Education and Science tedmem

Vol 39 (2014) No 175 285-295

Comparison of Balanced and Unbalanced Designs Based on Generalizability Theory via the Data of Intramuscular Injection Station [*]

Funda Nalbantoğlu Yılmaz [1], Ezel Tavşancıl [2]

Abstract

The purpose of this study is to compare the reliability and estimates of variance results by employing the balanced and unbalanced designs that were formed by rating the groups of participant students, both equal and different in numbers, considering their performance on the same task provided that the number of the participating students stays the same. The study group of the research are 240 first year students, who have taken the Intramuscular Injection Station in the Structured Objective Clinical Test of the Medicine Faculty of Hacettepe University. Eight raters was employed in assessing the performances of the students in the station. The variance values retrieved from the G studies done by each rater's rating the students equal in number (balanced) and those unequal in number (unbalanced) were found to be parallel in both designs. Morover, considering the G and Phi coefficients obtained from the senarios in the balanced and unbalanced (s:r) x t designs were observed to be close.

Keywords

Generalizability theory

Balanced design

Unbalanced design

Article Info

Received: 11.06.2012

Accepted: 10.23.2014

Online Published: 11.10.2014

DOI: 10.15390/EB.2014.2408

Introduction

Performance measurement, at all levels of education and in personnel selection, is significant for the assessment of skills that require cognitive and muscular coordination. It is especially crucial that performance on a particular skill be detected in the higher education programs where vocational skill sets for a given profession are taught.

Performance is required to finalize a tas kor process (Wiggins, 1993). It includes various activities such as essay writing, playing a musical instrument, making presentations. Thus student produces a protuct belonging to a performance condition or makes a relevant skills (Arias, 2010).

In medical training, one of the performance tests measuring how the students perform their clinical skills is the Objective Structure Clinical Examination (OSCE). This is a performance test comprised of multiple stations where the raters evaluate different skills in each of these stations. (Elçin, Odabaşı and Sayek, 2005). In OSCE, the stations are individual tests locations set up in order to measure the clinical skills of the candidate.

[*] This article was written as part of Funda Nalbantoğlu Yılmaz's PhD Thesis, completed through the Ankara University, in the Measurement and Evaluation Program.

[1] Nevşehir Hacı Bektaş Veli University, Faculty of Education, Department of Educational Science, Turkey, fundan@nevsehir.edu.tr

[2] Ankara University, Faculty of Educational Science, Department of Measurement and Evaluation, Turkey, etavsancil@gmail.com

OSCE, likewise all performance detection tools, are affected by various factors. These factors may stem from the environment where the examination is being conducted, test duration, the measurement tool or the strictness or generosity of the rater. The reliability and validity of the measurement results may differ due to above-mentioned deviations, hence the consistency of decisions which will be made accordingly. Therefore, it is significant that the reliability of the performance ratings be verified prior to usage.

Reliability in performance measurement is usually attained by methods based on classical test theory, item response theory and generalizability theory.

According to the classical test theory, reliability is how well the observed scores reflect the actual scores. The reliability index in the classical test theory is equal to the proportion of the variance of true scores to the variance of observed scores. Such a reliability index, however, is not practical since the true scores of the individuals are unknown. (Webb, Shavelson and Haertel, 2006). Therefore, the reliability coefficient in the classical test theory is determined indirectly by means of different methods based on various assumptions.

In cases where deviations deriving from different sources might affect the test results, it is necessary to evaluate its reliability by means of various reliability methods. In the classical test theory, there is at least one reliability prediction per source of error and this reliability varies according to the relevant sources of error. In the classical test theory, deviations affecting the test results may derive from different sources despite the fact that there is only one true score and non-distinctive error term (Cronbach, Gleser, Nanda and Rajaratnam, 1972). However, the classical test theory considers all sources of error as errors stemming from a single variable source and presumes that the sources of error do not affect one another. The generalizability theory, unlike the classical test theory, is comprised of a conceptual framework and method which deals with the multiple sources of error in the measurement process as well as the interaction of each of these error sources with one another (Brennan, 2001; Cronbach, Gleser, Nanda and Rajaratnam, 1972; Shavelson and Webb, 1991). Therefore, a rather comprehensive and realistic error description can be carried out via the generalizability theory in the performance assessment.

The generalizability theory is a statistical theory which enables the detection of the degree of error sources of the observed scores as well as the assessment and research of reliability in the measurement of behavior (Brennan, 2001; Cronbach, Glaser, Nanda and Rajaratnam, 1972; Shavelson and Webb, 1991). The generalizability theory has two basic functions, first of which is to evaluate the quality of the measurement process, second to make assumptions on how to increase its reliability (Wing and Chiu, 2001). In order to be able to conduct these studies, work patterns are designed first. And, according to the research data, these designs turn out balanced sometimes and unbalanced other times.

In the balanced design, the number of observations is equal at all levels of the variable (Brennan, 2001). For instance, in case the test is implemented on students in different schools, in the (p:s) x i design pattern (p: student, s:school, i: item) where the students are listed alongside schools and each student receives each item, participation by same of amount of students ensures a balanced situation. If the test contained three sub-tests and same amount of items are included in each sub-test, it would also mean that the design is balanced. However, in practice, unbalanced designs are more common. The reason why is that it is not always probable for the same amount of students to participate at the test in each school or the number of items to be equal in each of the sub-tests, if any. It was stated that the number of observations was the same as per each variable level in the balanced design patterns. However, the number of observation outcomes per variable may not be the same in all cases. Unbalanced designs emerge in cases of missing data or when the number of observations per variable level differs (Brennan, 2001).

The data will be cited as unbalanced in the event that the number of observations on the cells belonging to the conditions of a variable is not equal (Kaufmann and Schering, 2007). Certain data should be excluded in order to create balanced data out of data which the number of observation outcomes on all conditions of a variable is not equal. This is not a practical method, since a significant portion of data may be required to be excluded in certain cases which would then lead to much data loss. In such cases, instead of excluding data, it is necessary to calculate the variance components with unbalanced designs.

The literature review on the studies on the generalizability theory shows that most research conducted overseas utilizes balanced data structures besides studies where both structures including balanced and unbalanced are used together (Jeon, Lee, Hwang and Kang, 2009; Lee and Frisbie 1999; Ødegård, Hagtvet and Bjørkly, 2008; Sharma and Weathers, 2003; Wei and Haertel, 2011). Meanwhile, literature review on the studies on the generalizability theory in Turkey indicates that most research has a tendency towards comparison of the generalizability theory with the classical test theory and/or the Rasch model (Atılgan, 2008; Deliceoğlu ve Çıkrıkçı Demirtaşlı, 2012; Güler and Gelbal, 2010; Güler, 2011; Taşdelen, 2009; Yelboğa and Tavşancıl, 2010; Yılmaz Nalbantoğlu and Gelbal, 2011). And also, all of the studies are balanced designs. In addition, it is known that most unbalanced designs are more compatible with actual data. Under the implementation conditions, however, one may not always expect the data collected by the researchers to be in a balanced structure. Especially in case of tests such as the OSCE where multiple performance status of the students are measured and/or the number of students is high; employing multiple raters to assess the students' performance lead to certain constraint of time and manpower. Due to such restrictions, single rater is placed per station in the OSCE examination and the raters take turns at certain intervals. And, because the raters grade the students on a rotation basis; the total number of students graded by each rater varies and this results in an unbalanced structure as far as the number of students graded per rater. In this regard, it is asserted that this study will contribute to the implementation of the OSCE examinations by determining the reliability of the current implementations as well as by assessing the effect of the cases where raters grade equal and unequal number of students on reliability. In addition, the research is thought to contribute to investigate the reliability taking into account the interacton of error sources and effect of raters according to different scoring situation in the fine arts, physical education and so on. other researchers in different fields that decision are made based on performance. Thus, it is considered essential to carry out comparative studies because the number of studies utilizing the generalizability theory in Turkey is limited and that exemplary implementations utilizing designs in an unbalanced structure, which is a more common case in practice, are unavailable.

In this study, it is aimed to determine how the reliability is obtained from balanced and unbalanced designs, whether or not it affects the reliability coefficient calculated by unbalanced structure data according to the balanced structure data, whether or not differences will exist between the estimated variance components for designs. In this regard, the general objective of this research is to compare the results of the analysis of the generalizability theory with both the balanced and unbalanced designs, which are attained through rating of equal and different number of students on the same tasks, on condition that the number of participants remained the same, with the data pertaining to the implementation of Intramuscular Injections station as part of an OSCE examination, which is a test that evaluates the vocational skills of medical students. The following questions are asked for this end:

1. What is the percentage for explaining the total variance and variance components obtained through the generalizability theory via balanced and unbalanced (s:r) x t (s: student, r: rater, t: task) designs?
2. How do the G and Phi coefficients predicted from decision D-studies in accordance with scenarios set forward by increasing or decreasing the number of raters and students in both designs vary?

Method

Type of Study

In this study, the results of balanced and unbalanced designs comprised of data of clinical examination conducted to evaluate students' performance while the raters graded on a rotation basis are analyzed and the results of both designs are compared. From this aspect, the study is a basic research.

Study Group

The study group was comprised of 240 first year students who duly filled their assessment forms and sat at the Intramuscular Injection station in scope of the Objective Structure Clinical Examination conducted at the School of Medicine of Hacettepe University in the academic year 2010-2011. 12 raters were used for assess the performance of students in respective station under OSCE. The number of students varied for each rater. The data related to the station is unbalanced in terms of number of students for each rater. Aimed to comparison of the result obtained from balanced and unbalanced designs in the study. Therefore, eight raters were used in the study in order to create data structures for two design. The raters who graded the Intramuscular Injection skills of the students are assigned by the Department of Medical Training and Informatics and all raters are selected from the relevant field.

Research Data

In this research, the data obtained from the Intramuscular Injection station at the Objective Structure Clinical Examination taken by the students of the Department of Medical Training and Informatics at the School of Medicine of Hacettepe University in the first semester of the academic year 2010-2011 are used.

At an Intramuscular Injection station, students take turns to execute Intramuscular Injection on a medical mannequin. Each student is given equal amount of time at this station where a rater is ready to evaluate the skills of the student at that station. The rater, within the stipulated timeframe for him/her, evaluates the skills performed by the students who take turns at the station by using the Intramuscular Injection Execution Skill Evaluation Form consisting 17 tasks prepared by the Department of Medical Training and Informatics for this station. The rater who is placed at the station on a rotation basis leaves the station at the end of the allocated timeframe for him/her and is replaced by another rater at the station. Therefore, the raters at the station swap posts at certain intervals and each rater grades different group and number of students.

The balanced and unbalanced cases of the (s:r) x t designs – whereby the student (s) and rater (r) variables are nested and the tasks (t) which are same for all students are crossed with these variables – have been analyzed using the research data, each rater grading only part of and different groups of students on a rotation basis to attain generalizability theory analysis. The difference between the balanced and unbalanced (s:r) x t designs that are analyzed in scope of the study stems from the fact that the number of students graded by each rater is equal in a balanced design yet unequal in an unbalanced one. In case of the balanced (s:r) x t design, eight raters grade the performance of equal number of students i.e. 30 each on the 17 tasks. Meanwhile, in case of the unbalanced (s:r) x t design, out of 240 students; 23 students are graded by the 1st rater, 25 by the 2nd, 26 by the 3rd, 28 by the 4th, 30 by the 5th, 31 by the 6th, 36 by the 7th, and 41 students are graded by the 8th rater, respectively on the 17 tasks.

Analysis of Data

In the analysis of the data, the G_String (G-string-IV, Version 6.1.1.) interface software based on urGENOVA (Brennan, 2001), which is commonly used in the generalizability theory analysis of balanced and unbalanced data, was used.

Results

Below are the findings sorted according to the subgoals set forward in the framework of the general objective of the research:

The Percentage for Justifying the Total Variance and Variance Components Obtained via Balanced and Unbalanced Designs

Table 1 shows the percentage for explaining the total variance and variance predicted according to the (s:r) x t designs which are attained through rotational rating of equal and different number of students on the same tasks.

Table 1. The percentage for explaining the total variance and variance estimated from the balanced and unbalanced (s:r) x t designs

			Balanced (s:r) x t Design		Unbalanced (s:r) x t Design	
Source of Variation	Variance Component	df	σ^2	%	σ^2	%
t	σ^2_t	16	0.00141	2.6	0.00142	2.5
r	σ^2_r	7	0.00043	0.8	0.00032	0.6
s : r	$\sigma^2_{s:r}$	232	0.00199	3.6	0.00193	3.4
tr	σ^2_{tr}	112	0.00514	9.4	0.00541	9.6
ts : r	$\sigma^2_{ts:r,e}$	3712	0.04571	83.6	0.04712	83.8
Total		4079		100		100

s: student, r: rater, t: task

According to data shown on Table 1, the percentage for explaining the total variance by means of variance components of the tasks is %2.6 in case of the balanced (s:r) x t design, while %2.5 for the unbalanced (s:r) x t design. It is found that the predicted variance components from both designs with respect to the tasks are low. Therefore, one may assert that the tasks do not differ in terms of difficulty-easiness in both designs and that the tasks are equally challenging for students.

The variance component predicted for the rater main effect is estimated to be 0.00043 in case of the balanced (s:r) x t design, while 0.00032 for the unbalanced (s:r) x t design. Moreover, the percentage for explaining the total variance of the rater effect predicted through both designs is %0.8 in case of the balanced (s:r) x t design while %0.6 for the unbalanced (s:r) x t design. The predicted variance values and the percentages justifying the total variance are found to be low and close to one another. In this regard, one may assert that the rater changeability does not affect the performance of students in both designs likewise the attitude of the rater towards grading.

The percentage of the variance component of the (s:r) variable whereby the students are clustered with the raters in case of the balanced and unbalanced (s:r) x t designs attained through the raters grading students on a rotational basis on the same tasks are predicted to be 0.00199 (%3.6) for the balanced design, while 0.00193 (%3.4) for the unbalanced one. The outcome shows that the s:r variance predictions and their percentage for explaining the total variance are low and close to one another in both designs. This shows that the students' skills of administering Intramuscular Injection do not vary in either design likewise the attitude of the rater remains unchanged towards one student to the next.

The variance components pertaining to the interaction of the task x rater (txr) in case of the balanced and unbalanced (s:r) x t designs are predicted to be 0.00514 (%9.4) for the balanced design, while 0.00541 (%9.6) for the unbalanced one. These findings indicate that variance components in both designs are close to one another and that in both designs the effect of common factor of task x rater is higher in proportion to other variance components, except for surplus variance. Therefore, one may claim that differences exist in both designs due to the task x rater common factor.

Table 1 shows that in case of the balanced (s:r) x t design, where raters grade equal number of students, the residual variance's (σ^2(ts:r)) percentage of explaining the total variance is %83.6, meanwhile in case of the unbalanced (s:r) x t design, where raters grade unequal number of students, the surplus variance's (σ^2(ts:r)) percentage of explaining is at %83.8 of the total variance. In both designs, the surplus variance turned out high. The fact that surplus variance is high in (s:r) x t designs is an indicator that the student x task interaction, the differences stemming from student x task x rater interaction and/or other unknown sources of variables might be high. Therefore, one may claim that certain effects of the student x task interaction, the student x task x rater interaction and/or other unknown sources of variables exist in both designs.

Comparison of D-Studies in Both Designs by Increasing or Decreasing the Number of Raters and Students

The balanced and unbalanced cases compared in line with the objective of the research stem from the fact that the numbers of students graded by the raters are equal and unequal, respectively. In a decision study conducted in this respect, the tasks are determined to be the object of measurement in the rotational rating of students by different raters on the same tasks and they are considered to be the universe score in the (s:r) x t design where all variables are coincidental, meanwhile the raters and students are considered to be the universe of generalization. In this regard, in case of designs where raters grade equal number (balanced) and unequal number (unbalanced) of students, the G and Phi coefficients obtained from decision studies while increasing or decreasing the number of raters and students are shown on Table 2.

Table 2. Comparison of the G and Phi Coefficient Obtained from Decision Studies in accordance with Scenarios with Balanced and Unbalanced (s:r) x t Designs

Balanced (s:r) x t Design					Unbalanced (s:r) x t Design					
n_{s+}	n_r	$n_{s:r}$	G	Phi	n_{s+}	n_r	$n_{s:r}$	$\check{n}_r$	G	Phi
					120	6	16,18,20,21,22,23	5.92	0.520	0.507
120	6	20	0.533	0.515	120	8	10,12,14,15,15,17,18,19	7.73	0.562	0.549
120	8	15	0.579	0.563	120	8	13,13,15,15,15,15,17,17	7.93	0.568	0.556
120	10	12	0.612	0.596	120	10	8,9,10,11,12,12,12,14,15,17	9.55	0.592	0.580
240	6	40	0.574	0.556	240	6	33,38,40,41,43,45	5.95	0.562	0.548
240	**8**	**30**	**0.629**	**0.612**	**240**	**8**	**23,25,26,28,30,31,36,41**	**7.73**	**0.613**	**0.600**
240	10	24	0.667	0.651	240	8	26,28,30,30,30,30,33,33	7.96	0.618	0.605
360	6	60	0.589	0.571	240	10	15,17,20,22,24,26,27,28,30,31	9.56	0.648	0.636
360	8	45	0.647	0.630	360	6	55,58,60,60,62,65	5.98	0.578	0.565
360	10	36	0.687	0.672	360	8	38,40,43,45,47,48,49,50	7.94	0.636	0.623
					360	8	41,43,45,45,45,45,48,48	7.98	0.637	0.624
					360	10	22,26,31,31,35,36,38,39,45,57	9.36	0.664	0.652

n_{s+} : number of total studens n_r: number of raters, $\check{n}_r$: $n^2_{s+}/\sum n^2_{s:r}$

As shown on Table 2, the G coefficient obtained from the design balanced by eight raters each grading 30 students (n_{s+}=240) is predicted to be 0.629, while the Phi coefficient 0.612. In a balanced (s:r) x t design, the G coefficient obtained from the design balanced by raters grading unequal number of students once is predicted to be 0.613, while the Phi coefficient be 0.600. In case of designs attained by raters grading equal or unequal number of students, the G coefficient in the balanced design in proportion to the unbalanced is calculated to be 0.016, while the Phi coefficient is 0.012 higher than the unbalanced design. Therefore, it was found that coefficients from both designs using the implementation data are close to one another.

Table 2 shows that the G coefficient is predicted to be 0.618 in case of the unbalanced design in accordance with the scenarios where 240 students are graded by eight raters and the difference among the number of students graded by each rater is decreased in proportion to the unbalanced data (n_{s+}:240, n_r:8, $\check{n}_r$:7.96), while the Phi coefficient is predicted to be 0.605. These values are compared with the coefficient obtained from the balanced and unbalanced designs and it was found that they are lower (n_{s+}:240, n_r:8) than balanced design, while higher (n_{s+}:240, n_r:8, $\check{n}_r$:7.73) than unbalanced design. This means that the G and Phi coefficients show a tendency to increase when the variance among raters is decreased as far as the number of students graded per rater.

In case of both scenarios, while keeping the number of total students (n_{s+}:240) intact yet decreasing the number of raters (n_r:6), the G coefficient is predicted to be 0.574 in balanced design and the Phi coefficient be 0.556. Under the above-mentioned conditions yet in case of unbalanced design, the G coefficient is predicted to be 0.562 and the Phi coefficient be 0.548. The G coefficient in balanced design is predicted to be 0.667 and 0.648 in the unbalanced one, while the Phi coefficient 0.651 in the balanced and 0.636 in unbalanced designs where the number of raters is increased (n_r:10) on condition that the number of students remained intact. These findings show that the amount of increase by increasing or decreasing the number of raters have been lower compared to the implementation (n_{s+}:240, n_r:8) data. Besides, it may be asserted that the difference between the coefficient obtained from both designs where the number of raters are increased or decreased while keeping the total number of students the same and the each rater graded equal and unequal number of students.

Table 2 further shows that the G coefficient is predicted to be 0.579 in the balanced designs, while the Phi coefficient be 0.563 whereby the number of the raters remained intact (n_r:8) yet the students decreased (n_{s+}:120). In case of the unbalanced design (n_{s+}: 120, n_r:8, $\check{n}_r$:7.73); the G coefficient is predicted to be 0.562 and the Phi coefficient 0.549; and with the imbalance decreased as per the number of students graded by raters (n_{s+}: 120, n_r:8, $\check{n}_r$:7.93); the G coefficient is predicted to be 0.568, and the Phi coefficient 0.556, respectively. The G coefficient in proportion to the balanced design is predicted to be 0.647 and the Phi coefficient 0.630 where the number of raters remained unchanged yet the students increased (n_{s+}: 360); and in proportion to the unbalanced design (n_{s+}:360, n_r:8, $\check{n}_r$:7.94), the G coefficient is predicted to be 0.636 and the Phi coefficient 0.623, respectively. Moreover, the G coefficient is predicted to be 0.637 and the Phi coefficient 0.624 where the imbalance is decreased (n_{s+}: 360, n_r:8, $\check{n}_r$:7.98) as per the number of students graded by raters. Therefore, it was found that the coefficient from both designs, where the total number of students are increased or decreased while keeping the total number of raters intact, are close to one another.

The G coefficient is predicted to be 0.687 in the balanced (s:r) x t design, while the Phi coefficient be 0.672 while the G coefficient 0.664 and the Phi coefficient 0.652 in the unbalanced (s:r) x t design whereby the both number of the raters and the students are increased together (n_{s+}: 360, n_r:10). Accordingly, one may assert that the G and Phi coefficients slightly increase when the number of students and raters are increased together and that the coefficient in both design types renders close results.

In scope of the D-study in case of both designs, the G and Phi coefficients obtained from the unbalanced design attained through eight raters grading unequal number of students with a total of 240 on 17 tasks are found to be higher than the G and Phi coefficients obtained from the design where six raters graded a total of 120 students on the same tasks as before. The same situation occurred among coefficient where eight raters graded 360 students in an unbalanced design and eight raters graded 120 students, as well. This means that coefficient differ in both designs and that much data loss occur thus reliability decreases in case of designs where the reliability coefficient is calculated by reducing available data in order to balance the unbalanced design at hand.

Discussion, Conclusion and Suggestions

The results of the percentage for explaining the total variance and the variance obtained from the balanced and unbalanced (s:r) x t designs via G-study conducted at the Intramuscular Injection station in scope of an OSCE examination are as follows:

When the percentage for explaining the total variance and the variance components obtained from the balanced and unbalanced (s:r) x t designs via G-study conducted at the Intramuscular Injection station while each rater graded equal or unequal number of students are analyzed; it was detected that percentage for explaining the total variance and the variance components rendered similar results in case of both designs. This finding corresponds with the results of the research conducted by Shavelson and Webb (1981) claiming that the variance predictions would show similarity in case of balanced and unbalanced designs depending on whether the number of raters, coming from different geographical backgrounds and grading vocational skills, are equal or unequal. Furthermore, Sharma and Weathers (2003) also found that variance predictions are greatly similar in scope of their study which calculated the variance predictions obtained from the balanced design where equal number of participants is used from each country and unbalanced where different number of participants is used.

Considering the variance predictions on both designs, it is concluded that the tasks at the Intramuscular Injection station does not differ in terms of difficulty-easiness, the tasks are equally challenging for the students, the rater variability does not affect students' performance, the rater does not cause difference in terms of grading and the rater attitude does not differ towards one student to the next. The finding that the rater variability does not affect students' performance, the rater does not cause difference in terms of grading corresponds with the findings of the studies conducted by Yılmaz Nalbantoğlu and Gelbal (2011), Yılmaz Nalbantoğlu and Başusta Uzun (2012) where different skills at different stations are evaluated in scope of the Hacettepe University OSCE examination.

The results pertaining to G and Phi coefficients obtained from balanced and unbalanced (s:r) x t designs in accordance with the scenarios set forward by the decision study at the Intramuscular Injection station in scope of the OSCE examination are as follows:

As a result of the D-study, it is found that the G and Phi coefficients through both designs are not high. One may claim the reason is that the difficulty level of the tasks is not changed, the students' performance at the tasks does not vary and that the group is homogenous as far as the measured attributions are concerned.

The coefficient calculated according to the scenarios where the raters grade equal or unequal number of students of the same tasks on condition that the total number of data remains the same in case of the balanced and unbalanced designs renders similar results in both cases.

This finding shows consistency with the studies conducted by Jeon, Lee, Hwang and Kang (2009), Lee and Frisbie (1999), Malhotra and Sharma (2008), Shavelson and Webb (1981) where reliability coefficient obtained from balanced and unbalanced designs showed similar results in both designs. However, the D-study shows that the data used in this study is inherently unbalanced and balancing its unbalanced nature by means of data reduction leads to data loss thus difference in the reliability coefficient obtained from both designs and that a higher rate of reliability is attained in the unbalanced case. In addition, it was detected that the reliability coefficient shows a decreasing tendency as the number of students graded by each rater differs while the total number of students remains the same, in other words as the variability among the number of students graded by the raters increases.

Consequently, in scope of the study where psychomotor skills are commonly required; the reliability coefficient is found to be similar in cases where the raters graded equal (balanced) and unequal (unbalanced) number of students on a rotation basis with the total number of data equal; while data reduction or an increase in terms of imbalance leads to difference in the reliability coefficient in both designs.

The following suggestions are given in line with the results of the study:

The reliability coefficient shows slight difference in case the raters grade equal or unequal number of students, on condition that the number of data remains the same. The reason why the difference between the balanced and unbalanced situations is low might be the fact that the number of data remains the same in both designs. In case of a slight data loss, the unbalanced design renders higher reliability in comparison with the balanced one. Therefore, one must not forget that the reliability decreases if much data loss occurs in order to carry out generalizability theory analysis on balanced designs by means of data reduction. Therefore, it would be more appropriate to conduct analysis by using the unbalanced designs rather than calculating reliability through balanced designs which have been formed out of unbalanced designs by means of data reduction, if the data is in an inherently unbalanced structure.

It was found that the G and Phi coefficients show a decreasing tendency, as the variability between the number of students graded by each rater, while the number of students remains the same, in other words as the variability among the number of students graded by the raters increases. Therefore, one must avoid much difference, in terms of the number of students, among raters while evaluating students’ performance at the Intramuscular Injection station or any other stations in scope of OSCE examination. In addition, in the different educational researches that investigated the reliability of performance based, when the variability increases in the conditions of a facet in a unbalanced design it should be noted that the reliability is downward trend. Therefore, in the data collection phase the observations of conditions of facet should be kept close to each other as possible.

In another study that can be performed in different performance situations, comparison of balanced and unbalanced designs can be researched by using different sources of variability.

References

Arias, R. M. (2010). Performance Assessment. *Papeles del Psicologo*, *31*(1), 85-96.

Atılgan, H. (2008). Using Generalizability Theory to Assess The Score Reliability of The Special Ability Selection Examinations for Music Education Programmes in Higher Education. *International Journal of Research & Method in Education, 31*(1), 63-76.

Brennan, R. L. (2001). *Generalizability Theory.* New York: Springer- Verlog.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: Wiley.

Deliceoğlu, G., Çıkrıkçı Demirtaşlı, N. (2012). Futbol Yetilerine İlişkin Dereceleme Ölçeğinin Güvenirliğinin Genellenebilirlik Kuramına ve Klasik Test Kuramına Dayalı Olarak Karşılaştırılması. Güvenirliklerinin Karşılaştırılması. *Spor Bilimleri Dergisi*, 23(1), 1-12.

Elçin, M., Odabaşı, O., ve Sayek, İ. (2005). Yapılandırılmış Objektif Klinik Sınavlar. *Hacettepe Tıp Dergisi, 36,* 1-2.

Güler, N., ve Gelbal, S. (2010). Studying Reliability of Open Ended Mathematics Items According to Classical Test Theory and Generalizability Theory. *Educational Sciences: Theory and Practice, 10*(2), 989-1019.

Güler, N. (2011). Rastgele Veriler Üzerinde Genellenebilirlik Kuramı ve Klasik Test Kuramı'na Göre Güvenirliğin Karşılaştırılması. *Eğitim ve Bilim, 36,* 162: 225-234.

Jeon, M. J., Lee, G., Hwang, J. W. & Kang, S. J. (2009). Estimating Reliability of School-Level Scores Using Multilevel and Generalizability theory Models. *Asia Pacific Education Rev., 10*, 149-158.

Kaufman, J. & Schering, A. (2007). *Analysis of Variance ANOVA. Wiley Encyclopedia of Clinical Trials,* John Wiley & Sons, Inc.

Lee, G. & Frisbie, D. A. (1999). Estimating Reliability Under a Generalizability Theory Model for Test Scores Composed of Testlets. *Applied Measurement in Education, 12* (3), 237 -255.

Malhotra, M. K. & Sharma, S. (2008). Measurement Equivalance Using Generalizability Theory: An Examination of Manufacturing Flexibility Dimensions. *Decision Sciences, 39*(4), 643-669.

Ødegård, A., Hagtvet, K. A. & Bjørkly, S. (2008). Applying Aspects of Generalizability Theory in Preliminary Validation of the Multifacet Interprofessional Collaboration Model (PINCOM). *International Journal of Integrated Care, 8,* 1-11.

Sharma, S. & Weathers, D. (2003). Assessing Generalizability of Scales Used in Cross National Research. *International Journal of Research in Marketing, 20,* 287-295.

Shavelson, J. R. & Webb, N. M. (1981). Generalizability Theory:1973-1980. *British Journal of Mathematical and Statistical Psychology, 34,* 133-166.

Shavelson, J. R. & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Newbury Park. CA: Sage Publications.

Taşdelen, G. (2009). *Nedelsky ve Angoff Standart Belirleme Yöntemlerinin Genellenebilirlik Kuramı ile Karşılaştırılmasına İlişkin Bir Araştırma.* Yüksek Lisans Tezi, Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

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

Wei, X. & Haertel, E. (2011). The Effect of Ignoring Classroom-Level Variance in Estimating the Generalizability of School Mean Scores. *Educational Measurement: Issues and Practice, 30*(1), 13-22.

Wiggins, G. (1993). Assessment: Authenticity, Context and Validity. *Phi Delta Kappan, 75,* 200-214.

Wing, C. & Chiu, T. (2001). *Scoring Performance Assessments Based on Judgements: Generalizability Theory.* Kluwer Academic Publishers, Boston.

Yelboğa, A. ve Tavşancıl, E. (2010). Klasik Test ve Genellenebilirlik Kuramı'na Göre Güvenirliğin Bir İş Performansı Ölçeği Üzerinde İncelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 10*(3), 1825-1854.

Yılmaz Nalbantoğlu, F. ve Gelbal, S. (2011). İletişim Becerileri İstasyonu Örneğinde Genellenebilirlik Kuramı'yla Farklı Desenlerin Karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 41*, 509-518.

Yılmaz Nalbantoğlu ve Başusta Uzun (2012). "Genellenebilirlik Kuramıyla Dikiş Atma ve Alma Becerileri İstasyonu Güvenirliğinin Değerlendirilmesi" III. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi, Abant İzzet Baysal Üniversitesi, Bolu.