

Gentle Introduction of Generalizability Theory Analysis in OSCE Using EduG for Medical Educators

Winy Setyonugroho

Faculty of Medicine and Health Sciences of the Universitas Muhammadiyah Yogyakarta,
Indonesia

Corresponding author:

Winy Setyonugroho
wsetyonugroho@umy.ac.id
wsetyonugroho@gmail.com

Faculty of Medicine & Health Sciences, Universitas Muhammadiyah Yogyakarta, Indonesia

Generalizability Theory (G-Theory) has become the gold standard of reliability analysis in OSCEs. However, little progress has been made in popularizing G-Theory. This method allows various variance calculations to determine and quantify the source of an error. EduG is the only GUI software able to calculate G-Theory. The objective of this project is to introduce the G-Theory analysis of an OSCE using EduG. This assessment is a one day OSCE consisting of 10 stations per circuit. There are 3 buildings used in this exam. Each building held one complete circuit. G coefficient is 0.76. The majority of error (81.4%) was due to the interaction between students (nested in location) and observations. There is no error coming from the observation itself. Mean from building A, B, and C are 57.41, 57.97, and 61.56, respectively. First conclusion, the possible source of error is greater in building C. Secondly, EduG is capable calculating the reliability analysis in OSCE using G-Theory.

Keywords : Generalizability Theory, OSCE, reliability analysis

1. Introduction

A quick search of the keywords in Pubmed of Generalizability Theory (G-Theory) and Objective Structure Clinical Examination (OSCE) in the past 10 years giving us 14 result only. Even after the recommendation of using G-Theory to analyze OSCE in 2012, there is still far too little attention by researchers or medical educators in the use of this method.¹

One of the greater challenges is the fact that G-Theory is still considered a black art. Learning about the theory itself often scares away most educators. The available G-Theory books are not helping this method become popular, since they are mostly too "statistic". Meanwhile, in the writer's opinion, the majority of medical educators do not like numbers and formula.

G-Theory itself is a powerful tools needed to evaluate measurement quality and give the accurate information to improve the assessment procedures.^{1,2} Compared to Classical Test Theory (CTT), G-Theory is advantageous in its ability to quantify the source of errors.³

So far, Cronbach Alpha remains the common method to assess the OSCE's reliability.⁴ The results are specific and can only be applied to the specific study. Therefore, it cannot be generalized to other studies. On the other hand, G-Study is able to calculate multiple variances at the same time, as well as estimate the source of errors.³

To analyze an OSCE with G-Theory, a computer software is needed. To date, EduG is the only software that was made by primarily utilizing the power of modern operating systems using Graphic User Interface. And yet, it is rare to find an article explaining the use of EduG within a medical education topic.

Since G-Theory is the gold standard of reliability analysis for OSCE, therefore the objective of this paper is to explain the procedures of conducting an analysis using Generalizability Theory on one type of assessment tool, Objective Structured

Clinical Examination (OSCE), using the EduG software. In this project, the scenario is to be set to the minimal complexity in order to avoid confusion, whilst still able to demonstrate the power of G-Theory.

2. Experimental Design

2.1. The assessment

In this paper, the author uses the simplest OSCE scenario to give a better understanding towards the basic principles of using EduG to analyze OSCE and draw a conclusion from the analysis afterwards.

Any terms regarding the OSCE, is referred to in the AMEE Guideline.^{5,6} The scenario of the OSCE is a 10 stations exam per circuit. Since the exam must only must only take up a single day, there is a need of more than one building. Three buildings in total are used in this exam within the university grounds. Each building is administered by one circuit of OSCE. The participating students are randomly assigned to one of the circuits.

2.2. Analysis

The G-Theory analysis and Decision study (D-study) was performed using EduG version 6.1-e. The data was prepared using LibreOffice Calc and SPSS v15. The ANOVA was calculated using SPSS.

3. Result and Discussion

3.1. Result

The EduG software was made by the Swiss Society for Research in Education Working Group. Edometrics - Quality of measurement in education. This software is available for free at <https://www.irdp.ch/institut/english-program-1968.html>, though it is only available for Windows® OS. For this paper, the writer used Linux and ran the EduG within Windows using a virtual machine.

3.1.1. Study Design

The first step in running G-Theory analysis using EduG is deciding the measurement design. This step is important and the most confusing for novice users attempting to run the G-Theory analysis for the first time. Imagine the number of possibilities for an error to occur that influences the assessment. This error would then translate into facets. However, not all errors are able to be calculated since it is possible that the information related to the assessment is not available. Therefore, to decide what facet to include in the calculation design is a compromise between the data availability and the source of error we want identify. Ideally, the design should include all the possible errors in calculation, since the more detailed the information, the easier it would be to pin point the exact source of error. However, as the purpose of this paper to introduce G-Theory, the writer will limit the facet to minimal of 3 facets. It is not possible to include only 2 facets in calculation, since we would not be able to take advantage of the G-Theory and it would basically be the same as using CTT.¹

The focus of this analysis is to identify the error that influences the process of producing a score in the assessment. Hence the main purpose that the students become the object of study. The students become facet differentiation.

According to the OSCE scenario, there are 10 stations for each circuit. Each station contains specific test objectives. In this case, the stations are the main suspects of the source of errors. In CTT, the internal consistency of stations are commonly measured using Cronbach Alpha.^{1,7} It is for this specific reason that the stations are defined as facet instrumentation.

Turning now to the test locations. In this scenario, there are 3 buildings that were used as test locations. With 3 different locations comes the probability that each building could have a different effect to a student's performance, since there will be different rater for each station (i.e. station 1 in building A, B, and C). Due to this matter, we will count the building as a facet instrumentation.

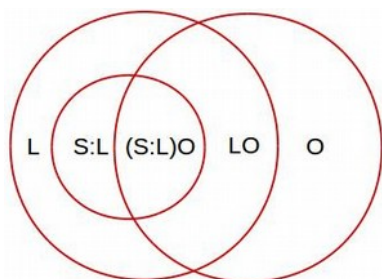


Figure 1. Variance Partition Diagram for the (S:L)O model, where S, L, and O represent Students, Locations, and Observation stations, respectively. All facets are random

The students were assigned to only one out of the three locations, therefore they would inherit the unique environment of the location they were nested to. It is from this condition that

we can expect an error to arise. Note that in this paper the stations are represented by the letter O and the letter S used to represent the students.

3.1.2. Data Preparation

As shown in Figure 2, the OSCE results should be managed as the illustration above. The ID of a student is the column the very left followed by the score the student pertained for each station in the column to its right. The next are the next students.

Readers might question the unbalanced data. EduG is only capable of calculating balanced data, therefore, unbalanced data needs to be balanced beforehand. For those who are familiar with SPSS, we can utilize a "stratified random sample" menu to create a balanced data from several groups of data or use a spreadsheet program as shown as in writer's blog (<http://pakwinny.staff.umy.ac.id/2017/05/25/random-discharge/>).⁸ In many circumstances, balancing the data of an unbalanced data is acceptable and the discussion regarding balancing data was explained by the creator of this software in EduG book.⁹

Score	Student ID
74.8	Stud_001
99.3	Stud_001
69.5	Stud_001
60.2	Stud_001
73.3	Stud_001
65.4	Stud_001
54.1	Stud_001
63.6	Stud_001
66.9	Stud_001
59.6	Stud_001
70.8	Stud_002
72.9	Stud_002
49.4	Stud_002
59.8	Stud_002
75.2	Stud_002
54.5	Stud_002
62.8	Stud_002
56.3	Stud_002
62.5	Stud_002
55.0	Stud_002

Figure 2. Illustration of the data in long format to be use as an input in EduG

Before proceeding to calculate in EduG, it is important to note that in EduG, only station marks need to be exported in long data format, as illustrated in Figure 3. There are two ways to change the data format from wide to long data format that were used in our data preparation:

First, for those who familiar with SPSS, menu Restructure is able to produce data needed for EduG.
Second, by using VBA script in MS Excel as provided below.

```

Sub RangeToColumn()
Dim varray As Variant
Dim i As Long, j As Long, k As Long
Application.ScreenUpdating = False
k = 1
varray = Range("C2:F61").Value
For i = 1 To UBound(varray, 1)
For j = 1 To UBound(varray, 2)
Cells(k, 9).Value = varray(i, j)
k = k + 1
Next
Next
Application.ScreenUpdating = True
End Sub

```

The VBA script above served the writer well. In order to use it, the reader needs to know how to use VBA script in excel. Edit the parameters to match with the reader's data :

First, replace line number 6, `Range("C2:F61")`, with your own data range. Afterwards, change line number 9, `Cells(k, 9)`, with the column (k) number in which you are prepared to write the results in.

Once the data is ready, save into .txt or .csv format. This file is ready for EduG.

A	B	C	D	E	F	G	H	I	J	K	L	M
ID	Location	Total Score	Station1	Station2	Station3	Station4	Station5	Station6	Station7	Station8	Station9	Station10
Stud_001	Building A	68.4	74.8	99.3	69.5	60.2	73.3	65.4	54.1	63.6	66.9	59.6
Stud_002	Building A	61.3	70.8	72.1	49.4	59.8	75.2	54.5	62.8	56.3	62.5	55.0
Stud_003	Building A	61	75.3	80	40	57.6	64.2	57.0	54.7	58.1	55.5	52.8
Stud_004	Building A	57.9	65.3	68	40	58.7	60.1	54.6	53.7	46.3	57.9	52.5
Stud_005	Building B	65.3	70.4	68.5	67.3	69.3	49.7	71.9	61.6	55.7	75.4	71.7
Stud_006	Building B	63.4	68.0	70.6	60.9	75.0	63.1	65.1	52.3	57.4	60.3	67.1
Stud_007	Building B	63.2	75.8	64.3	70.8	65.1	56.6	46.6	54.8	66.5	76.2	67.0
Stud_008	Building B	61	62.2	67.2	49.8	69.1	56.2	68.6	62.9	60.8	54.1	60.8
Stud_009	Building C	70.2	81.4	71.2	68.6	69.3	71.1	69.8	52.0	80.9	63.7	86.0
Stud_010	Building C	67.5	79.1	66.2	53.9	73.6	62.8	71.7	49.1	79.0	75.5	70.2
Stud_011	Building C	65.4	81.9	64.7	67.0	73.0	70.0	59.6	54.6	48.6	81.8	64.0
Stud_012	Building C	60.6	75.6	65.6	56.5	47.9	57.5	59.6	50.4	64.1	77.5	62.7

Locations

Figure 3. Illustration of the raw data structure of the recorded OSCE's score.

3.1.3. Running The EduG

EduG can only run in Windows OS. Therefore, other OS users would require the use of a virtual machine. Start the EduG and the blank windows will appear. Create new file and choose the file directory in which this project will be saved. The EduG worksheet is then ready to be used. The next steps require filling in the parameter according to the OSCE conditions. It is important to note that before we can start to use EduG, we have to decide the study design and prepare the file, since different design will require different file structures.

Fill in the parameters in EduG according to Table 1. Note that the students are nested within the location. The students were basically divided into 3 big groups, based on the buildings (A, B, C). This nested code should be written correctly in "Label".

Table 1. Data input to EduG according to study design

Facet	Label	Level	Universe
Locations	L	3	3
Students : Locations	S:L	30	INF
Observation	O	10	INF

Before importing our data, it is mandatory to decide the measurement design, as can be seen in Figure 4, the Measurement Design is S:L/O. The Differentiation facet was located in the left side of "/". When we have more than one facet on each side, it doesn't matter which is labeled first, since it is reversible. The important thing is its positioning, either to the left or right side of "/".

Do not make a mistake when imputing the "Level". It is basically the amount of each facet. The locations are 3, since we conducted the exam in 3 different locations. Are there only 30 students? NO. The actual total number of students (n) are 30 x 3 (remember, they are divided into 3 buildings). The total data value cells are 3 x 30 x 10 = 900. So, your input file should contain exactly 900 data values or when in excel, it will be 900 lines in one column. If those two numbers (total level and data input) do not match, EduG will give a warning and prohibit to the next step.

Having discussed how to setup the EduG, it is now time to input the data and run the analysis. There are two ways to input data, direct insert the data or import data from an external file. As described previously about data file creation, we will use that

file as the data source. Click the "import a file with raw data" button, then choose your previously prepared file. You are now ready to analyze the data. Once more, re-check to make sure your data structure is correct by confirming the data requirement from the "insert data" button. Remember, EduG is only able to check the amount of data, not the data structure.

Let us now turn to our main purpose, data analysis. Choose an .rtf file format as an output and after "compute", the file will open automatically. Close the file and run "mean" calculation to give a better view of the G-Theory analysis. Do not click replace. That will create a new file and your G-Theory calculation will disappear. Choose "Add" to add the mean calculation at the end of the file.

3.1.4. Calculation Results

The result consists of two parts, which in this paper is called Relative and Absolute part. In this paper, since we only concentrate on finding the source of errors, only the Relative part is shown.

Table 2. "Relative error variance" part of EduG calculation output. (The other part is "Absolute error variance part is not being use in this study design purpose).

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative
L	1.54		
S:L	26.51		
	O	
	LO	1.63	18.6
	SO:L	7.13	81.4
Sum of variances	28.05		8.76	100%
Standard deviation	5.3		Relative SE: 2.95963	
Coef_G relative	: 0.76			

It can be seen from the data in Table 2 that the G coefficient is 0.76. What is interesting about the data in this table is that 81.4 % of the error came from SO:L. One interesting result is that there is no source of error from the Observation (stations).

Whilst the mean from building A, B, and C are 57.41, 57.97, and 61.56, respectively.

3.2. Discussion

Before readers begin to analyze the results of the G-Theory, warnings were mentioned by Cardinet (2012) : Firstly, before studying G-Theory, readers should have an adequate knowledge of ANOVA. Secondly, there is no exact guide to interpret the results of the G table. An understanding of what we analyze and a lot of practice makes it perfect.^{1,9}

First things first, the G-coeff shows a quite high reliability (0.76) for an examination. Although the researcher mentioned that the good reliability for high-stake assessments should be 0.80 or above.^{3,9} These results also give another meaning that there is a 24% chance of error in this assessment. Using G-Theory, by grouping individual measurements, it is possible to pinpoint the source of errors.¹ Therefore, the effort to create a better examination in the future will be easier, since we can correctly fix the problem(s). Quantifying the source of errors cannot not be performed in CTT.^{3,10}

Lets focus on the 3rd and 5th column of Table 2. Due to there being no error in Observation, we can be assured that the station design (i.e. the rubrics and case scenario) is perfect. Moving on to the next line, the LO or stations in each building give off a little bit of high error. But remember, there is no error in the stations. So, what is the problem ? The answer is in SO:L, when a the factor of students come into the stations. It gave off a 81.4% error. Remember the assumption towards normal distribution when the students were randomly assigned to one of three locations? There should not be any differences in the assessment process of each building. The students should have equal capabilities and the results of the three locations should be more or less the same. But that was not the case in this project. So, what is the problem among the three facets? It isn't the location nor the students, but the distribution of the students to one of the locations.

Why are three factors needed to cause a big error ? We know that in OSCE, a rater is assessing the students. The problem of this OSCE is how the group of raters in one building rate differently compared to the other two groups. Consistency of the raters is known to be one of the main problems in OSCE, due to lack of briefing and training for the OSCE.¹¹

The proof that this is the source of error can be seen from the mean of building C (61.56) that is notably higher compared with the means of building A and B (57.41 and 57.97, respectively). In this case, we can suspect that there is a cause to this significant difference between the mean of building C with the mean of buildings A and B. Wait, do we have an F test in EduG? Even though EduG provides an ANOVA table, it does not provide an F test. The F test should be performed using another software.⁹ Writer performed an ANOVA, as can be seen in Table 3, and the Tukey comparison afterwards revealed that the score in building C is statistically significance higher than buildings A and B.

Table 3. Analysis of Varians (ANOVA) for the OSCE (compare the effect of locations and stations).

Effect	Sum of Squares	df	Mean Square	Significance
Locations	3043.598	2	1521.799	0.093
Stations	20516.423	9	2279.603	0.005
Location*Stations	10054.179	18	558.566	0.000

Another way to prove the error in building C is by running separate tests with level reduction. Each analysis is testing the reliability of the exam for each building. The results of G coefficient are 0.82, 0.84, and 0.68 for buildings A, B, and C, respectively. These results further support the idea that the rater is the source of error, since it is highly possible that the source of error caused low reliability in building C.


Further investigation by running the D-study or Optimization revealed the source of error is in line with that of previous studies. Increasing the number of stations will hypothetically increase the reliability of the assessment, that the reliability of 10 is lower than 15 stations and 20 is the highest (G= 0.76, 0.83 and 0.87, respectively).¹² Unfortunately, increasing the number of the stations will increase the cost of running an OSCE due to increment of the resources, equipment, staff, etc.^{13,14}

4. Conclusion

There are two findings in this project. Firstly, the G-Theory analysis was able to pinpoint the source of error. Secondly, this project confirmed that using EduG to performed G-Theory analysis in OSCE is not as difficult as most would expect it to be. Hopefully, the writing of this paper will motivate medical educators to help analyze future OSCEs.

References

- Bloch, R., Norman, G.: Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Med. Teach.* 34, 960–992 (2012). doi:10.3109/0142159X.2012.703791
- Chiu, C.W.-T.: Scoring Performance Assessments Based on Judgements: Generalizability Theory. Springer Science & Business Media (2001)
- Setyonugroho, W., Kropmans, T., Kennedy, K.M., Stewart, B., Dalen, J. van: Calibration of Communication Skills Items in OSCE Checklists according to the MAAS-Global. *Patient Educ. Couns.* doi:10.1016/j.pec.2015.08.001
- Patricio, M., Juliao, M., Fareleira, F., Young, M., Norman, G., Vaz Carneiro, A.: A comprehensive checklist for reporting the use of OSCEs. *Med. Teach.* 31, 112–124 (2009). doi:10.1080/01421590802578277
- Khan, K.Z., Ramachandran, S., Gaunt, K., Pushkar, P.: The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. *Med. Teach.* 35, e1437–e1446 (2013). doi:10.3109/0142159X.2013.818634
- Khan, K.Z., Gaunt, K., Ramachandran, S., Pushkar, P.: The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration. *Med. Teach.* 35, e1447–e1463 (2013). doi:10.3109/0142159X.2013.818635
- Streiner, D.L.: Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *J. Pers. Assess.* 80, 99–103 (2003). doi:10.1207/S15327752JPA8001_18
- Field, A.: *Discovering Statistics using IBM SPSS Statistics.* SAGE (2013)
- Cardinet, J., Johnson, S., Pini, G.: *Applying Generalizability Theory Using Edug.* Taylor & Francis (2012)
- Sudweeks, R.R., Reeve, S., Bradshaw, W.S.: A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assess. Writ.* 9, 239–261 (2004). doi:10.1016/j.asw.2004.11.001
- Besar, M.N.A., Siraj, H.H., Manap, R.A., Mahdy, Z.A., Yaman, M.N., Kamarudin, M.A., Mohamad, N.: Should a Single Clinician Examiner be used in Objective Structure Clinical Examination? *Procedia - Soc. Behav. Sci.* 60, 443–449 (2012). doi:10.1016/j.sbspro.2012.09.404

- 
12. Brannick, M.T., Erol-Korkmaz, H.T., Prewett, M.: A systematic review of the reliability of objective structured clinical examination scores. *Med. Educ.* 45, 1181–1189 (2011). doi:10.1111/j.1365-2923.2011.04075.x
13. Gormley, G.: Summative OSCEs in undergraduate medical education. *Ulster Med. J.* 80, 127–132 (2011)
14. Cusimano, M.D., Cohen, R., Tucker, W., Murnaghan, J., Kodama, R., Reznick, R.: A comparative analysis of the costs of administration of an OSCE (objective structured clinical examination). *Acad. Med.* 69, (1994)