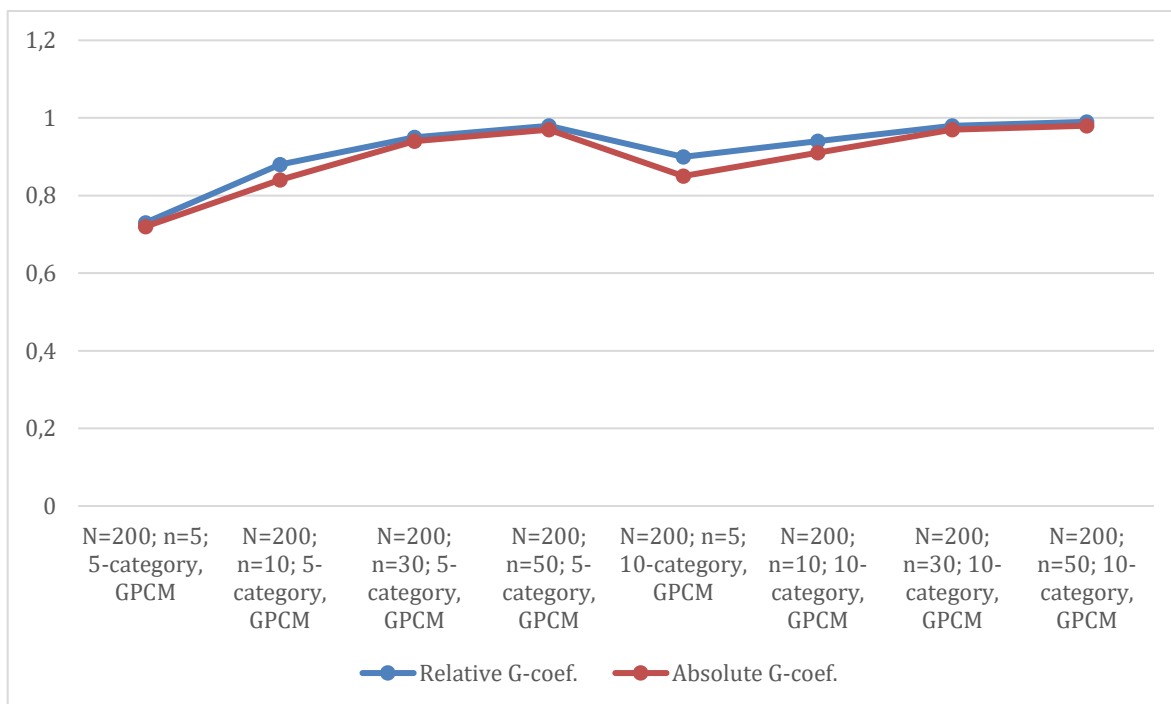


Figure 11. Graph of Relative and Absolute G-coefficient in Test Length Modification

Combination of the GPCM Model (Fixed Number of Categories and N, Modified N)

If the sample size is modified, the highest G-coefficient is obtained with a sample size of 1000, a test length of 5 items, and 5 categories. The G-coefficient values for each combination are shown in Table 13 and Figure 12.

Table 13. Results of the G Study for the GPCM Polytomous with Test Length (n) = 5 and Modified Sample Sizes (N) of 5 and 10

	Relative G-coefficient	Absolute G-coefficient
N=200; n=5; 5-category, GPCM	0.73	0.72
N=300; n=5; 5-category, GPCM	0.76	0.72
N=500; n=5; 5-category, GPCM	0.75	0.73
N=1000; n=5; 5-category, GPCM	0.89	0.84
N=5000; n=5; 5-category, GPCM	0.8	0.72
	Relative G-coefficient	Absolute G-coefficient
N=200; n=5; 10-category, GPCM	0.90	0.85
N=300; n=5; 10-category, GPCM	0.92	0.88
N=500; n=5; 10-category, GPCM	0.93	0.90
N=1000; n=5; 10-category, GPCM	0.93	0.89
N=5000; n=5; 10-category, GPCM	0.92	0.90

In Table 13 above, the sample size is modified for a fixed test length of 5; the GPCM model with 5 categories and a sample size of 1000 yields the highest G-coefficient. Meanwhile, the 10-category sample size of N=500 yields the highest absolute G-coefficient. Although not significantly different from N=1000 and N=5000 (Figure 12).

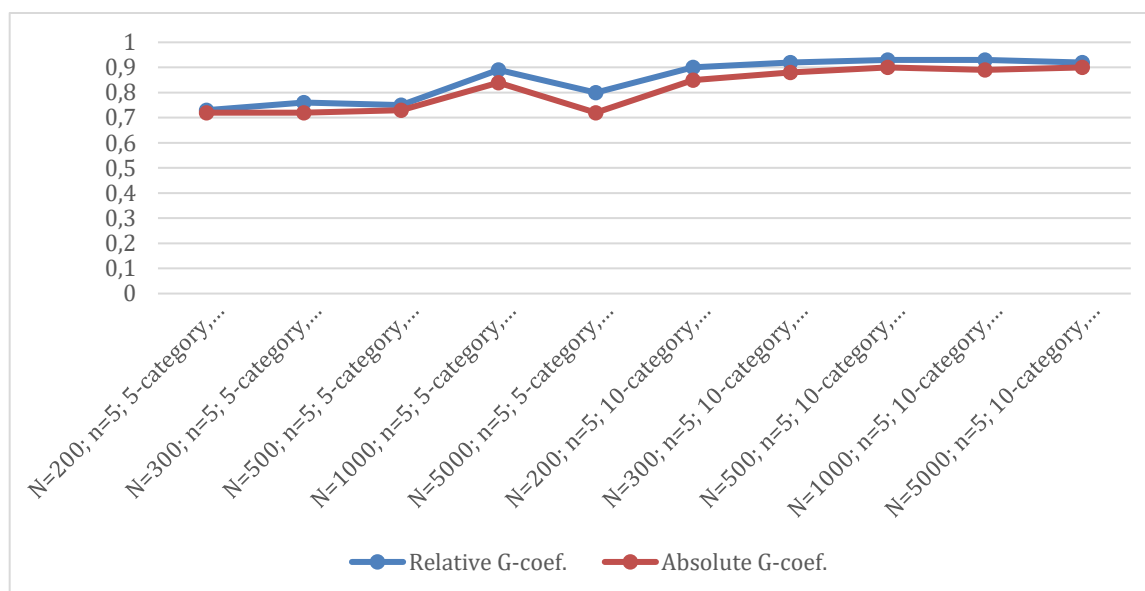


Figure 12. Graph of Relative and Absolute G-coefficients for Modified Sample Sizes

DISCUSSION

Research related to test length and sample size is important. Both need to be carefully designed to produce reliable assessments (Ngozika, 2026). Using Generalizability Theory to guide test development also ensures accurate and reliable evaluation of student performance.

In dichotomous data, the best combination for obtaining the highest relative and absolute G-coefficient values in the 1PL model for a test length of 50 is a sample size of 5000. For the 2PL parameter model with a test length of 50, the sample size that produces the highest relative and absolute G-coefficient values is 1000. Furthermore, for the 3PL parameter model and a test length of 50, the highest relative and absolute G-coefficient values are in sample sizes of 200, 300, and 500.

In polytomous data, two models can be distinguished: the GRM and the GPCM. In the GRM, the highest G-coefficient is observed when $N=200$, $n=100$, and the number of categories is 5. Meanwhile, across the 10 categories, the highest G-coefficient is observed at $N=200$, with a test length of 100. Furthermore, if the test length is 5, then for response categories of 5 and 10, the sample size with the highest G-coefficient value is when $N=300$. This indicates that if researchers are interested in creating 5 test items for the GRM, a sample size of 300 people would be preferable. Based on these findings, the best combination is $N = 300$, test length = 5 items, with 5 or 10 response categories. An example of the application of the GRM parameter model is a questionnaire with a Likert scale, where a sample size of 300 is a good number to get optimal results (Kyriazos, 2018).

Meanwhile, in the GPCM model, for $N=200$ and $n=5$, the 10-category has the highest G-coefficient values compared to the 5- and 7-category. When $N=200$, for categories of 5 and 10, a test length of 50 has the highest G-coefficient in the GPCM model. This can serve as a reference for researchers wishing to create test forms using the GPCM scoring system. The results show that a 10-category design with a test length of 50 and a sample size of 200 yields the highest G-coefficient values among the combinations. Finally, when $n=5$ and the response category is 10, the sample size that yields the highest value is $N=500$, although this is not significantly different from $N=1000$ and $N=5000$. Based on these findings, the best combination for the GPCM model is when the number of samples (N) = 200, test length (n) = 50 items, and 10 response categories, because this combination produces the highest G-coefficient value compared to other combinations. However, if researchers want to use a test length less than 50, $n = 5$ with 10 response categories produces the highest G-coefficient value with $N = 500$. The sample sizes can be reduced if the test length is increased by more than 5 and less than 50. An example of applying the GPCM model is a mathematics essay test, which is scored from 0 to 10 based on the completeness of the solution steps. In addition, practical tests can also use assessments

with a score category of 0–10 (very poor to very good). The test length for both examples can be a minimum of 5, depending on the researcher's needs and the eligible sample size.

These patterns align with the fundamental reasoning of D-studies. Any increase in the number of items, the sample size, or the number of response categories only reduces the standard error. The generalizability coefficient is essentially the ratio of universe-score variance to the sum of universe-score variance and error variance (Cronbach et al., 1972). This explains why, in D-studies, both absolute and relative standard errors of measurement decrease as the number of items increases. The nearly identical results observed at $N=1000$ and $N=5000$ for the GPCM model with $n=5$ items illustrate how the G-coefficient in this study sometimes reaches a plateau or shows only minimal improvement once the sample size exceeds a certain threshold. In other words, Generalizability Theory (GT) does not automatically confer an advantage simply by having a large dataset.

The $N=300$ threshold that repeatedly appears in the GRM analysis results aligns with long-standing sample size recommendations in psychometric research. A parameter-recovery study of the Graded Response Model (GRM) conducted by Reise & Yu (1990) using the MULTILOG software found that the correlation between estimated and true parameters exceeded 0.85 when the sample size reached at least 500. Similarly, Kyriazos (2018) suggested a sample size of approximately 300 respondents as a practical benchmark for stable parameter estimation in scale-based instruments. This is generally consistent with recent findings indicating that while the GRM benefits from larger sample sizes for long instruments with many categories, it achieves stability at $N=300$ for short tests. Studies comparing the performance of the GRM and GPCM—examining variables such as sample size, instrument length, and missing data—have presented simulations allowing for more direct comparisons. These studies found that both models require a minimum sample of 300 respondents and an instrument length of at least five items to yield adequate item parameter estimates, and that the GPCM is relatively more stable than the GRM across various instrument lengths (Dai et al., 2021). These findings are consistent with current research results showing that the highest G-coefficient values for the GPCM are more stable than those for the GRM across variations in the number of categories and test length. The study's conclusion—that a structure with 10 categories does not necessarily yield better results than one with 5 categories—also aligns with a more recent Monte Carlo study by Ikeda (2026) that found that increasing the number of response categories does not automatically improve the quality of GRM parameter estimation when the sample size and number of items are already adequate.

Rather than presenting abstract statistics, these findings offer specific design recommendations for assessment developers. A moderate sample size of 200–300 respondents is sufficient when calibrating dichotomous multiple-choice instruments using the 1PL/Rasch model, as adding thousands more respondents yields only marginal improvements in reliability coefficients. This is advantageous when data collection is costly or time-constrained, aligning with the minimum length guidelines suggested by Ngozika (2026) for effective multiple-choice assessments. Developers should aim for a sample size closer to 1,000 if the 2PL model is required due to varying item discrimination.

According to a generally accepted rule of thumb in questionnaire-based research, Kyriazos (2018) A sample of approximately 300 respondents suffices for short Likert-type scales analyzed using the GRM. For rubric-scored performance tasks—such as essays or practical tests—a 10-category rubric with around 50 items and 200 respondents yields the highest consistent reliability estimates under the GPCM model. However, a shorter rubric with 5 items and the same category range can still perform well with a sample size of $N=500$. These standards are directly applicable to instruments such as the AKM and TKA in Indonesia, which now combine dichotomous and polytomous item formats within a single test administration. Consequently, calibration plans must consider sample size requirements for each format individually, rather than relying on a single, universal rule.

This study makes three specific contributions to the psychometric literature. First, rather than examining these model groups in isolation—as has been the case in most prior Generalizability Theory (GT) research—this study presents a comparative simulation benchmark of G- and D-coefficient performance across various dichotomous (1PL, 2PL, 3PL) and polytomous (GRM, GPCM) models within a unified EduG-based

framework. Second, rather than relying on general guidelines drawn from unrelated statistical traditions, this study generates practical recommendations on sample sizes and test lengths for each model type, thereby providing an empirically grounded starting point for researchers and test developers. Third, this study extends the applied relevance of GT to contemporary educational measurement policy in Indonesia by linking its recommendations to mixed-format assessments, such as the AKM and TKA. This offers a reference point that can directly inform the future calibration and validation of large-scale, mixed-format instruments.

This study has certain limitations that should be noted. First, the data utilized in this research are simulated datasets generated under a standard normal distribution of examinee ability. While this layout provides a controlled environment for testing parametric behavior, empirical settings in educational and psychological testing often encounter skewed, multi-modal, or non-normal distributions. Second, the facets examined were constrained to test length and sample size within single-facet designs. Future studies should evaluate the robustness of these interactions using multifaceted designs (e.g., incorporating raters or testing occasions) and empirical data to further confirm the stability of the G and D study coefficients.

CONCLUSION

This study shows that test length, sample size, model parameters, and the number of score categories influence the Generalizability (G) coefficient and D-study values. For dichotomous data, increasing test length has a greater impact on the G-coefficient than increasing the sample size. The 1PL and 2PL models tend to produce higher G-coefficients than the 3PL. For dichotomous data, test length (n) significantly affects the G-coefficient value. The best combination for achieving the highest reliability (above 0.85) is to use a test length of 50 items. At this test length, the sample size does not significantly change. Researchers are advised to choose the 1PL or 2PL model with a sample size of at least 200 people to achieve efficiency comparable to a large sample.

For polytomous data, the GRM model with a small number of items ($n = 5$) indicates that the optimal sample size is $N = 300$ to obtain a G-coefficient above 0.7. Reliability values will be optimal (above 0.90) if the test length is increased to 50 or 100 items, especially for the response categories of 5 and 10. The GPCM model produces higher G-coefficients than the GRM, especially when the number of score categories and test length are increased. Test lengths of 50–100 items generally yield excellent reliability, with G-coefficients above 0.90. For polytomous data, the GPCM model generally produces higher and more stable G-coefficient values than the GRM. The highest reliability values are obtained with a combination of 10 response categories, a test length (n) of 50, and a sample size (N) of 200. If researchers use only short tests ($n = 5$), a sample size of 1000 (for 5 categories) or 500 (for 10 categories) can yield the maximum G-coefficient results.

The findings of this study indicate that test instrument design needs to consider the optimal combination of test length, sample size, scoring model, and response categories to achieve high measurement reliability. Furthermore, the use of Generalizability Theory through EduG software has proven effective in evaluating the quality of measurement instruments for both dichotomous and polytomous data. Overall, the length of the test and the characteristics of the parameter model (especially the choice of categories in polytomous data) play a more dominant role in increasing the reliability coefficients G and D than a massive increase in sample size.

This study theoretically confirms that the strength of GT lies in its ability to identify sources of error. It demonstrates that a large sample size is not necessarily more important than having a sufficient number of high-quality items. This approach addresses the previously mentioned research gap by enabling the analysis of assessments involving both dichotomous and polytomous scoring. Furthermore, empirical studies often struggle to secure large samples of participants willing to complete assessments containing a high number of items. Using EduG, G-coefficient, and D-coefficient analyses can be performed with simulated data. The findings can assist developers of Likert-based questionnaires, national examinations, or classroom quizzes in designing smaller-scale, more cost-effective data collection processes.

REFERENCES

- Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education*, 4(1). <https://doi.org/10.1080/2331186X.2017.1416898>
- Brennan, R. L. (2001). Generalizability theory: Statistics for social science and public policy. In *New York: Springer-Verlag*. (Vol. 30).
- Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1), 1–21. <https://doi.org/10.1080/08957347.2011.532417>
- Brennan, R. L., & Liao, J. R. (2020). *Generalizability Theory References: The First Sixty Years* (Issue 53).
- Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying Generalizability Theory using EduG*. Routledge. <https://doi.org/10.4324/9780203866948>
- Clauser, B. E. (2008). A Review of the EDUG Software for Generalizability Analysis. *International Journal of Testing*, 8(3), 296–301. <https://doi.org/10.1080/15305050802262357>
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley & Sons, Inc. <https://doi.org/10.3102/00028312011001054>
- Dai, S., Vo, T. T., Kehinde, O. J., He, H., & Xue, Y. (2021). Performance of Polytomous IRT Models With Rating Scale Data: An Investigation Over Sample Size, Instrument Length, and Missing Data. *Frontiers in Education*, 6(September), 1–18. <https://doi.org/10.3389/feduc.2021.721963>
- Eason, S. (1989). Why Generalizability Theory Yields Better Results than Classical Test Theory. *The Annual Meeting of the Mid-South Educational Research Association*, Little Rock.
- Fiangga, S., & Sari, Y. M. (2017). Analisis Generalisabilitas Multifaset pada Instrumen Penalaran Matematika SMP. *Jurnal Elemen*, 3(2), 118. <https://doi.org/10.29408/jel.v3i2.398>
- Fitriyah, I. M. (2023). *Pengembangan Soal AKM Numerasi Berbasis Komputer untuk Kelas XI SMA*. Universitas Negeri Yogyakarta.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory* (D. S. Foster (ed.)). Sage Publication.
- Han, K. T., & Amherst, M. (2007). WinGen: Windows Software That Generates Item Response Theory Parameters and Item Responses. *Applied Psychological Measurement*, 31(5), 457–459. <https://doi.org/10.1177/0146621607299271>
- Handayani, R., Rahariyoso, D., Falani, I., Oktadinata, A., Daya, W. J., & Ockta, Y. (2024). Analyzing The Reliability of Thesis Assessment Instruments: Generalizability Theory. *Journal of Education, Teaching, and Learning*, 9(1), 47–55. <https://doi.org/10.26737/jetl.v9i1.5715>
- Ikedo, T. (2026). A Monte Carlo simulation study of sample size requirements for the Graded Response Model. *PLOS ONE*, 21(4), 1–15. <https://doi.org/10.1371/journal.pone.0347684>
- Irianty, R., Selang, R. A. R., Ihsan, H., Supriyati, Y., & Falani, I. (2025). Generalizability Theory Analysis of Local Curriculum Validation Instruments: Item Effects and Rater Consistency. *Al-Ishlah: Jurnal Pendidikan*, 17(4), 6963–6970. <https://doi.org/10.35445/alishlah.v17i4>
- Istiyono, E. (2016). The application of GPCM on the MMC test as a fair alternative assessment model in physics learning. *Proceedings of the 3rd International Conference on Research, Implementation and Education of Mathematics and Science (ICRIEMS)*, May, 25–30.
- Karami, H. (2012). The Relative Impact of Persons, Items, Subtests, and Academic Background on Performance on a Language Proficiency Test. *Psychological Test and Assessment Modeling*, 54(3), 211–226.
- Kartono. (2008). Equating the combined dichotomous and polychotomous item test model in an achievement test. *Jurnal Penelitian Dan Evaluasi Pendidikan*, 12(2), 302–320.
- Kyriazos, T. A. (2018). Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General. *Psychology*, 9, 2207–2230. <https://doi.org/10.4236/psych.2018.98126>
- Lakes, K. D., & Hoyt, W. T. (2013). Applications of Generalizability Theory to Clinical Child and Adolescent Psychology Research. *J Clin Child Adolesc Psychol*, 38(1), 144–165.

<https://doi.org/10.1080/15374410802575461>

- Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/BF02296272>
- Nababan, D., Nurfahrunnisa, A., & Rachmaniar. (2025). Analisis Kualitas Butir Soal Tes Kemampuan Akademik (TKA) SMA di Kota Merauke. *Journal of Authentic Research*, 4(2), 2075–2083. <https://doi.org/10.36312/ep2nry15>
- Ngozika, O. E. (2026). Determining Minimum Test Length for Reliable Assessment Using Generalizability Theory: Evidence from an Economics Multiple-choice Test. *African Educational Research Journal*, 14(1), 128–133. <https://doi.org/ISSN: 2354-2160>
- Ostini, R., & Nering, M. L. (2006). *Polytomous Item Response Theory Models*. SAGE Publications, Inc.
- Parriott, D. (2017). Using generalizability theory to investigate sources of variance of the Autism Diagnostic Observation Schedule-2 with trainees, in *Dissertation Abstracts International Section A: Humanities and Social Sciences*.
- Reise, S. P., & Yu, J. (1990). Parameter Recovery in the Graded Response Model Using MULTILOG. *Journal of Educational Measurement*, 27(2), 133–144. <https://doi.org/10.1111/j.1745-3984.1990.tb00738.x>
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory - A Primer*.
- Suliyannah, A., B. D., Jauhariyah, M. N. R., Misbah, Mahtari, S., Saregar, A., & Deta, U. A. (2021). A bibliometric analysis of minimum competency assessment research with VOS viewer related to the impact on physics education in 2019-2020. *Journal of Physics: Conference Series*, 2110(1). <https://doi.org/10.1088/1742-6596/2110/1/012022>
- Susongko, P. (2016). Validation of science achievement test with the Rasch model. *Jurnal Pendidikan IPA Indonesia*, 5(2), 268–277. <https://doi.org/10.15294/jpii.v5i2.7690>
- Susongko, P. (2015). Studi generalisabilitas tes tipe dua facet dengan menggunakan analisis varian tiga jalur. *Prosiding Seminar Nasional Pendidikan*, 1–9.
- Teker, G. T., Guler, N., & Uyanik, G. K. (2015). Comparing The Effectiveness of SPSS and EduG Using Different Designs for Generalizability Theory. *Kuram ve Uygulamada Eğitim Bilimleri*, 15(3), 635–645. <https://doi.org/10.12738/estp.2015.3.2278>
- White, M. C. (2017). *Generalizability of Scores from Classroom Observation Instruments*. University of Michigan.