Journal of Education, Teaching, and Learning
Volume 9 Number 1 2024, Special Issue. Page 47-55
p-ISSN: 2477-5924 e-ISSN: 2477-8478

ANALYZING THE RELIABILITY OF THESIS ASSESSMENT INSTRUMENTS: GENERALIZABILITY THEORY

Reli Handayani[1)], Ernanda[2)], Dwi Rahariyoso[3)], Rachmawati[4)], Ilham Falani[5)]

[1)] *Universitas Jambi, Jambi, Indonesia*
E-mail: reli_handayani@unja.ac.id
[2)] *Universitas Jambi, Jambi, Indonesia*
E-mail: ernanda@unja.ac.id
[3)] *Universitas Jambi, Jambi, Indonesia*
E-mail: dwirahariyoso@unja.ac.id
[4)] *Universitas Jambi, Jambi, Indonesia*
E-mail: Rachmawati@unja.ac.id
[5)] *Universitas Jambi, Jambi, Indonesia*
E-mail: ilhamfalani@unja.ac.id

Abstract. This study aims to analyze the reliability of the thesis assessment instrument using Generalizability Theory (G-Theory). This research is a quantitative study that applies G-Theory. G-Theory is unique and useful in evaluating the sources of actual variance and error in any measurement design. Thus, G-Theory contributes to improving reliability and validity in assessment. In this study, data were collected using a thesis assessment instrument that had been tested on thesis assessments conducted at the Faculty of Teacher Training and Education, Universitas Jambi. G-Theory analysis in this study includes several main points, namely G-Study, Optimization (D-Study), and G-facets analysis. The results of the analysis show that the assessment instrument has a fairly high level of reliability with a G coefficient of 0.78. This indicates that the instrument can be used effectively in the thesis assessment. In addition, based on the results of the analysis carried out, recommendations were obtained to reassess item number two, increase the number of raters to six people to increase the reliability of the instrument to 0.80 to meet the satisfactory category. The resulting recommendations can be used by developers to improve the reliability of instruments in future assessment activities. This research is also expected to be a reference for future researchers in applying G-Theory in analyzing the reliability of measurement instruments.

Keywords**:** Generalizability Theory; Accuracy; Assessment; Thesis

I. INTRODUCTION

Thesis as one of the compulsory courses for students in higher education is a prerequisite for students to obtain a bachelor's degree. The thesis specifically refers to a scientific paper in the form of a detailed description related to the results of student research in the undergraduate program (S1) which discusses a problem/phenomenon in a particular field of science using research methodology and applicable rules. To complete a thesis, students are required to have high-level analysis, synthesis, and evaluation skills using existing knowledge. Then students apply it to new situations that have never been encountered before, namely the ability and critical thinking skills (Ponder et al., 2004; Wass et al., 2001).

Conceptually and qualitatively, the thesis differs from other scientific publications in several aspects that are necessary for academic accomplishment. They are objective, procedural, empirical, rational, and original. The thesis has significance and qualifications that exceed essays in general due to five fundamental criteria (Quality Assurance Team, 2021: 2). In general, the standards proposed correspond to the ideal scientific presentation of the thesis work. It is believed that these criteria will be sufficient to meet the academic standard.

Not only that, but having the outcomes of academic and scientific study accountable before the board of examiners also plays a significant role in enhancing the quality of a student's thesis work. Two factors are taken into consideration when it comes to the thesis examination: the students' comprehension of the prepared thesis topic and their oral communication skills (Joughin, 1998). It is necessary for students to possess the skills in formulating logical and scientific arguments and present them in writing. Additionally, students must demonstrate that their prepared

draft thesis represents their own original work and displays their mastery of the subject matter and literature (Swift & Douglas, 1997). According to Jackson & Tinkler (2001), thesis exams can assess students'. In accordance with Jackson & Tinkler (2001), thesis exams can gauge a student's comprehension of the literature, research gaps, and ability for conducting research. In other words, thesis examinations allow lecturers to assess students in terms of knowledge, understanding, and application of complex and abstract concepts.

The assessment process in thesis exams is vulnerable to issues of reliability and bias. This is consistent with Paul's (1994) claim that a thesis examination's assessment is highly subjective. Because there are still disparities in the range of thesis score with intervals of more than 10, this also occurs at FKIP Universitas Jambi. This issue arises because the thesis examiner board, which is composed up of several lecturers, finds it challenging to maintain uniformity in grading among . Furthermore, the absence of standard thesis assessment criteria causes thesis assessments to vary across lecturers and study programs. This clearly does not adhere to subchapter on Procedure of the Thesis Examination in Unja Thesis Writing Guidebook, where point (h) states that "the difference in the scoring scale of thesis from each examiner cannot be more than 10 (ten)" (Quality Assurance Team, 2021: 8-9). In this case, the thesis examining board's assessment has been set up at specific intervals to ensure that there is no significant discrepancy.

The thesis gets 10 (ten) course credits, which can be split down into 2 (two) course credits for the thesis research proposal and 8 (eight) course credits for thesis writing, as further detailed in the Thesis Writing Guidebook (Quality Assurance Team, 2021: 3). The final thesis has a significant impact on the student's grade point average (GPA) given the credits mentioned. Ideally, students benefit from having their performance fairly assessed. On the other hand, inaccurate assessments have the potential to seriously harm the students. As a consequence, both lecturers and students require precise, transparent, and consistent assessment during the thesis examination. The purpose of this study is to provide an overview of the criteria for assessment and guidelines that are currently approved for every study program in the Faculty of Teacher Training and and Education, Universitas Jambi. The description of measurable and accurate criteria is expected to be the basis for creating an effective thesis assessment model in eliminating inconsistencies in the thesis assessment.

Using the Generalizability Theory (G-Theory) is one technique being used to increase the thesis assessment's accuracy. Implementation of G-Theory, in the analysis of thesis assessment reliability offers several advantages. First, G-Theory allows researchers to consider multiple sources of variability in measurement, such as subjects, items, and raters, thus providing a more comprehensive picture of reliability. Second, G-Theory provides a framework for designing and improving measurement instruments, by considering how changes in design can affect reliability. Third, G-Theory allows researchers to generalize assessment results to a broader context, which is important in academic research (Brennan, 1992; Hapsan & Rosnawati, 2023; Shavelson & Webb, 2012; N. M. Webb et al., 2012). As a result, applying G-Theory can increase the reliability and validity of the thesis assessment. G-Theory is a tool that needs to be considered in the thesis or thesis evaluation process.

Numerous previous studies relevant to the investigation of the thesis assessment instrument's or thesis's reliability have been conducted. Ramadhan (2019) and Suwita (2020) developed an information system for thesis assessments. Additionally, Lumaurridlo's (2019) uses G-Theory to estimate the reliability of the Munaqosah assessment. Khosim (2022) examined the generalizability theory as well as the effects of learning styles, motivation, and self-regulation on the academic achievement of Perak's local students. Through the application of G-theory, Retnowati research (2009) created an instrument for evaluating children's artwork in primary schools. Nurmala & Retnowati's (2013) examines the development of thesis assessment instruments in the history department, of Padang State University. Safitri et al.'s research (2024) estimated measurement error in thesis assessment using generalizability theory analysis. The results of the literature study show that research related to the thesis assessment instrument or thesis has not specifically applied G-Theory, especially at the Faculty of Teacher Training and Education (FKIP), Universitas Jambi.

Based on the background described above, this study aims to analyse the reliability of the thesis assessment instrument or thesis by applying G-Theory. The problem formulations presented in this study are how to implement G-Theory in analysing the reliability of the thesis assessment instrument or thesis at FKIP Universitas Jambi. Furthermore, how is the reliability of the thesis assessment instrument at FKIP Universitas Jambi? It is expected that this research will improve the thesis assessment's measurement accuracy. Furthermore, the quality of the thesis assessment in the future will be better.

II. METHODS

The research method used in this study is a quantitative method by applying Generalizability Theory (G-Theory). (Cardinet et al., 2010; Soesana et al., 2023) This research was conducted at the Faculty of Teacher Training and Education, Universitas Jambi, which is located on Jl. Jambi - Muara Bulian No.KM. 15, Mendalo Darat, Kec. Jambi Luar Kota, Muaro Jambi Regency, Jambi.

A. Sampling Technique

The participants in this study were students of the Faculty of Teacher Training and Education, Universitas Jambi who were undergoing the Thesis or Thesis Examination process in the even semester of 2023/2024. Meanwhile, the selection of participants in this study was carried out using a purposive sampling technique. The purposive sampling

Journal of Education, Teaching, and Learning
Volume 9 Number 1 2024, Special Issue. Page 47-55
p-ISSN: 2477-5924 e-ISSN: 2477-8478

technique is a non-probability sampling technique, in which researchers select subjects based on the specific attributes or qualities they have. This method is used when the researcher has a specific goal and needs subjects with specific attributes to achieve that goal. In the case of this study, the specific attributes in question are students of the Faculty of Teacher Training and Education, Universitas Jambi who are undergoing the Thesis or Thesis Examination process in the even semester of 2023/2024. By using a purposive sampling technique, seven participants were obtained from seven different study programs.

B. Data Collection Technique

The data collected in this study are data related to the results of the thesis assessment of students who complete the thesis examination stage. Meanwhile, the data collection technique used in this study involves the use of a thesis assessment instrument that has been developed and validated. This instrument was used to assess students' thesis assignments. Each participant was assessed by board of examiners consisting of 2 supervisors and 3 examiners. The thesis assessment instrument consists of 8 items that assess aspects: research background, theoretical studies, research methods, research results, research benefits, writing systematics, attitude, and argumentation.

C. Data Analysis Technique

The data analysis conducted in this study is an analysis of the reliability of the thesis assessment instrument using the Generalizability Theory (G-Theory). (Cardinet et al., 2010; Soesana et al., 2023) Generalizability Theory (G-Theory) is essentially a method for estimating measurement accuracy under conditions where measurements have various sources of error. (Cardinet et al., 2010; Susongko, 2010) One of the most important and most basic points of view of the G-theory model is that it exposes various sources of measurement error. G-theory provides a broad conceptual framework and a powerful set of statistics to address a wide range of measurement issues. G-theory is often considered a further development of Classical Test Theory (CTT) and Analysis of Variance (ANOVA). (Brennan, 2010; Briesch et al., 2014; Crocker & Algina, 1986; Hove et al., 2022). G-theory not only has advantages in the estimation of measurement dependability but can also estimate the contribution of measurement error, allowing it to be used to improve measurement procedures in future applications (Cardinet et al., 2010; Soesana et al., 2023).

The Generalizability Theory (G-Theory) used in this study is the Two Facet Design (p x i x h). p is the thesis rater (supervisor/examiner), i is the item on the thesis assessment instrument, h is the student assessed during the thesis assessment. The (p x i x h) design illustrates that each rater (supervisor/examiner), assesses each item, for each thesis student. In standard ANOVA terminology, effects within a design can be identified as main effects or interaction effects (Cardinet et al., 2010). In the context of a Two-Facets design, the main effect is indicated by p,i,h, while the interaction effect is denoted by pi,ph,ih,pih. The illustration of the linkage can be seen in the following diagram.

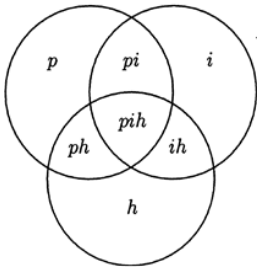

Fig. 1 Venn diagram of the interrelationship between influences in the Two-Facets design $(p \times i \times h)$

In the Venn diagram in Figure 1, each main influence is represented by a circle. The interaction influences are represented by the intersection of the circles. The total number of influences in a Two-Facets design is represented by the number of distinct areas within the Venn diagram (Brennan, 2010).

The following are the mathematical formulas for scores and coefficients in Generalization Theory (G-Theory) adapted from Brennan (Brennan, 2010; Briesch et al., 2014; Chiu, 2001a).

G-Study

$$X_{pir} = \mu \quad \text{(grand Mean)} \quad (1)$$

$$+(\mu_p - \mu) \quad \text{(person effect)} \quad (2)$$

$$+(\mu_i - \mu) \quad \text{(item effect)} \quad (3)$$

$$+(\mu_r - \mu) \quad \text{(rater effect)} \quad (4)$$

$$+(\mu_{pi} - \mu_p - \mu_i + \mu) \quad \text{(person*item interaction)} \quad (5)$$

$$+(\mu_{pr} - \mu_p - \mu_r + \mu) \quad \text{(person*rater interaction)} \quad (6)$$

$$+(\mu_{ir} - \mu_i - \mu_r + \mu) \quad \text{(rater*item interaction)} \quad (7)$$

$$+(X_{pir} - \mu_{pi} - \mu_{pr} - \mu_{ir} + \mu_p + \mu_i + \mu_r - \mu) \quad \text{(residual)} \quad (8)$$

varians

$$\sigma^2(X_{pir}) = \sigma_p^2 + \sigma_i^2 + \sigma_r^2 + \sigma_{pi}^2 + \sigma_{pr}^2 + \sigma_{ir}^2 + \sigma_{pir,e}^2 \quad (9)$$

D-Study

$$\text{Relative error variance} = \sigma_\delta^2 = \frac{\sigma_{pi}^2}{n_i'} + \frac{\sigma_{pr}^2}{n_r'} + \frac{\sigma_{pir,e}^2}{n_i' n_r'} \quad (10)$$

Absolute error variance

$$= \sigma_\Delta^2 = \frac{\sigma_i^2}{n_i'} + \frac{\sigma_r^2}{n_r'} + \frac{\sigma_{ir}^2}{n_i' n_r'} + \frac{\sigma_{pi}^2}{n_i'} + \frac{\sigma_{pr}^2}{n_r'} + \frac{\sigma_{pir,e}^2}{n_i' n_r'} \quad (11)$$

$$\text{Koefisien Generizability} = \rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2} \quad (12)$$

$$\text{Koefisien Dependability} = \Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\Delta^2} \quad (13)$$

$$\text{Standar Error Measurement} = \text{SEM} = \sqrt{\sigma_\Delta^2} \quad (14)$$

The mathematical equation above describes the process of estimating the measurement error. The G-Theory mathematical equation above was implemented into the analysis of the reliability of measuring ethnomathematics knowledge in this study. The mathematical calculation of G-

Journal of Education, Teaching, and Learning
Volume 9 Number 1 2024, Special Issue. Page 47-55
p-ISSN: 2477-5924 e-ISSN: 2477-8478

Theory in this study was carried out using the help of EdUG software (Society & Group, 2010). EduG is software that applies Generalizability Theory with a simple and easy-to-use display (Clauser, 2008).

The purpose of this study is to determine the extent to which the assessment results are reliable and to identify sources of variation in the data that may affect the reliability of the results. Thus, this research not only provides insight into the effectiveness of the assessment instruments used, but also makes an important contribution to improving the quality of thesis assessment at the Faculty of Teacher Training and Education, Universitas Jambi.

III. Result and Discussion

A. Result

Facets in G-Theory are synonymous with factors in the Analysis of Variance (ANOVA). The term was first introduced by Guttman to distinguish psychometric and factor analysis contexts (Cardinet et al., 2011; Chiu, 2001b).

If we can identify all the factors that potentially contribute to variation in a set of measurements, then we can "partition" the total variance to reflect the different sources of variance. The purpose of ANOVA is to assess the relative levels of the identified sources of variation (scores), through the estimation of variance components (Cardinet et al., 2011). The variance model is represented in the score decomposition equation. For example, for a two-facets design, where in this study each student (S) was independently assessed by each of the thesis raters (R).

TABLE I
Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (Levels to Exclude)
Persons	P	7	INF	
Items	I	8	INF	
Raters	R	5	INF	

The table above defines a set of information used to describe the data structure of the two facets of design in this study. The first column in the Observation and Estimation Designs table is the Facet Column, in this column, certain characteristics of the facets identified in this study are mentioned, namely Students (person), Items, and Raters. Then the facets are labeled as follows, P, I, and R. Furthermore, the Level Column represents the number of levels in each facet, which are 7, 8, and 5, respectively. The Univ Column represents the universe size of each facet, where the three facets are infinite (INF). Once established, the observation design cannot be changed within the baseline established in the table, although it can be temporarily reduced if analysis is required involving only a subset of the stated facet levels. However, the "Reduction" field allows the user to exclude part of the data set to be analyzed (Society & Group, 2010).

1) Analysis of variance (ANOVA): The output of EduG at the initial stage of G-Study is the variance analysis of the inputted data, as for the results of the variance analysis as presented in the following table.

TABLE II
Analysis of Variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	38.23571	6	6.37262	0.12470	0.12470	0.12470	19.3	0.08007
I	10.62857	7	1.51837	0.02470	0.02470	0.02470	3.8	0.02102
R	2.55000	4	0.63750	-0.00676	-0.00676	-0.00676	0.0	0.00851
PI	28.62143	42	0.68146	0.06815	0.06815	0.06815	10.5	0.02998
PR	25.05000	24	1.04375	0.08788	0.08788	0.08788	13.6	0.03648
IR	8.76429	28	0.31301	-0.00395	-0.00395	-0.00395	0.0	0.01270
PIR	57.23571	168	0.34069	0.34069	0.34069	0.34069	52.7	0.03695
Total	171.08571	279					100%	

The table above displays the results of the analysis of variance for each source of variance from the defined measurement design which includes: P, I, R, as well as the interaction between their facets: PI, PR, IR. The results of the analysis of variance that have been produced by EduG as in the table above are then used by EdUG as a reference for calculating the G-Study and D-Study in this study (Briesch et al., 2014; N. Webb et al., 2006).

2) Generalizability Study (G-Study): The next stage in Generalizability Theory is the Generalizability Study (G-Study). In this stage, EdUG performs calculations based on the ANOVA results obtained in the previous stage to conclude the quality of measurement for the selected differentiation facet. The G-Study outputs produced by EdUG are presented in the following table,

Journal of Education, Teaching, and Learning
Volume 9 Number 1 2024, Special Issue. Page 47-55
p-ISSN: 2477-5924 e-ISSN: 2477-8478

TABLE III
G STUDY TABLE

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.12470		.....		.....	
	.....	I	.....		0.00309	8.2
	.....	R	.....		(0.00000)	0.0
	.....	PI	0.00852	24.6	0.00852	22.6
	.....	PR	0.01758	50.8	0.01758	46.6
	.....	IR	.....		(0.00000)	0.0
	.....	PIR	0.00852	24.6	0.00852	22.6
Sum of variances	0.12470		0.03461	100%	0.03770	100%
Standard deviation	0.35313		Relative SE: 0.18605		Absolute SE: 0.19417	
Coef_G relative	0.78					
Coef_G absolute	0.77					

Grand mean for levels used: 3.44286
Variance error of the mean for levels used: 0.02585
Standard error of the grand mean: 0.16077

Estimate of Phi(lambda)
Cut Score = lambda = 3
Estimate of Phi(lambda) = 0.88667

The first point to note in the table above is that in the bottom row there is a generabilibity coefficient, the Coef_G Relative indicates the reliability and precision of the measurements made. While Coef_G Relative indicates sources of variance affecting the relative measurement scale, Coef_G absolute also accounts for additional sources of error associated with the absolute measurement scale. Both provide values above 70, 0.78 and 0.77 respectively. In addition, the table above also provides information on the value of the standard error of measurement both in relative and absolute terms, which are 0.18605 and 0.19417, respectively.

Based on the above points, the researcher tries to identify which of these facets is the biggest source of error in the measurements made. One of the best features that Generalizability Theory has but Classical Test Theory (CTT) cannot do is the ability to determine the exact location of the error and quantify the error. Therefore, we instrument developers can correct errors more precisely to improve measurement accuracy in the future.

In the table above, the various facets and their interactions are separated into differentiation facets (first two columns) and instrumentation facets. There are two Source of variance columns that show how the sources of variance are divided by measurement design. The contribution of each type of variance to the error variance (relative or absolute) is detailed in the %relative or %absolute column. The "%absolute" column represents how much absolute error variance is divided based on its contribution as a source of error. This information allows us to identify the sources of variance that have the greatest negative impact on the precision of the measurement. This information is very useful in conducting D-studies to improve measurement precision. The total sum of the variances row is an estimate of the true variance and error variances for relative measurement and absolute measurement. In the table above, it can be seen that PI; PIR; and PR contribute the largest error contribution in the measurements made on both relative and absolute measurement scales with consecutive values of 24.6%; 24.6%; 50.8% (relative measurement) and 22.6%; 22.6%; and 46.6% (absolute measurement).

At the bottom of the table EduG displays the Estimate of Phi(lambda). Estimate of Phi(lambda) is a coefficient estimation procedure for absolute measurement scales, this type of measurement is in demand or used in many educational measurements. This measurement scale considers the difference between the score obtained by a test taker and the minimum score criteria that must be met. In the EduG Estimate of Phi(lambda) in this study, a cutting Score of 3 was set, this number was taken from the smaller integer closest to the grand mean of 3.44, resulting in an Estimate of Phi(lambda) of 0.88667.

3) Study Design (D-Study): After reviewing the initial results of the G-Study analysis, the researcher conducted further analysis with the aim of improving the quality of the measurement instrument. EduG offers two features that may be used: Optimization and G-facets analysis.

Performing optimization, also known as a Design Study (D-Study), allows users to see the possible effects on relative and absolute coefficients of changing the number of levels of one or more facets, and/or changing the characteristics of the facets. EduG responds to optimization (D-Study) by recalculating the new contribution of each source of error to measurement error based on the variation of facets. EduG also estimates Coef_G and other parameters as a result of the alternative Design Study (D-Study).

The optimization facility allows researchers to vary the number of levels observed for one or more facets of instrumentation, and to see the potential effects of such

Journal of Education, Teaching, and Learning
Volume 9 Number 1 2024, Special Issue. Page 47-55
p-ISSN: 2477-5924 e-ISSN: 2477-8478

changes on measurement reliability. Researchers can increase the number of levels observed from larger contributors to error variance and, if this contributes to cost-effectiveness, researchers can reduce the number of levels observed from instrumentation facets that contribute little to error variance. The results of the optimization (D-study) with the help of EDuG can be seen in the following table.

TABLE IV
OPTIMIZATION

	G-Study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	**Univ.**	**Lev.**	**Univ.**	**Lev.**	**Univ.**	**Lev.**	**Univ.**	**Lev.**	**Univ.**	**Lev.**	**Univ.**
P	7	INF	7	INF	7	INF	7	INF	7	INF	7	INF
I	8	INF	8	INF	8	INF	8	INF	8	INF	8	INF
R	5	INF	3	INF	4	INF	5	INF	6	INF	7	INF
Observ.		280		168		224		280		336		392
Coef_G rel.		0.78274		0.70568		0.75195		0.78274		0.80471		0.82117
rounded		0.78		0.71		0.75		0.78		0.80		0.82
Coef_G abs.		0.76786		0.69357		0.73820		0.76786		0.78898		0.80480
rounded		0.77		0.69		0.74		0.77		0.79		0.80
Rel. Err. Var.		0.03461		0.05201		0.04114		0.03461		0.03026		0.02716
Rel. Std. Err. of M.		0.18605		0.22805		0.20282		0.18605		0.17397		0.16480
Abs. Err. Var.		0.03770		0.05510		0.04422		0.03770		0.03335		0.03025
Abs. Std. Err. of M.		0.19417		0.23473		0.21030		0.19417		0.18263		0.17391

The right side of Table 4 shows the number of alternative observation levels of each generalization aspect and/or alternative sampling status (by changing the universe size). In Column Option 1, the facet R level is modified to 3, Option 2 facet R level is modified to 4, Option 3 facet R level corresponds to the initial level, Option 4 facet R level is modified to 6, Option 5 facet R level is modified to 7. EduG allows researchers to make modifications 5 times. Based on the table above, it can be seen that there is an increase in the relative Coef_G and absolute Coef_G values as the modified facet R level increases. In the modifications made, the relative and absolute Coef_G meet the satisfactory criteria in option 4 where the size of facet R becomes 6, which results in relative and absolute Coef_G being 0.80 and 0.79. With relative and absolute measurement standard errors reduced to 0.034 and 0.037.

The results from the table above show that the assessment conducted by 5 people is not satisfactory in terms of measurement reliability. A minimum of six assessments are required to meet the satisfactory criteria for the reliability of the relative measurement scale (0.80) and the absolute measurement scale (0.79). In some ways, optimizing measurement precision by increasing the number of facet levels is a straightforward strategy, as it can result in increased measurement reliability without detailed knowledge of what negatively impacts this.

4) G-Facets Analysis: G-Facets analysis generalizes item analysis, a standard procedure in psychometrics. Its purpose is to compare the G coefficient values obtained when each level of the facet under study is excluded from the analysis.

G-Facets Analysis can show more precisely than optimization the extent to which each level of instrumentation facet affects the relative and absolute coefficients. EduG facilitates G-Facets Analysis through features accessible on the workscreen. EduG allows researchers to select specific facets to analyze from a list of predefined instrumentation facets.

EduG's G-facets analysis output provides the relative and absolute G coefficient values obtained when each level of the selected instrumentation facet is analyzed. To "improve" measurement precision, researchers can exclude certain facets from repeating the analysis, to eliminate interaction effects. The results of the G-facet analysis with the help of EduG are shown in the following table.

TABLE V
G-FACETS ANALYSIS

Facet	**Level**	**Coef_G rel.**	**Coef_G abs.**
I	1	0.77965	0.76320
	2	**0.81861**	**0.80031**
	3	0.72323	0.70690
	4	0.75414	0.73166
	5	0.77098	0.74829
	6	0.77806	0.77604
	7	0.77341	0.75449
	8	0.76372	0.74633
R	1	0.69956	0.68407
	2	0.73499	0.72457
	3	**0.84007**	**0.82362**
	4	0.65430	0.63629
	5	**0.80581**	**0.79585**

In Table 5 above, it should be noted that Item number 2 tends to reduce reliability. Without item number 2, the reliability of the remaining 7 items rises to 0.8186 for relative measurement and 0.8003 for absolute measurement. Furthermore, in facet R, it is obtained from the table above that reliability increases if eliminating rater 3 or rater 5, this can be seen from the relative and absolute Coef_G increasing to $\geq$ 0.80.

Journal of Education, Teaching, and Learning
Volume 9 Number 1 2024, Special Issue. Page 47-55
p-ISSN: 2477-5924 e-ISSN: 2477-8478

B. Discussion

The relative and absolute Coef_G from the G-Study analysis results give values above 70, 0.78 and 0.77 respectively. From these results, we can understand that there is still an error of about 22%. This indicates that the measurement carried out with the number of students, items, and assessors is not fully satisfactory ($\geq$80, satisfactory) (Brennan, 1992; Cardinet et al., 2011; Fiangga & Sari, 2017; Safitri et al., 2024). Based on further identification of error sources, it can be seen that PI; PIR; and PR contribute the largest error contribution in the measurements taken. PI is the interaction between Person and Item contributing 24.6% of the error. We can interpret that there is a problem in the person being assessed with the instrument item used. Furthermore, with PIR which is the interaction between person, item, and rater contributes 24.6% error. This shows that there is a problem with the person being assessed by being assessed with the assessment instrument items by the raters. Furthermore, the largest percentage of error contribution was in IR at 50.8%. This adds more information that the interaction between person and rater further adds to the measurement problems. This problem is very likely caused by the sampling of 7 different people from different study programs under the Faculty of Teacher Training and Education, Universitas Jambi. This sampling needs to be a future consideration for future researchers in analyzing the reliability of assessment instruments using generalizability theory. In this study, the sample taken was one person from each study program so that increasing reliability through the D-study feature is difficult to do, because eliminating one person means eliminating representation from one study program (Cardinet et al., 2011; Falani & Kumala, 2017; Setyonugroho, 2017; Society & Group, 2010).

In this study, efforts to improve the reliability of assessment instruments using optimization (D-Study) and G-facets Analysis focused on instrumentation facets including Items (I) and Raters (R). This is because the person (facet differentiation) in this study cannot be modified due to the limited number of samples as described in the previous paragraph.

The D-Study results show that modifications made to the rater facet level provide results that can increase the reliability of the measurement instrument (Susilaningsih, 2014). The modification is in option 4, where the relative and absolute Coef_G meets the satisfactory criteria in option 4 where the size of facet R becomes 6, which results in relative and absolute Coef_G being 0.80 and 0.79. With the standard error of relative and absolute measurements reduced to 0.034 and 0.037. This is one of the important information in the development of the thesis assessment instrument in the future to increase reliability (Hapsan & Rosnawati, 2023; Safitri et al., 2024).

Furthermore, efforts to increase the reliability of the thesis assessment instrument can also be seen from the results of the G-facets analysis conducted (Briesch et al., 2014; N. Webb et al., 2006). The results of the first G-facets analysis provide information that eliminating item number 2 on the instrument can increase the reliability of the measurements made, the reliability of the remaining items of the 7 items rose to 0.8186 for relative measurements and 0.8003 for absolute measurements. However, the elimination of this instrument item needs to consider the representativeness of the assessment indicators (Cardinet et al., 2011), it is necessary to look again at the lattice of the thesis assessment instrument used as presented in the table below.

TABLE VI
ASSESSMENT INSTRUMENT LATTICE

Dimension	Indicator	Item	Format
Thesis/thesis assesment	Background and research problem	1	Rating Scale
	Theoretical review	2	Rating Scale
	Research method	3	Rating Scale
	Results, discussion of research results, conclusion, and suggestions	4	Rating Scale
	Benefits	5	Rating Scale
	Attitude	6	Rating Scale
	Systematics and writing	7	Rating Scale
	Argumentation	8	Rating Scale

In the research instrument lattice table above, it can be seen that item number 2 is about the Theoretical Review indicator. This indicator is only represented by one item, so the elimination of this item can make the measurement of the measurement dimension incomplete. However, for items related to theoretical studies, it is necessary to review the scoring criteria, so that the assessment given by the rater can be more precise to increase the accuracy of the measurements that will be carried out in the future.

IV. CONCLUSIONS

Analysis of the reliability of the thesis assessment instrument is important considering that the thesis is one of the graduation requirements to obtain the final degree of education at the undergraduate level. The implementation of Generalizability Theory provides advantages compared to Classical Test Theory, namely its ability to identify sources of variation in the assessment that can affect the reliability of the thesis assessment results. Based on the results obtained from the G-Study and D-Study information that has been carried out, it is necessary to revise or review the rubric for the assessment of the thesis assessment instrument, especially on item number two. In addition, information was also obtained that to be able to improve the measurement required a minimum of six raters, one more person was needed so that the measurement results met the satisfactory criteria.

REFERENCES

Brennan, R. L. (1992). Generalizability Theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34.

Brennan, R. L. (2010). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education*, *1*(1),

1–21. https://doi.org/10.1080/08957347.2011.532417

Briesch, A., Swaminathan, H., Welsh, M. E., & Chafouleas, S. M. (2014). Generalizability Theory: A Practical Guide to Study Design, Implementation, and Interpretation. *Journal of School Psychology*, *52*(1), 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

Cardinet, J., Johnson, S., & Pini, G. (2010). Applying Generalizability Theory Using EduG. *In Routledge. Routledge of the Taylor& Francis Group*. https://revistas.ufrj.br/index.php/rce/article/download/1659/1508%0Ahttp://hipatiapress.com/hpjournals/index.php/qre/article/view/1348%5Cnhttp://www.tandfonline.com/doi/abs/10.1080/09500799708666915%5Cnhttps://mckinseyonsociety.com/downloads/reports/Educa

Cardinet, J., Johnson, S., & Pini, G. (2011). Applying Generalizability Theory using EduG. In *Routledge*. Routledge of the Taylor& Francis Group. https://revistas.ufrj.br/index.php/rce/article/download/1659/1508%0Ahttp://hipatiapress.com/hpjournals/index.php/qre/article/view/1348%5Cnhttp://www.tandfonline.com/doi/abs/10.1080/09500799708666915%5Cnhttps://mckinseyonsociety.com/downloads/reports/Educa

Chiu, C. W.-T. (2001a). Scoring Performance Assessments Based on Judgements: Generalizability Theory. In *Springer Science & Business Media* (Vol. 50). Springer Science & Business Media.

Chiu, C. W.-T. (2001b). Scoring Performance Assessments Based on Judgements. In *In Scoring Performance Assessments Based on Judgements*. https://doi.org/10.1007/978-94-010-0650-7

Clauser, B. (2008). A Review of the EDUG Software for Generalizability Analysis. *International Journal of Testing*, *8*(3), 296–301.

Crocker, L. M., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. ERIC.

Falani, I., & Kumala, S. A. (2017). Kestabilan Estimasi Parameter Kemampuan pada Model Logistik Item Response Theory Ditinjau Dari Panjang Tes. *SAP (Susunan Artikel Pendidikan)*, *2*(2). https://doi.org/10.30998/sap.v2i2.2028

Fiangga, S., & Sari, Y. M. (2017). Analisis Generalisabilitas Multi Faset pada Instrumen Penalaran Matematika SMP. *Jurnal Elemen*, *3*(2), 118–129.

Hapsan, A., & Rosnawati, R. (2023). *Koefisien Realibilitas dengan Teori Generalizabilitas (G-Theory) two-Facet i: h: p design*.

Hove, D. Ten, Jorgensen, T. D., & van der Ark, A. (2022). Interrater Reliability for Multilevel Data: A Generalizability Theory Approach. *Psychological Methods*, *4*(4), 650.

Jackson, C. P., & Tinkler, P. (2001). Back to Basics: a Consideration of the Purposes of the PhD viva. *Assessment & Evaluation in Higher Education*, *26*(4), 355–366.

Joughin, G. (1998). Dimensions of Oral Assessment. *Assessment & Evaluation in Higher Education1*, *23*(4), 367–378.

Khosim, F. (2022). *Pengujian Teori Generalizabiliti serta Pengaruh Motivasi, Pembelajaran Regulasi Kendiri dan Gaya Pembelajaran terhadap Pencapaian Akademik Murid Orang Asli di Perak*. Universiti Utara Malaysia.

Lumaurridlo. (2019). Estimasi Keandalan Penilaian Munaqosah. *Jurnal Tawadhu*, *3*(1), 665–673.

Nurmala, M. D., & Retnowati, T. H. (2013). Pengembangan Instrumen Penilaian Skripsi Mahasiswa. *Jurnal Evaluasi Pendidikan*, *1*(1), 25–33.

Paul, V. K. (1994). Assessment of Clinical Competence of Undergraduate Medical Students. *The Indian Journal of Pediatrics*, *61*, 145–151.

Ponder, N., Beatty, S., & Foxx, W. (2004). Doctoral Comprehensive Exams in Marketing: Current Practices and Emerging Perspectives. *Journal of Marketing Education*, *26*(3), 226–235.

Ramadhan, W. F. R. (2019). *Sistem Informasi Penilaian Tugas Akhir*. Universitas Islam Indonesia.

Retnowati, T. H. (2009). Pengembangan Instrumen Penilaian Karya Seni Lukis Anak di Sekolah Dasar. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *13*(1).

Safitri, I., Rosnawati, R., Ansyari, R., & Abada, R. (2024). Estimasi Kesalahan Pengukuran dalam Penilaian Sidang Skripsi: Generalizability Theory Analysis. *Afeksi: Jurnal Penelitian Dan Evaluasi Pendidikan*, *5*(1), 162–168.

Setyonugroho, W. (2017). Gentle Introduction of Generalizability Theory Analysis in OSCE Using EduG for Medical Educators. *Advanced Science Letters*, *23*(12), 12656–12659.

Shavelson, R. J., & Webb, N. M. (2012). Generalizability theory. In *In Handbook of Complementary Methods in Education Research* (pp. 309–322). Routledge.

Society, S., & Group, E. W. (2010). *Edug User Guide*. Edumetrics.

Soesana, A., Subakti, H., Karwanto, Kuswandi, A. F. S., Sastri, L., Falani, I., Aswan, N., Hasibuan, F. A., & Lestari, H. (2023). *Metodologi Penelitian Kuantitatif*. Yayasan Kita Menulis.

Susilaningsih, E. (2014). Instrumen Penilaian Praktikum Kimia dan Estimasi Reliabilitasnya dengan Koefesien Generalisabilitas. *Prosiding, Seminar Nasional Kimia Dan Pendidikan Kimia VI UNS*.

Susongko, P. (2010). Studi Generalizabilitas Tes Tipe Dua Facet dengan Menggunakan Analisis Varian Tiga Jalur. *OSEATEK*, *07*.

Suwita, F. S. (2020). Pengembangan Sistem Informasi Tugas Akhir dan Skripsi (SIMITA) di Universitas Komputer Indonesia (UNIKOM). *Jurnal Teknologi Dan Informasi*, *10*(1), 71–82.

Swift, J., & Douglas, A. (1997). *The viva voce: A Research Guide*. Research Training Initiative, Birmingham Institute of Art & Design, University of Central England.

Wass, V., Vleuten, V. der, Shatzer, J., & Jones, R. (2001). Assessment of Clinical Competence. *The Lancet*, *357*(9260), 945–949.

Webb, N. M., Shavelson, R. J., & Steedle, J. T. (2012).

Generalizability Theory in Assessment Contexts. In *In Handbook on measurement, assessment, and evaluation in higher education* (pp. 152–169). Routledge.

Webb, N., Shavelson, R. J., & Haertel, E. H. (2006). 4 Reliability Coefficients and Generalizability Theory. *Handbook of Statistics*, *26*, 81–124.