Asian Journal of Education and Training
Vol. 4, No. 2, 85-90, 2018
ISSN(E) 2519-5387
DOI: 10.20448/journal.522.2018.42.85.90

Using Generalizability Theory to Assess the Score Reliability of Communication Skills of Dentistry Students

N. Bilge Uzun[1]
Mehtap Aktaş[2]
Semih Aşiret[3]
Seha Yorulmaz[4]

[1,2,3,4]*Department of Measurement and Evaluation in Education, Faculty of Education, Mersin University, Mersin, Turkey*
[1]*Email: n.bilgeuzun@gmail.com Tel: 5332354008*
[2]*Email: mhtpaktas@gmail.com Tel: 05059184362*
[3]*Email: semihasiret@gmail.com Tel. 5072325248*
[4]*Email: sehayormaz@gmail.com Tel: 0554 829 3226*


(✉ Corresponding Author)

Abstract

The goal of this study is to determine the reliability of the performance points of dentistry students regarding communication skills and to examine the scoring reliability by generalizability theory in balanced random and fixed facet (mixed design) data, considering also the interactions of student, rater and duty. The study group of the research consists of the 16 dentistry students who took communication skills elective course in Hacettepe University in 2014-2015 academic year. In the study, the variability sources are taken as student (s), rater (r), occupation (o) and task (t). 16 students are assigned with 16 different duties and whether the students fulfill the duties is rated by 8 raters from two different occupation groups, 4 standard patients and 4 academicians. On these facets using different designs G study has been performed. Within the scope of this study the sxrxt design where all variability sources are crossed; sx(o:r)xt design where the occupation nested on rater as the result of scoring of the students by two different occupation groups as standard patient and academicians; sxaxt and sxspxt facets where rater facet is taken as fixed facet according to occupation groups (sp: Standard Patient, a: Academician) were examined. For each design estimated variance components are discussed separately.

Keywords: Communication skills, Performance evaluation, Reliability, Generalizability theory.

Contents

Citation | N. Bilge Uzun; Mehtap Aktaş; Semih Aşiret; Seha Yorulmaz (2018). Determinants of Job Satisfaction of Colleges of Education Lecturers: A Study of Nasarawa State College of Education, Akwanga. Asian Journal of Education and Training, 4(2): 85-90.
History:
Received: 29 January 2018
Revised: 16 February 2018
Accepted: 20 February 2018
Published: 23 February 2018

Publisher: Asian Online Journal Publishing Group

Contribution/Acknowledgement: All authors contributed to the conception and design of the study.
Funding: This study received no specific financial support.
Competing Interests: The authors declare that they have no conflict of interests.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study was reported; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained.
Ethical: This study follows all ethical practices during writing.

1. Introduction

In measurement studies the goal is to provide that the acquired score regarding the measured property to be as close as possible to the actual score. The proximity of the value obtained from measurement results to the true value of the measured property is obtained by capability of making error free measurements thus called reliability. Reliability coefficient takes various values between "0" and "1" and defined as a measurement of measurement results being free of random errors. Measurement results always carry some degree of error. As it will affect the accuracy of the decisions taken by the help of the measurement tools used to minimize this error, it shall be accurate to say that the reliability studies have a critical importance. In this regard, reliability has a critical importance also in performance based measurements in education activities. Performance is a concept which defines how well people use their own potential or actual knowledge, abilities in order to reach their goals or expectations (Helvaci, 2002; Erşangur, 2003). According to Deliceoğlu (2009) performance is the percentage of ability to use the capacity of a person in completing a job in a given time period with success.

The reliability coefficient is generally low in measurement of performance. In performance evaluation, the reason why this coefficient is low is that the scoring in general is done dependent on the rater judgements, in a word subjectivity. It is clear that the rater would not be the only source of error that affects the reliability in performance measurements because there are various sources of variability affecting these measurements. In general, there are different sources of variability in measurements based on performance. One of these sources of variability is the individuals. Individual source of variability is the abilities of individuals that show their talent or knowledge related to a problem, task or product. It is expected that the abilities of individuals evaluated would be different from one and other. In evaluations based on performance, raters are also a variability source. In performance evaluation raters have a very important place. This is because there are different approaches and responses of raters that affect their scoring. Another variability source is the task or item that the individual expected to fulfill. The items and tasks related to performance should be clear and accurate as to provide individuals to make their performances visible and also should carry the details that will bring up the differences between individuals. The number and parameters of tasks or items within the performance are among the properties that affect the performance of the individuals. In cases that the performance measurement is necessary, it is required to examine these variability sources that form the measurement and their interactions with each other. It is assumed that the measurable ability of an individual can be estimated by the performance he/she shows in the task requested to be performed. However, it should not be dismissed that these performances of individuals are in interaction with the raters being in different strictness levels or items and duties being in different difficulties (Güler, 2008). When these defined variability sources are considered, rater becomes an important source of error that decreases the reliability of the evaluation of performance, where interaction with factors such as duty or time can also be an error source as important. Thus it is not sufficient for rating reliability just to provide interrater or intrarater consistency and also the interaction between sources of error should be considered in calculation of reliability. However all the methods used for calculation of reliability of performance evaluation handle together the interaction both with rater and other error sources. The reliability of performance evaluation functions with methods based on the three basic theories of measurement, classical test theory (CTT), item response theory (IRT), and generalizability theory (GT) (Güler, 2008). In performance evaluation, one of the advantages that the generalizability theory provides over classical test theory is that it is possible to evaluate together the errors from many sources in generalizability theory (Brennan, 2001). Thus a more realistic error determination is made by generalization theory and it is possible to reach more realistic results and decisions.

When studies in literature are examined, it is seen that generalizability theory is used often in determination of reliability in performance evaluation (Ludlow, 2001; Barneveld, 2005; Lee and Kantor, 2007; Atılgan, 2008; Güler and Gelbal, 2010; Yelboğa and Tavşancıl, 2010; Yılmaz and Gelbal, 2011; Yılmaz and Başbaşa, 2015). Different measurement tools can be used in performance measurements in the field of medicine. Besides which ability is measured by which tool, the used tools to be valid and reliable is important and also it is required that the measurements taken should be reliable (Kurtz *et al*., 1998). The reliability of the measurement tools used in fields like medicinal training where especially the evaluations made often rely on performance evaluation should be as high as possible. In medicine training; patient-doctor interaction is a very important process that is in the center of the applications based on performance and where the doctor listens to complaints of the patient and help his/her problems (Good medicine application program booklet, p:10), communication in medicinal training is a professional ability and should be developed. The teaching of communication skills in medicinal training cannot be considered as a selective application. It should be taught as the other occupational abilities within the medicinal training.16 In this regard, on the subject of student evaluation in medicinal training, the third step of the clinical evaluation pyramid, performance evaluation (knowledge and ability, attitude to be shown with application under surveillance and within training environment) defined by Miller that is used often carries a critical importance. Moreover, the Communication Skills Evaluation that is in the scope of this study can be examined under this step. The communication skills course can be considered as a sort of preparation for real life conditions. Especially for performance weighted aspects, it is important to provide genuine conditions in these trainings. Thus, standard patients are used in these trainings, so that possibilities of diversification of interviews within a standard content, providing their repetition, creation of special cases and hard conditions can be offered (Good medicine application program booklet, p:155).

In fields like medicinal training where applications based on performance evaluation come forth, the studies done regarding the extensive reliability analysis of measurements via handling multivariable sources of error together are too scarce in our country (Yılmaz and Gelbal, 2011; Yılmaz and Başbaşa, 2015). When the performance is evaluated, it is considered that the extensive reliability analyses to be performed for used measurement tools and studies that handle multivariable sources of error to be done are important and this study will provide contribution to the literature in this regard. In the scope of this research, communication skills elective course is considered. Mentioned course consists of evaluation of patient interview of students in occupational sense.

In this study, it is aimed to have the ratings of performance evaluation tasks obtained as the result of rating of raters from more than one, different occupational origins to be examined by generalizability theory and to be

analyzed with multifacet fixed designs. Broadly, it separates obtained measurement results into different variance sources, evaluates and defines them, then provide the generalization of the observed ratings of the individuals subject to measurement to the universal ratings that correspond to actual rating concept in classical test theory.

Findings of this research are limited to sixteen dentistry students who took communication skills elective course in 2014-2015 academic year in Hacettepe University and eight raters coming from two different occupational groups that rate these students; and the data from communication skills checklist consisting of sixteen duties that aim to measure communication skills of these students.

2. Method

2.1. Research Design

This research consisting of determination of the reliability of the performance scores acquired in the practice examination in the communication skills elective course is a scanning type of research.

2.2. Study Group

The study group of this research is formed by the sixteen dentistry students who took communication skills elective course in 2014-2015 academic year in Hacettepe University. In evaluation of the performance conditions of the students' eight raters, four academicians and four standard patients who participate in trainings were used.

2.3. Data Collection Tools

In evaluation of the performance of the students regarding communication skills, the "Communication Skills Evaluation Form" prepared by Hacettepe University Medicinal Education and Informatics Department was used. In this evaluation form, there are sixteen tasks. Tasks were scored as whether the student had the related skill (1), or not (0).

2.4. Data Collection

In this context, each of the raters scored different numbers of students' recursively. Four academicians and four standard patients who participated in trainings, totally eight raters scored the 16 students for same tasks. In the study, student (s), rater (r), occupation (o) and task (t) were handled as variability sources. Sixteen students were assigned to sixteen different tasks and whether the students succeed the task or not was scored by eight raters from two different occupation groups which one is composed of four standard patients and other is composed of four academicians. On these facets, G study was conducted by using different designs.

Within the scope of this study, the sxrxt design where all variability sources were crossed; different from the all crossed design, sx(o:r)xt design where the occupation nested on rater as the result of scoring of the students by two different occupation groups as standard patient and academicians; rater facet from sxrxt design was taken as fixed facet with regard to occupation groups (sp: Standard Patient, a: Academician) and as a result sxaxt and sxspxt designs were examined.

2.5. Data Analysis

In the analysis of the data, eduG software was used. For each design, estimated variance components were discussed

3. Findings

The variance components estimated as the result of the G study conducted via sxrxt design totally crossed according to each of the 8 raters rating 16 students in line with 16 tasks are given in Table 1.

Table-1. Estimated Variance Components for oxpxg Design

Variance Source	Squares total	Freedom Degree	Squares Average	Corrected Components	%
s	18.851	15	1.256	0.007	2.8
r	13.933	7	1.990	0.006	2.3
t	181.069	15	12.071	0.091	35.2
sr	25.246	105	0.240	0.008	3.0
st	50.297	225	0.223	0.013	5.1
rt	37.152	105	0.353	0.015	5.7
srt,e	186.542	1575	0.118	0.118	46.0
Total	513.093	2047			100
G coefficient	0.73				
Phi coefficient	0.44				

As is shown Table 1, there are totally seven variance components in G study conducted with sxrxt design. When the estimated variance components of these variance sources are examined, the variance component of the main effect of the task variability source is 0.091 and explains 35.2% of the total variance. This shows that difference between difficulty levels of the tasks is considerably high (Shavelson and Webb, 1981; Cardinet *et al.*, 2010). The rate of definition of the estimated variance for relative student and rater main effect of the total variance shows that the group of students is a homogenous group regarding performance and there is no effect due to rater difference. Furthermore, it can be seen that the variance percentage of the residual component with the biggest variance is 46%. Hence, it can be said that 46% of the total variance is due to unexplained systematic or unsystematic errors.

The G and Phi coefficients of the totally crossed design are calculated as 0.73 and 0.44 respectively. When it is considered that the scoring are done in absolute, it can be said that the reliability could be low regarding the Phi coefficient.

The estimated variance components of the sx(o:r)xt design different from the all crossed design where the occupation nested on rater as the result of scoring of the students by two different occupation groups as standard patient and academicians are given in Table 2.

Table- 2. Estimated Variance Components for ox(m:p)xg Design

Variance Source	Squares total	Freedom Degree	Squares Average	Corrected Components	%
s	18.851	15	1.257	0.007	2.5
r	11.412	3	3.804	0.006	2.2
o:r	2.522	4	0.630	0.001	0.4
t	181.070	15	12.071	0.091	35.0
sr	15.299	45	0.340	0.005	1.7
st:r	9.947	60	0.166	0.004	1.5
st	50.297	225	0.224	0.011	4.2
rt	17.643	45	0.392	0.001	0.4
ot:r	19.510	60	0.325	0.014	5.3
srt	92.021	675	0.136	0.016	6.1
sot:r,e	94.522	900	0.105	0.105	40.6
Total	513.093	2047			100
G coefficient	0.66				
Phi coefficient	0.38				

When Table 2 is examined, as it is in the totally crossed design, it is seen that the second biggest variance component is the task component with 0.091. The variance of task variability source explains 35% of the total variance. Hence, it can be said that the difference between difficulty levels of the duties is high for this design. Similarly, the residual component variance with 40.6% is the most contributing one to the total variance. When components in totally crossed design and this nested design are compared, it is seen a nearly 6% reduction. It is thought that this reduction is due to "srt" common effect, however, the explained ratio of the total variance of the other components is seen to be relatively low. This situation shows that the design where the occupation is nested to rater against the totally crossed design does not reveal a big difference in estimated variance components. When compared to totally crossed design, it can be said that the G coefficient (0.66) and Phi coefficient (0.38) of this design is reduced.

"sxrxt" design where rater is taken as fixed facet according to occupation groups (sp: Standard Patient, a: Academician) is conducted as totally crossed two different designs as sxaxt and sxspxt and G study was performed. The obtained variance components of these two patterns are given in Table 3 comparatively.

Table-3. Estimated variance components for Fixed facet Crossed designs

Source of variance	Oxsxg			Oxaxg		
	Squares Average	Corrected Components	%	Squares Average	Corrected Components	%
s	0.467	0.004	1.7	1.139	0.012	4.8
r	0.730	0.002	0.8	0.678	0.001	0.2
t	7.715	0.117	47.0	5.064	0.072	27.9
sr	0.144	0.002	1.0	0.301	0.011	4.1
st	0.160	0.014	5.6	0.186	0.014	5.3
rt	0.181	0.005	1.9	0.409	0.017	6.7
srt,e	0.104	0.104	42.0	0.131	0.131	51.0
G coefficient	0.57			0.69		
Phi coefficient	0.27			0.54		

When Table 3 is examined, it is seen that while the variance component of students is 0.004 in sxrxt design and it is 0.012 in sxaxt design. In this case, it can be said that when academicians scored students occupies, difference between students'scores is greater. The percentage of explaining total variance of the task variability source is 47% for sxspxt design and 27.9% for oxaxg design. It can be said that the difference between difficulty of the tasks is greater than when the patients score. The percentage of explaining total variance of the rater-student common effect for sxrxt design is 1%, and for sxaxt design it is 4.1% and it can be explained as the interaction between the academician and the student is relatively more than the interaction between the patient and the student. Similarly, when "rt" common effect is examined in Table 3, it can be inferred as that the standard patients compared to academicians differ less in rating tasks. However, it should also be considered that the contribution the estimated variance of this common effect to the total variance is relatively low. It is seen that the percentage of definition of the total variance of the residual variance source for sxspxt design (51%) is higher than that of the sxaxt design (42%). It can be inferred that the residual variability source is considerably big in fixed facet designs as in the random facet designs.

When reliability coefficients are compared for both with fixed facet designs, the score reliability of academicians (Phi coefficient = 0.54) is higher than the score reliability of the standard patients (Phi coefficient = 0.27).

4. Discussion, Conclusion and Implications

The performances of dentistry students regarding OSCE Communication skills have been examined in the frame of generalizability theory using multifacet designs and the results obtained are specified below.

Students are taken as the measurement object in totally crossed and nested designs. For both designs, the estimated variance component of students has very low effect in explaining the total variance. When the related literature is investigated, there is a common idea that the measurement object should have a significant effect in

explaining the total variance. However, in the literature, there are studies which report that variance percentages of the measurement object is low in the cases that the measurement object does not differ regarding the measured attributes.19 In this study, it is concluded that the students do not differ significantly in performances of students regarding the communication skills are succeeded.

When the studies conducted in all designs are examined, it is inferred that the variance of the task component is too large. This inference indicated that the tasks for the communication skills differ in regarding the difficulty levels with itself. This finding is consistent with findings what has been reported by Yılmaz and Gelbal (2011) and Kozan *et al.* (2015). In this and similar studies, the findings related to tasks indicate that the weighted scoring of the tasks that form the skill in fields like Medicine can be more suitable. Similarly, the control lists, as they do not give information about to what degree the performance levels or criteria are met (Moskal and Leydens, 2000) in field specific studies, it is suggested to the field experts that the rubrics that facilitate to inform about the level of the performance to be developed.

The studies in the literature, it is reported that the rater variability source main effect is relatively low effect on explaining total variance and concluded that raters are consistent in scoring the students with regard to tasks internally. These results are parallel with the results from the studies in which the different abilities within Hacettepe University OSCE examination are evaluated of Yılmaz and Gelbal (2011); Yılmaz and Tavsancıl (2014) and Yılmaz and Başbaşa (2015).

Unexplained variance is too large in all designs. This situation reveals the requirement that the different variability sources should be included in the designs.

When the raters with respect to the occupation groups are taken as fixed, according to the scoring by academicians, the duties are differed more regarding the difficultness-easiness in cases the students are rated by standard patient. When residual effects are examined according to the occupation groups, the unexplained variance ratio revealed higher in cases that the academicians score to task of the students. One reason of this difference is thought to be due to the differentiation between tasks. Other reason of this difference is due to the academicians participating in scoring are experts in field of medicine. Field experts evaluating these tasks in more equivalent qualities and giving similar importance to all the tasks forming the skill could have increased this difference.

Overall, it is said that Phi coefficients in all designs are low. One reason of this is that unexplained variance is large thus it has too much errors. Another reason is that the task variability source has a large variance value. Beside these, putting out the differences between performances of communication skills of students who are measurement object is also a factor which decreases the reliability. However, it is concluded that the reliability is higher in fixed facet studies when the academicians score the students. It is thought that the reason why academicians evaluated the students more reliably is due to the doctor academicians acting more consistently in sorting tasks and scoring.

Researches can perform similar studies to this one where different raters are used, different variability sources are considered in different conditions with mixed designs. At the same time in performance evaluation the generalizability theory is important as it can perform a more realistic and extensive reliability analysis via evaluating many error sources and their interactions together and therefore in order to perform extensive reliability analysis, to collect information on student performances, evaluation, tasks, raters and to perform decision studies that shed light to future studies, it is suggested that the generalizability theory to be used in performance evaluation studies.

References

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. *View at Google Scholar | View at Publisher*

Barneveld, C.V., 2005. The dependability of medical students' performance ratings as documentedon in-training evaluations. Academic Medicine, 80(3): 309-312. *View at Google Scholar | View at Publisher*

Brennan, R.L., 2001. Generalizability theory. New York: Springer- Verlog.

Cardinet, J., S. Johnson and G. Pini, 2010. Applying generalizability theory using EduG. USA: Taylor and Francis Group, LLC.

Deliceoğlu, G., 2009. Comparison of the rating scale of football competence based on generalizability and classical test theory. Unpublished PhD Thesis, Ankara University, Ankara.

Erşangur, C., 2003. An implementation of the participatory management to increase the performance of employees in the operation and to evaluate their performance and in Niğde industry enterprises. Graduate Thesis, Niğde: Niğde University Institute of Social Sciences.

Güler, N., 2008. A research on classical test theory generalizability theory and rasch model. Unpublished PhD Thesis, Hacettepe University, Ankara.

Güler, N. and S. Gelbal, 2010. Studying reliability of open ended mathematics items according to classical test theory and generalizability theory. Educational Sciences: Theory and Practice, 10(2): 1011-1019. *View at Google Scholar*

Helvaci, M.A., 2002. The importance of performance evaluation in performance management process. Ankara University Journal of Educational Sciences, 35(1-2): 155-169.

Kozan, S., A. Uçar, E. Balta, N.B. Principal and O. Odabaşı, 2015. The determination of the suitable condition numbers to be used in the assessment of communication skills of the dentist students by the generalization theory. The Congress of Clinical and Communication Skill Training in Health Sciences, Hacettepe University, Ankara.

Kurtz, S.M., J.D. Silverman and J. Draper, 1998. Teaching and learning communication skills in medicine. Oxford: Radcliffe Medical Press.

Lee, Y.W. and R. Kantor, 2007. Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G theory. International Journal of Testing, 7(4): 353-385. *View at Google Scholar | View at Publisher*

Ludlow, C.B., 2001. Using running records as a benchmark reading assessment: Reliability in assessing reading progress. The Degree of Doctor, BrighamYoung University, Provo.

Moskal, B.M. and J.A. Leydens, 2000. Scoring rubric development: Validity and reliability. Practical Assessment, Research & Evaluation. Retrieved from http://PAREonline.net/getvn.asp?v=7&n=10 [Accessed September 08, 2016].

Shavelson, J.R. and N.M. Webb, 1981. Generalizability theory:1973-1980. British Journal of Mathematical and Statistical Psychology, 34(2): 133-166. *View at Google Scholar | View at Publisher*

Yelboğa, A. and E. Tavşancıl, 2010. An examination of reliability based on classical testing and generalizability theory on a business performance scale. Theory and Practice in Educational Sciences, 10(3): 1825-1854.

Yılmaz, N.F. and N.B. Başbaşa, 2015. Assessment of sewing and picking skills station reliability with generability theory. Journal of Measurement and Evaluation in Education and Psychology, 6(1): 107-116.

Yılmaz, N.F. and S. Gelbal, 2011. Comparison of different patterns with generality theory in the case of communication skills station. Hacettepe University Journal of Education Faculty, 41: 509-518.

Yılmaz, N.F. and E. Tavsancıl, 2014. Comparison of balanced and unbalanced patterns in generalizable theory with intramuscular injection station data. Education and Science, 39(175): 285-295. *View at Google Scholar | View at Publisher*

Bibliography

Kara, Y. and H. Kelecioğlu, 2015. Examining the effect of scoring qualities on the determination of cutting scores by using generalizability theory. Journal of Measurement and Evaluation in Education and Psychology, 6(1): 58-71.