USING GENERALIZABILITY THEORY IN RELIABILITY ESTIMATION OF MEASUREMENTS OF HIGHER-ORDER COGNITIVE SKILLS

Abstract

The purpose of this research is to estimate reliability with Generalizability Theory in assessing high-order cognitive skills of a particular performance. For this purpose, a performance task belonging to science and technology lessons was applied to 50 students and the performances of the students were graded by 10 teachers with a rubric, prepared for the same aim one of which was analytic and the other was holistic. The collected data were analyzed by G-Theory using the EDU-G statistical package program. G study done with students, raters, forms were created by means of variability sources \((s x r x f)\) 2 facet crossed design and D studies were conducted with the created scenarios. According to the G study results, it was seen that the students- who were the objects of measurement- were the greatest variability source, and the other influences stemmed from the raters and forms respectively. Based on the results of D-study, 6 raters and 1 form yielded the optimum results.

Keywords: Inter-rater reliability, Generalizability Theory
ÜST DÜZEY ZİHİNSEL ÖZELLİKLERİN ÖLÇÜMLERİNİN GÜVENİRLİK KESTİRİMİNDE GENELLENEBİLİRLIK KURAMININ KULLANIMI

Öz

Anahtar kelimeler: Puanlayıcılar Arası Güvenirlik, Genellenebilirlik Kuramı

1. INTRODUCTION

Students are expected to answer the questions in the measurement tool administered to them in a limited period of time in classroom settings without using other resources in measurements performed in traditional approaches of measurement. However, such measurement results do not prove that students act in compliance with the responses they give in real life situations. For instance, the behaviors students are expected to gain as a requirement for the knowledge, skills and efficacies in lab applications course can be measured through a multiple-choice test prepared for the course. High marks received from the test by students can indicate their knowledge but the situation does not indicate for certain that they have the skills and efficacy to be able to do the experiments appropriately. In a similar vein, traditional pen-and-paper tests are inadequate in measuring the complicated, gradual and applied properties to be instilled in students in such courses as music, drawing and physical education. In such cases, what students can do rather than what they know is the focus of interest. Performance-based assessment in such cases came to the fore. Performance-based assessment is used in several fields such as finance, economy (Saçıcı & Kaya, 2014). For field of education, the meaning of performance based assessment differs a bit. Assessment of situation which is performance-based gains currency as an alternative to traditional assessment approaches such as pen-and-paper tests which are restricted in measuring the properties which focus on how students are to behave in real life situations, which are complex and which require application.

Performing measurement and assessment based on the way individuals are expected to behave in real life circumstances is desired in performance-based assessment (Filzpatrick & Morrison, 1971). The process of assessment - which focuses on how individuals behave in learning environments and on what they do – considers the concept of performance very important. Kutlu, Doğan and Karakaya (2009) define performance as the effort in performing the...
tasks based on higher-order cognitive skills, and as the output in consequence of the effort. It is expected that performance is versatile, that students can apply it in real life situations by using higher order cognitive processes, and that it requires skills and abilities beside knowledge (Wiggins, 1993; Kutlu et al., 2009). In such evaluation, the focus is on how students will use their higher order cognitive skills in real life situations (Airasion, 1994).

A performance task is described, it is assigned to students, it is fulfilled by students, and the process as well as the outcome is scored by raters in a process of performance-based assessment to be conducted in an educational environment. Checklists and rubrics suiting to the structure of the performance are employed in such performance-based assessment activities as different from classical measurement tools (Airasion, 1994).

Rubrics are the scoring schemata containing the criteria for a particular task and the degree of the properties for the criteria (Brookhart, 1999; Andrade, 2000). On defining rubrics in terms of the process of performance-based assessment, they can be defined as rating tools in which the criteria expected in performing a particular performance task and the degree of the criteria in students’ achievement are presented with explanations (Kutlu et al., 2009). Rubrics offer benefits such as expressing clearly the tasks assigned to students, students’ becoming aware of the behaviors expected of them, giving detailed feedback to students, and distinguishing performance in a valid and reliable way (Andrade, 2000; Wiggins, 1993). Rubrics are divided into two as analytic and holistic (Merter, 2001).

Analytic rubrics assess the qualities required by a performance task by dividing them into sub-categories. The achievement levels students are expected to attain are described in details and are scored in the sub-categories. The achievement scores received from the sub-categories are added up and thus scores for the performance are calculated. Holistic rubrics, on the other hand, give scores without dividing the performance into parts (Merter, 2001; Nitko, 2001; Moskal, 2000).

It is demanded that the scores received from the rubrics used in performance-based assessment are reliable – that is to say, they should be error-free. Reliability of such instruments is achieved through raters’ opinions (Kutlu et al., 2009). Reliability estimations based on raters’ opinions are described as inter-rater reliability in the literature, and are actualized through many techniques such as correlation, variance analysis, fit coefficients, log-linear analysis, etc. Reliability estimation through the above mentioned techniques makes estimations by regarding only one source of error as the source of variability. Generalizability Theory, which considers diverse sources of error- that is to say, sources of variability merging into the measurement process- in addition to the knowledge provided by these techniques and which takes into consideration the interactions of variability sources, has been used in recent years in performance-based situation determining studies (Brennan, 1996; Brennan & Johnson, 1995; Gao, Shavelson & Baxter, 1994; Çakıcı Eser & Gelbal, 2014; Naumenko, 2015; Brennan, 2000).

Generalizability Theory is based on studies performed by Cronbach, Nanda & Rajaratnam (1972), Feld & Brennan (1989), and Shavelson & Webb (1991). The theory differs from Classical Test Theory, which is based on only one source of variability in reliability estimations, and it considers the contributions of multiple sources of error to error in one single analysis (Brennan, 2000). Multiple sources of error are analyzed on the basis of variance analysis. With the variance analysis of such variability/ error sources as items, forms, raters, etc., how influential they are in an error is evaluated. Two separate studies are conducted in reliability estimations in Generalizability Theory.
The first study for the purpose of generalization is the G study, and the second is the D study which aims to make a decision on the basis of the G study. Studies have crossed and nested designs based on the sources of variability. This study employs crossed design since all conditions overlap.

Sources of variability should be determined in Generalizability studies. Because the basic aim of measurement is to evaluate the differentiation between individuals in terms of the measured quality, individuals are the object of measurement. In this context, individuals are not considered to be a source of error in most G studies. Abundance of variability stemming from individuals is an expected situation. The main source of error is the sources of variability such as items, forms and raters which are present in the measurement process, which are not desired but which cause variability. Since a property of individuals is measured in this study, the object of measurement is individuals and the sources of variability are raters and the analytic and holistic rubrics used. The study, which is to be conducted in crossed design with relevant sources of variability, searches for answers to the following questions:

1. What are the results for the G study performed with students, raters and rubrics (sxrxf) in crossed design?
2. What are the G and Phi coefficients obtained from scenarios created by changing the number of raters and rubrics?

**Purpose of Study**

This study aims to estimate the components of the variance in the total variance for the students, raters and rubrics (s x r x f) design with the G study. And it aims to analyze the effects of changes in the number of raters and rubrics on the results with the D study.

2. **METHOD**

2.1. **Type of Study and the Study Group**

This study is based on conducting the G and D studies in Generalizability Theory with real data and on the analysis of the results. It is a descriptive study in this respect.

A pre-defined performance task was assigned to 50 primary school students, and the performance of the students fulfilling their task was graded by 10 teachers at two different times with rubrics one of which was analytic and the other of which was holistic. Thus, the study group was composed of 50 primary school students and 10 primary school teachers.

2.2. **Data Collection Tools and Data Collection Process**

This study was conducted with a performance task and two different rubrics. The performance task prepared by Kutlu, Doğan and Karakaya (2009) in relation to the unit of “Let’s Do the Puzzle of Our Body” in the course book for 4th grade Science and Technology course was assigned to the students. The list of evaluation criteria in addition to the detailed description of the performance task was distributed to the students. Detailed explanation was offered to the students about the performance task and the process of grading, and the students were allowed 10 days to do the task.

10 teachers were requested to perform the grading in this study. They graded the students at different times. Besides, each teacher was asked to grade the students by using two rubrics one of which was analytic and the other of which was holistic. A two-week interval was left between the two ratings in order to eliminate the effects of recalling.

Of the two rubrics used in the study, the analytic rubric was created on the basis of the rubric prepared by Kutlu, Doğan and Karakaya (2009) in relation to the performance task they
developed. The final shape was given to the rubric by consulting the opinions of 5 measurement and evaluation experts. The sub-categories such as scope, content, use of materials, use of time, creating graphs and creating tables were included in the measurement tool/rubric. Although each sub-category was scored between 1 and 3, the maximum score receivable was 18 while the minimum score receivable was 6. Holistic rubric was developed by the researchers based on the generalization of criteria in the sub-categories of the analytic rubric for the same purpose. Whereas the maximum score receivable from this tool was 4, the minimum score was 1.

2.3. Research Design

This study used crossed design with the students, raters and rubrics (sxrf) sources of variability. The design is shown visually in Table 1.

Table 1. Research Design and the Sources of Variability

<table>
<thead>
<tr>
<th>Sources of variability</th>
<th>Types of variability</th>
<th>Research design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students (S)</td>
<td>Object of measurement</td>
<td></td>
</tr>
<tr>
<td>Raters (R)</td>
<td>Raters’ inconsistencies in rating the students</td>
<td></td>
</tr>
<tr>
<td>Forms (F)</td>
<td>Inconsistencies between rubrics</td>
<td></td>
</tr>
<tr>
<td>S x R</td>
<td>Effects of students and raters</td>
<td></td>
</tr>
<tr>
<td>S x F</td>
<td>Effects of students and rubrics</td>
<td></td>
</tr>
<tr>
<td>R x F</td>
<td>Raters’- rubrics’ effects</td>
<td></td>
</tr>
<tr>
<td>S x R x F.e</td>
<td>Residual effects/random errors</td>
<td></td>
</tr>
</tbody>
</table>

G coefficient $E^2_F = 0.97$ Phi coefficient $E^2_\phi = 0.96$ 100%

2.4. Analysis of the Data

EDU-G 6.0 statistical package was used in the analysis of the data. Raters’ variance component, rubrics’ variance components G and Phi coefficients were estimated in the analysis performed through crossed design (sxrf) obtained by grading 50 students by 10 teachers according to 2 rubrics.

3. RESULTS

The findings are presented in the order of research questions. First, the findings on the question “What are the results for the G study performed with students, raters and forms/rubrics (s x r x f) in crossed design?” are presented. Table 2 sums up the results for ANOVA conducted to analyze the variance components in the total variance.

Table 2. ANOVA Results

<table>
<thead>
<tr>
<th>Sources of variability</th>
<th>Squares total</th>
<th>sd</th>
<th>Squares average</th>
<th>Estimated variances</th>
<th>Total variance percentages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students (S)</td>
<td>21804.1</td>
<td>49</td>
<td>444.9</td>
<td>21.6</td>
<td>67.6</td>
</tr>
<tr>
<td>Raters (R)</td>
<td>1383.6</td>
<td>9</td>
<td>153.7</td>
<td>0.9</td>
<td>2.8</td>
</tr>
<tr>
<td>Forms (F)</td>
<td>7.7</td>
<td>1</td>
<td>7.7</td>
<td>-0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>S x R</td>
<td>4734.5</td>
<td>441</td>
<td>10.7</td>
<td>2.5</td>
<td>7.9</td>
</tr>
<tr>
<td>S x F</td>
<td>385.2</td>
<td>49</td>
<td>7.8</td>
<td>0.2</td>
<td>0.7</td>
</tr>
<tr>
<td>R x F</td>
<td>520.2</td>
<td>9</td>
<td>57.8</td>
<td>1.0</td>
<td>3.3</td>
</tr>
<tr>
<td>S x R x F.e</td>
<td>2496.9</td>
<td>441</td>
<td>5.6</td>
<td>5.7</td>
<td>17.7</td>
</tr>
</tbody>
</table>

G coefficient $E^2_F = 0.97$ Phi coefficient $E^2_\phi = 0.96$ 100%
According to Table 2, variance component for students is 21.6 - the biggest value, and it explains 67.6% of the total variance. Because students are regarded as the object of measurement, highness of the variance in the total variance is a desired situation (Shavelson and Webb, 1991). Raters’ variance component is the fourth biggest variance (0.9). It is 2.8% of the total variance. This situation indicates that there are inconsistencies between raters - at least slightly.

The variance for the form was negative (-0.1), and it was estimated that it explained 0.0% of the total variance. Thus, interpretation of this situation may be that changing the form is not influential in results.

Values for interactions of the variables in addition to single effects are calculated in G study. The variance for students’ and raters’ joint effects was the third biggest value (2.5), and it explained approximately 8% of the total variance. While students-forms interaction explained 0.2% of the total variance, raters-forms interaction explained 1.0% of the total variance. The fact that the highest value in dual effects was found in students-raters interaction may be interpreted as that raters behaved differently in assessing the individuals. On the other hand, the fact that the variance percentage of the raters-forms interaction was smaller may be explained as that raters were consistent in themselves.

Surplus component had the second biggest value (17.7%) in the total variance. The high variance value of the surplus component can be regarded as an indicator that there are errors which cannot be measured in the process (Shavelson and Webb, 1991).

The results for the D study - in relation to the second research question - are summarized below. Table 3 shows the G and phi coefficients estimated with changes in the number of raters by keeping students and forms constant.

**Table 3. Results for Changes in the Number of Raters**

<table>
<thead>
<tr>
<th>Raters</th>
<th>Forms</th>
<th>Students</th>
<th>G</th>
<th>Phi</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>50</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>50</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>50</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>50</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>50</td>
<td>0.98</td>
<td>0.97</td>
</tr>
</tbody>
</table>

As is clear from the Table, the G coefficient was estimated to be bigger than the phi coefficient. When the number of forms and students is constant, the number of raters was changed and thus the K study was conducted. Accordingly, when there are 2 raters, the coefficients take on the smallest values whereas in the case of 12 raters the coefficients take on the biggest values. In other words, as the number of raters increases, the values of G and phi coefficients also increase. Although the number of raters raises reliability, high reliability can be estimated with 6 raters by reducing such variables as time, labor and cost; and on increasing the number of raters, there are no considerable changes in the coefficients. The interpretation for this situation can be that 6 raters at the optimum should be taken. Table 4 shows briefly the findings concerning the D study for changes in the number of rubrics by keeping the number of students and raters.
Table 4. Results for Changes in the Number of Rubrics

<table>
<thead>
<tr>
<th>Raters</th>
<th>Forms</th>
<th>Students</th>
<th>G</th>
<th>Phi</th>
<th>Raters</th>
<th>Forms</th>
<th>Students</th>
<th>G</th>
<th>Phi</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>50</td>
<td>0.95</td>
<td>0.94</td>
<td>6</td>
<td>1</td>
<td>50</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>50</td>
<td>0.97</td>
<td>0.96</td>
<td>6</td>
<td>2</td>
<td>50</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>50</td>
<td>0.98</td>
<td>0.97</td>
<td>6</td>
<td>3</td>
<td>50</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>50</td>
<td>0.98</td>
<td>0.97</td>
<td>6</td>
<td>4</td>
<td>50</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>50</td>
<td>0.98</td>
<td>0.98</td>
<td>6</td>
<td>5</td>
<td>50</td>
<td>0.97</td>
<td>0.96</td>
</tr>
</tbody>
</table>

According to Table 4, the G coefficient was estimated as higher than the phi coefficient in any case. The number of forms was changed by keeping the number of raters and students and thus the D study was performed. Accordingly, the coefficients took on the smallest values when there was 1 form whereas they took the biggest values when there were 5 forms. That is to say, when the number of rubrics increased, the G and phi coefficients also rose. It was observed in Table 3 that when the decision study was repeated with 6 raters, the use of 2 rubrics increased compared to the use of 1 tool of measurement and that there were no great changes when there were more than 2 tools of measurement.

3. CONCLUSIONS AND DISCUSSION

It is important that the measurement and evaluation results be reliable in the field of education. Since a great number of variability sources merge into the measurement and evaluation process, it is difficult to attain reliability of performance tasks. Using the G Theory showing the multiple sources of error/variability and their interactions offers advantages in the results of such evaluations (Shavelson and Webb, 1991).

It is expected in the variance analysis results in G studies that individuals—the target of measurement—have the highest percentage in the variance component. This situation shows that individuals differ in the property measured (Guler, 2008). It also indicates that other sources of variability are less influential in measurement. On examining the G study results, it is clear that the biggest source of variability is individuals—which means that results consistent with this generalization have been obtained.

It is clear that 2.8% of the total variance stems from raters. This was a finding indicating that raters did not grade consistently. It is expected that the percentage for this variable be as small as possible. Yet, the complexity of the performance tasks causes the raters to behave in different ways. It was seen that forms did not have any effects on the total variance (0.0%). This might have stemmed from the fact that both tools of measurement had been prepared with the same purpose. The main difference stemmed from the raters themselves, not from the forms they had used. Residual/Random errors had the second biggest share in the variability in the total variance. This situation indicates that there were errors merging into the process.

An examination of the first decision study showed that the highest G and Phi coefficients were obtained with the presence of 10 raters, but that the coefficients with the presence of 6 raters were also high ($E^G_P = 0.96$, $E^P_P = 0.95$). Due to the fact that it was difficult to find well-qualified raters, that working with more raters took too much time and cost too much, and that there might be increase in random errors in the process with the use of more raters; it was concluded that working with 6 raters would be better. Since rubrics were not influential on their own (0.0%) in the total variance in the G study, use of 1 form in the decision study yielded high
coefficients \( r_{EP} = 0.93; r_{E Observer} = 0.91 \), and because preparing more than one rubric was difficult, exhaustive and costly; it may be said that it would be sufficient to use only one rubric.

REFERENCES


versity in Education, 1 (1).


