AN EVALUATION OF THE RUBRIC USED IN DETERMINING STUDENTS' LEVELS OF DISCIPLINED MIND IN TERMS OF GENERALIZABILITY THEORY[*]

Özge CAN ARAN
Dr., Hacettepe University, Curriculum and Instruction, ozgecann@hacettepe.edu.tr

Neşe GÜLER
Assoc. Prof., Sakarya University, Measurement and Evaluation, gnguler@gmail.com

Nuray SENEMOĞLU
Prof., Hacettepe University, Curriculum and Instruction, profdrnuray@gmail.com

ABSTRACT: Nowadays, rapid changes in science and technology increase the demand of qualified individuals who have signs of disciplined mind which is highlighted in Howard Gardner's five minds as one type of mind. So, it is important to measure whether individuals have disciplined mind or not. Based on this idea, it is aimed to evaluate the reliability of rubric used in determining the seventh grade students' levels of disciplined mind in terms of generalizability theory in this research. It is possible in the Generalizability theory to calculate reliability both for the relative decisions (G coefficient) and the absolute decisions (Φ coefficient). The G and the Φ coefficients calculated with six tasks and three raters in this research were found as .86 and .79, respectively. In the generalizability study, attempts are made to predict the situations where the error can be reduced to the minimum for specific purposes through the decision study. According to decision study results, it was found that increasing the number of raters would not bring any benefits to the similar studies to be performed in the future; and it would not be a practical way in cases where it is difficult to find competent raters.

Key words: disciplined mind, reliability, generalizability theory, rubric

ÖĞRENCİLERİN DİSİPLİNLİ ZİHİN ÖZELLİKLERİNİ BELİRLEMEDE KULLANILAN DERECELİ PUANLAMA ANAHTARININ GENELLENEBİLİRLİK KURAMI AÇISINDAN DEĞERLENDİRİLMESİ

ÖZET: Günümüzde, bilim ve teknolojideki hızlı değişim Howard Gardner'ın beş zihin alanından birisi olarak vurguladığı disiplinli zihne sahip nitelikteki insana olan ihtiyacı arttırmaktadır. Bu nedenle bireylerin disiplinli zihne sahip olup olmadığını belirlemek önemlidir. Bu bağlamda, bu araştırmada yedinci sınıf öğrencilerinin disiplinli zihin özelliklerini belirlemek için kullanılan dereceli puanlama anahtarı güvenirliğinin Genellenebilirlik Kuramı ile kestirilmesi amaçlanmıştır. Genellenebilirlik kuramı ile göreli kararlar (G katsayısı) ve mutlak kararlar (Φ katsayısı) için güvenirlik katsayısının hesaplanması mümkün olup; bu çalışmada elde edilen puanların G ve Φ katsayısı, altı görev ve üç puanlayıcı için sırasıyla .86 ve .79 olarak bulunmuştur. Genellenebilirlik kuramında ayrıca, karar (K) çalışmaları yoluyla değişkenlik kaynaklarının sayısının değişmesi durumunda güvenirliğin tahmin edilmesi de mümkün olmaktadır. Bu araştırmadaki karar çalışması sonuçları, gelecekte yapılacak benzer nitelikteki çalışmalarda, puanlayıcı sayısını arttırmanın puanların güvenirliğini arttırmaya katkı sağlamayacağını göstermektedir. Nitelikli puanlayıcı bulmanın kolay olmadığı durumlarda, puanlayıcı sayısını arttırmanın uygulanma açısından da pratik olmayacağı düşünüldüğünde; daha fazla puanlayıcı ile benzer çalışmalar yapılmasının önerilemeyeceği sonucuna varılmıştır.

Key words: disiplinli zihin, güvenirlik, genellenebilirlik kuramı, dereceli puanlama anahtarı

1. Literature review

Rapid advances have been occurring today in science and technology. Improvements such as laser surgery, gene therapy, and fiber internet are the indicators of the rapid advances in science and technology. It is believed that science and technology course is important in raising people who can keep up with such developments in science and technology, generate and transfer knowledge into technology, and thus can bring the country's level of development to higher points; because it is believed that the development level of a country having individuals successful in the field of science studies will rise and reach the achievement level of first world countries easily. In raising individuals to contribute to their country's development, raising individuals with disciplined mind, which is highlighted in Howard Gardner's (2006) five minds as one type of mind.

1.1. Disciplined mind

The disciplined mind comprises employing the ways of thinking associated with major scholarly disciplines (history, math, science, etc.) and major professions (law, medicine, management, etc.); capable of applying oneself diligently, improving steadily, and continuing beyond formal education. It is thought to be important since individuals with a disciplined mind make efforts to learn in depth the discipline on which they are working, enjoy learning new things on the discipline, and can look at the discipline from a number of perspectives (Gardner, 2006). On analyzing the properties of disciplined mind, it may be said that they are the properties that successful scientists possess. Therefore, developing those properties is believed to be important in raising scientists who will make original discoveries and will thus contribute to their country. In order to develop those properties, firstly, it is useful to see at what level the students display those properties in science lesson. Gardner (2006) points out that those properties develop in adolescence. Hence, a measurement tool was developed in this research so as to exhibit the

[*] This manuscript is based on the data used in the pilot study of Özge CAN ARAN's PhD Thesis. This thesis was supported by grants from Hacettepe University Scientific Research Projects Coordination Unit.

level at which those properties develop in 7th graders. In developing this tool, Chun's (2010) "Collegiate Learning Assessment in the Classroom" (CLA), a measurement tool for "the assessment of what is learnt at University", was taken as a model.

1.2. Collegiate learning assessment in the classroom (CLA)

Students are presented scenarios from real life in the CLA. At the same time, students can also be given roles so that they can feel that they are within the scenarios (for instance, a student of journalism is asked to write an article for a magazine, or a biology student is asked to give advice to a relative suffering from cancer). The scenarios given to the students involve a challenging task from real life which does not have correct or incorrect answers, which is complex, for which the solution is not clear, and for which the knowledge is incomplete. Also in CLA, it is expected that students use critical thinking, analytic reasoning and problem-solving skills in facing the situations presented. In addition to this, it is also examined whether the students focus on analyzing, synthesizing and using the evidence for decision-making and for making judgments. Briefly, CLA assesses this set of skills through the use of performance task. Similarly, in this research, students were asked to write an article to demonstrate the positive and negative aspects of building nuclear power plants by using 4 different reports. The first of the 4 reports emphasized the positive sides of nuclear power plants, while the second stressed the negative sides, the third included a table showing that those power plants generated more energy than other types of power plants, and the fourth presented a map showing the regional distribution of nuclear power plants in the world. Chun (2010) referred to those scenarios as tasks. Since the Ministry of Education preferred the term performance task for such activities, the term performance task was used in this research for the article that the students were supposed to write. A rubric was designed so as to evaluate the article students were expected to write as the performance task by using disciplined mind. So the scores of students can be calculated with the help of this rubric.

1.3. Reliability of scores

The reliability of the scores obtained from all measurement tools, and hence the reliability of the scores obtained from observations in this research is one of the most important issues, which needs attention (Goodwin & Goodwin, 1991). Besides, since scoring by observation is a subjective process, the reliability of the scores is additionally important. For this reason, it is assured that mostly more than one rater give scores at the same time, as in this research, to make sure that the scores assigned to each student by observing their behaviors are more objective and more reliable.

On examining the research studies making evaluations based on measuring the performance, it was found that methods such as concordance indices for classical test theory (CTT), Pearson's correlations coefficient, t tests, or variance analyses were employed in calculating the inter-raters reliability (Goodwin, 2001; Güler & Gelbal, 2010; Yelboğa &Tavşancıl, 2010).

In statistical methods used in determining the reliability based on CTT, each source of error such as the measurement tool in the measurement process, raters, time in repetitive measurements, etc. can be considered separately; and reliability value can be calculated for each (Güler, Uyanık and Teker, 2012; Lei, Smith & Suen, 2007; Shavelson and Web, 2005). In brief, it is impossible in CTT to consider all the sources of error together and to derive one single reliability coefficient simultaneously (Baykul, 2000; Güler, 2011). Generalizability (G) theory, on the other hand, enables one to calculate reliability by considering all sources of error together and simultaneously (Cronbach, Gleser, Nanda & Rajaratman, 1972; Lucas, Brian, Arnetz & Arnetz, 2010). In addition to that G Theory, unlike CTT, makes it possible to predict not only the effect of one source of error, but also the effects of the interaction of those sources of error, to calculate the reliability both for relative and absolute decisions, and to form scenarios in which the desired reliability may be attained through different decision (D) studies (Güler, 2011; Shavelson & Webb, 1991; Yin & Shavelson, 2008).

1.4. Generalizability theory

The Generalizability (G) Theory has been developing since 1963 and has been implemented in several fields, especially in measuring the performance tasks. One of the most important reasons for the G theory to be more preferable than CTT in performance measurements is that it is assumed that the measurement tools in the G theory can be "randomly parallel" whereas in the CTT it is assumed that the measurement tools need to be "strictly" parallel (Hsu, 2012). In addition to that, the G theory removes the difference between reliability and validity which is available in the CTT in performance measurements; and thus makes it possible to determine how the sample in the study representing the population can be generalized in a valid way into all the probable measurement situations (Allal & Cardinet, 1997; Güler, 2011; Volpe, McConaughy & Hintze, 2009). G theory enables one to generalize the measurement results of a group of individuals –and even of only one individual- into much beyond the number of items, raters or situations through which the results are obtained (Brennan, 1992; Shavelson & Webb, 1991).

In the G theory, the whole of acceptable measurements where the entire probable circumstances of observation and the sources of variation beyond the measurement of a performance, task, etc. available in the measurement process is called the *universe*. Each source of variation such as the items (tasks), measurement tools, raters, or different measurement situations available in the measurement process in the G theory is called a *facet*. Facet can be interpreted as "the measurement situations having similarities". Each level on the facets is referred to as a *condition*. For instance, in the process of a 20-item multiple test given to students, items constitute a facet, and each item is one condition of the facet (Güler, Uyanık & Teker, 2012). Each facet can have infinite size. The source revealing the variability of concern (individuals, students, etc) is called the *object of measurement* constituting the real, systematic variability, rather than being called the source of variation (Kieffer, 1998; Musquash & O'Connor, 2006). Yet, the object of measurement does not always have to be composed of individuals; and items and situations can also be the object of measurement in conformity with the nature of the study (Brennan, 1992; Lei et al., 2007). Whereas the value of variance for the object of measurement is desired to be high, the value of variance for each source of variation is desired to be as low as possible (Alharby, 2006). The average which can be obtained from all the probable measurement situations of the object of measurement is called the universe score. The universe score reflects the real change that the researcher is interested in, and it is interpreted in a similar way to the real score variance in the CTT (Güler, Uyanık & Teker, 2012; Kieffer, 1998).

As different from CTT, two separate variances of error are available in the G theory. Thus, it is possible to calculate the reliability coefficient for the absolute decisions which are not available in CTT in addition to the generalizability coefficient which is found for relative decisions. The *generalizability* (*G) coefficient* predicted for the relative decisions is derived by considering the place in the ordering of other students' scores rather than considering how high the raw score of each students' object of measurement is (as can be remembered from what is stated above, the object of measurement does not need to be students or individuals). This coefficient is similar to the reliability coefficient in CTT. The dependability-phi (ϕ) coefficient predicted for absolute decisions is, however, a more strict value and exhibits both the degree of consistency of ordering of scores for students and the degree of consistency of the raw scores. The G coefficient may be preferred in performance measurements where a score above a certain cutoff score is important (for instance in driving tests, expertise examinations, etc). In cases where the place of score obtained in the ordering of scores is important, using the ϕ coefficient would be appropriate (Lee & Frisbie, 1999; Brennan, 1992). Both the generalizability (G) coefficient and the phi (Φ) coefficient receive values between 0 and 1. The Φ coefficient has a more strict value than the G coefficient. The G coefficient obtained in one-facet fully crossed designs is interpreted in a similar way to the Cronbach α coefficient in CTT (Musquash & O'Connor, 2006; Sudweeks, Reeve & Bradshaw, 2005).

The facets can be handled in a random or fix way. Whether a facet is to be handled random or in a fix way completely depends on the researcher's decision. If the conditions in the study are only a small sample of a much larger universe, or if it is possible to replace the conditions with other probable conditions on that facet, that facet is described as *random* (Güler, uyanık & Teker, 2012). For example, if the tasks available in the measurement of a performance are replaceable with other probable tasks to be available in a measurement to be made in the same field, the tasks in the study are considered as random. Studies depending on facets with random situations enable the researcher to make generalizations into the universe where all the conditions for that facet are available. On the other hand, if the researcher is concerned only with certain conditions depending on the facet included in the study, and if he or she does not aim to generalize into other conditions, the facet considered in this way is described as "*fixed*" (Crocker & Algina, 1986). A fixed facet can appear in two ways: 1. the number of conditions in a study is composed of the conditions chosen by the researcher in line with his/her purpose, and the researcher does not wish to make generalizations above those conditions. 2. The number of conditions in the study is very small and all of them are available in the process of measurement. If there is one or more fixed facets in a study, the variance of error will decrease; which will lead to an increase in dependability. However, this case will restrict the generalizability of the results (Güler, Uyanık & Teker, 2012; Brennan, 2011; Kieffer, 1998). The models of measurement having at least one fixed facet are called mixed models. Another point of importance here is that the G theory is a theory of measurement which is based on random facets. Hence, this case requires that at least one facet should be random; and it is not possible for all facets to be fixed (Güler, Uyanık & Teker, 2012).

The studies in the G theory can be described as crossed or nested designs. If all the conditions of a facet in a study are available in all conditions of the other source of variation, such a research design is called as a *crossed design*. If, for instance, all the students in a classroom (s) answer all the items (i) in a test and if the items are scored by the same raters (r), this design is referred to as a *crossed design* (and sometimes as a fully crossed design). Crossed designs are represented by the symbol "x". The representation of the crossed design in the example is: "*s x i x r*". If, however, one condition of a facet is available in only one condition of the other and if it is not available in the others, this is said to be a *nested design*. For example, if each student is asked different questions (q) in an interview and if each answer of each student (s) is scored by different raters (r), it means that *fully nested design* is used in this study. The representation of the fully nested design in the example is: "*s: q: r*". In some studies, both the crossed and the nested designed are used together and such designs are called *nested designs* (Brennan, 1992; Shavelson & Web, 1991). Although the G theory can be used in all of the designs mentioned here, the use of crossed (fully crossed) designs when possible in order to be able to make predictions for all facets provides an advantage in the G theory studies (Kieffer, 1998).

There are two kinds of studies of researching the dependability in the G theory: 1. Generalizability (G) study, and 2. Decision (D) study. The *G study* makes it possible to make predictions for all the sources of variation simultaneously and through the ANOVA method (Güler, 2009). By using the results obtained from the G study, attempts are made to predict the situations where the error can be reduced to the minimum for specific purposes in the D study. The results obtained from the *D study* help the researcher to make predictions about what conclusions he/she can reach on changing the number of items, raters and observations (Volpe et al., 2009). In a sense, the D study can be interpreted similarly to the purpose of using the Spearman Brown formula in CTT (Musquash & O'Connor, 2006). With Spearman Brown formula, it is possible to predict the reliability according to the change in the number of items in the tool of measurement. In the D study, however, the prediction is not restricted only to the number of items, and it enables one to predict the values receivable by reliability; generalizability and the phi coefficient simultaneously in one single study in case of differentiation of the conditions of all facets. Thus, the *D studies* help to predict the most effective and economical measurement situations for reliability (Lee & Fitzpatrick, 2003).

Research studies measuring students' skills of solving mathematical problems, scoring their reading, writing and musical skills, and researching the dependability of measurement results obtained from those performance tasks through the G theory are available in literature of education (Baker, Abedi, Linn & Niemi, 1995; Güler & Gelbal, 2010; Lane, Liu, Ankenmann & Stone, 1996; Mercer, Dufrene, Martell, Harpole, Mitchell & Blaze, 2012). This study also analyses the reliability of a measurement tool developed for the measurement of students' meta-cognitive behaviors in Science and Technology course through the G theory.

2. Materials and methods

2.1. Participants and application

The study group of the research was determined through typical situation sampling, one of the purposive sampling methods. The averages for high school placement test scores of the year 2009 for the schools in the central districts of Ankara, which were obtained from the Ministry of Education, formed the basis of typical sampling. The schools were ordered according to the score averages and the groups of 27% at the top and at the bottom were identified accordingly. Thus, a school which was in the middle according to the achievement scores was determined. Hence the study group was composed of 30 students of the seventh grade coming from schools of medium level achievement in terms of high school placement test achievement.

2.2. Measurement tool

The rubric prepared to evaluate the students' article was designed in six categories and according to four levels of achievement as 0, 1, 2, and 3 aiming to assess such criterion as using reports given to the students effectively (1 item), understanding in depth the reports (1 item), evaluating the reports objectively (1 item), establishing the cause-effect relationships correctly while constructing the article (1 item), making meaningful connections between the sentences or the paragraphs (1item) and including differing explanations in the article (1item). Two raters scored the articles based on this rubric.

2.3. Analysis of the data

The analyses of the scores according to the G theory was performed through the SPSS program developed by Musquash & O'Connor (2006) for the G theory, whereas the Cronbach alpha coefficients- which are the reliability for each rater- were obtained through the SPSS 16 package program according to the classical test theory.

3. Results and discussion

Firstly, the levels of fulfilling disciplined mind of 30 students (s) as the object of measurement were scored according to six tasks (t) by three raters (r). The three raters scored all the tasks performed by all the students, and the crossed design (*s x t x r*) was employed in the research. Thus, there are two facets in the research: tasks and raters. The results for the variance components obtained through generalizability analysis according to this design are shown in Table 1 below.

Table 1. Analysis of variance results and variance component estimates for students, tasks, raters and interactions

Source of variance	SS	df	MS	Variance Component Estimates	Percentage of Total Variance Estimates
s	210.128	29	7.246	.348	.349
t	101.039	5	20.208	.210	.211
r	1.544	2	.772	.000	.000
st	83.683	145	.577	.115	.115
sr	37.011	58	.638	.068	.068
tr	9.233	10	.923	.023	.023
str,e	67.544	290	.233	.233	.234
					100%

In Table 1, both key elements of ANOVA table and the variance component estimates are observed. Because, G theory focuses on the size of the variance component estimates, and not the statistical significance of the facets or their interactions, Table 1 does not include the significant test results (Goodwin & Goodwin, 1991). Also, there are the percentages of the each variance component to the total variance in the last column of the table. The first three estimates in that column are for the main effects of students, tasks and raters. While students (object of measurement) account for the largest percentage of the variance (34.9%), the main effect of the task accounts for 21.1% of the total variance and the main effect of the rater does not account for any variance (0.00%). These results exhibit the desired ideal case in measurement. It is desired that the variance stemming from the object of measurement is high and the values for the other sources of variance are low as possible (Brennan, 1992; Shavelson & Web, 1991). This case shows that the variability in measurement results is not dependent on the raters or the tasks. There was perfect consistency between the raters in this measurement process. On the other hand, it can be seen that two way interactions of student-by-task and task-by-rater account for 11.5% and 6.8% of the total variance, respectively. As it is seen clearly, the value of 11.5% demonstrates that each task's level of difficulty is the level of differentiation from student to student. This is inevitable in cases where the probability of differences that can stem from students' earlier experiences and attitudes is high. The fact that the 6.8% of the total variance stems from the student-rater interaction demonstrates that the raters' scoring did not differ much from student to student. As another interaction, task-by-rater yielded second smallest variance component estimates. This also indicates that the raters' scoring did not almost change from task to task. At last, the three way-interaction, students-by-tasks-by-raters, is also named as "residual" or "error" in the ANOVA model used here. If the measurement results are reliable in a research, this value of residual is desired to be as small as possible. According to Table 1, the three-way interaction accounted for 23.4% of the total variance. According to the G theory, this value of variance is desired to be as small as possible. This value signals that the change in scores might have emerged due to different sources of variation which were not available in the study.

Consequently, as is evident from Table 1, the researcher can see how much of the total variance is the result of the interaction of which source or sources- which is an advantage of the G theory (Güler, 2009).

The G coefficient, which is interpreted in a similar way to the reliability coefficient in classical test theory, is calculated in the G theory. As is explained in the introduction, it is possible in the G theory to calculate reliability both for the relative decisions (G coefficient) and the absolute decisions (phi coefficient). The G and the phi coefficients calculated with six tasks and three raters in this research were found as .86 and .79, respectively. Apart from that, the detailed calculation of the G and the phi coefficients using the values in Table 1 is shown in Table 2.

Table 2. Calculation of G coefficient

I. G coefficient for 6 tasks and 3 raters (n_t:6, n_r:3)

$$E\hat{\rho}^2 = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \frac{1}{n_t}\hat{\sigma}_{st}^2 + \frac{1}{n_r}\hat{\sigma}_{sr}^2 + \frac{1}{n_t n_r}\hat{\sigma}_{str}^2}$$

$$= \frac{.348}{.348 + \frac{1}{6}.115 + \frac{1}{3}.068 + \frac{1}{18}.233}$$

$$= \frac{.348}{.403}$$

=.86

II. Φ coefficient for 6 tasks and 3 raters (n_t:6, n_r:3)

$$\Phi = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \frac{1}{n_t}\hat{\sigma}_t^2 + \frac{1}{n_r}\hat{\sigma}_r^2 + \frac{1}{n_t}\hat{\sigma}_{st}^2 + \frac{1}{n_r}\hat{\sigma}_{sr}^2 + \frac{1}{n_t n_r}\hat{\sigma}_{tr}^2 + \frac{1}{n_t n_r}\hat{\sigma}_{str}^2}$$

$$= \frac{.348}{.348 + \frac{1}{6}.21 + \frac{1}{3}.0 + \frac{1}{6}.115 + \frac{1}{3}.068 + \frac{1}{18}.023 + \frac{1}{18}.233}$$

$$= \frac{.348}{.439}$$

=.79

As is pointed out in the introduction part of the research, by using the results obtained in the G study, attempts are made to predict the situations where the error can be reduced to the minimum for specific purposes through the D study in the G theory. The values receivable by dependability- that is to say, by the G and the phi coefficients- are predicted on decreasing or increasing the number of tasks and/or raters in the D study. The number of tasks included in the tool of measurement used in this study is clear. Yet, the extent to which dependability will change on raising or reducing the number of raters was researched in this study. The results of the D study are shown in Table 3.

Table 3. G and phi coefficients of D studies (n_t:6)

	1 rater	2 rater	**3 rater***	4 rater	5 rater
G coefficient	.74	.83	**.86**	.88	.89
Φ coefficient	.68	.76	**.79**	.81	.82

(*Number of raters in this study.)

As is apparent from Table 3, raising the number of raters above two does not increase dependability value significantly. Therefore, increasing the number of raters will not bring any benefits to the similar studies to be performed in the future; and it would not be a practical way in cases where it is difficult to find competent raters.

The Cronbach α values calculated according to CTT for the scores given by each rater included in the research and the G coefficient values calculated according to fully crossed design with one facet (*s x t*) in the same order according to the G theory were found as .79, .89, and .91. Through the D study, the fact that the G coefficient obtained with one rater was found as .74 can be interpreted the lower bound of these values.

4. Conclusion

The present study aims to investigate the reliability of a measurement tool developed for the measurement of students' meta-cognitive behaviors in Science and Technology course through the G theory. As is clear from the results of this research, in cases of measurement where several sources of variation such as tasks and raters are available, the G theory provides detailed information through one analysis. In this case, as detailed information provided by means of G theory is examined, it can be easily seen that the variance stemming from the object of measurement is high and the values for the other sources of variance are low as possible. This means that the majority of variability in measurement results is not dependent on the raters or the tasks. It can be also figured out that raters' scoring did not differ much from both student to student and the task to task. So it can be implied that there was perfect consistency between the raters in this measurement process.

Also, the value of three way-interaction, students-tasks-raters, is another noteworthy information ensured by G theory. Because it signals how much of the total variance by researcher is the result of the interaction of which source or sources. It also explains that the change in scores might have emerged due to different sources of variation which were not available in the study. In this research, it is considered that this value is in acceptable level. So it is possible to say that the change in the scores caused by different sources of variation which were not available in the study.

Moreover, G theory provides information source about G and the phi coefficients in this research. These coefficients are calculated by taking account all the sources of error together and simultaneously in order to predict not only the effect of one source of error, but also the effects of the interaction of those sources of error. In this research, the G and Φ coefficients confirms that the rubric used in determining the seventh grade students' levels of disciplined mind is a reliable measurement tool.

Taking all results obtained into consideration, G theory would serve as an appropriate method of determining reliability. Therefore it can be used in cases where individuals' behaviors are observed and evaluated as education and psychology especially, it is common to include more than one rater in order for the observation results to be objective. This study also is wished to help expand the application of G theory in this kind of evaluation research.

REFERENCES

Alharby, E. R. (2006). *A comparison between two scoring methods, holistic vs. analytic using two measurement models, the generalizability theory and the many facet rasch measurement within the context of performance assessment* (Unpublished Doctoral Dissertation). Pennsylvania State University.

Allal, L., & Cardinet, J. (1997). Generalizability theory. In J.P. Keeres (Ed.), *Educational research, methodology, and measurement: An international handbook* (2nd, pp. 737- 741). Cambridge, United Kindom: Cambridge University.

Baker, E. L., Abedi, J., Linn, R. L., & Niemi, D. (1995). Dimensionality and generalizability of domain-independent performance assessments. *Journal of educational research, 89*, 197–205.

Baykul, Y. (2000). *Measurement in education and psychology: Classical test theory and its application.* Ankara: ÖSYM, Turkey.

Brennan, R. L. (1992). *Elements of generalizability theory*. New York: Springer-Verlog.

Brennan, R. L. (2001). *Generalizability theory*. Iowa: ACT Publications.

Chun, M. (2010). *Taking teaching to (performance) task: Linking pedagogical and assessment practice*. Retrieved from http://www.learningace.com/doc/2844170/7d60cfc1e1ecc6db4a2df9e4e4038863/chun _change_takingteachingtotask.

Crocker, L., Algina, J. (1986). *Introduction to classical and modern test theory*. USA: Harcourt Brace Javanovich College Publishers.

Cronbach, J. L., Gleser, G. C., Nanda & Rajaratman, N. (1972*). The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley and Sons.

Gardner, H. (2006). *Five minds for the future*. USA: Harvard Business School Press.

Goodwin, L. D.&Goodwin, W. L. (1991). Research notes: Using generalizability theory in early childhood special education. *Journal of Early Intervention*, 15 (2), 193-204.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Psychical Education and Exercises Science*, 5(1), 13-14.

Güler, N. (2009). Generalizability theory and comparison of the results of g and d studies computed by SPSS and GENOVA packet programs. *Education and Science*, 34, 154, 93-103.

Güler, N. (2011). The comparison of the reliability according to generalizability theory and classical test theory on random data. *Education and Science*, 36, 162, 225-234.

Güler, N., Uyanık, G. K., & Teker, G. T. (2012). *Generalizability theory.* Ankara, Turkey: Pegem Akademi Publishing.

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

Hsu, L. (2012). Applications of generalizability theory to estimate the reliability o EFL learners' performance-based assessment: a preliminary study. *Educational Research*, 3(2), 145-154.

Kieffer, K. M. (1998). *Why generalizability theory is essential and classical test theory is often inadequate.* Paper presented at the annual meeting of the Southwestern Psychological Association. New Orleans, LA. USA.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, 33(1), 71-92.

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.

Lee, G.& Fitzpatrick, A. R. (2003). The effects of a student sampling plan on estimates of the errors for stuents passing rates. *Journal of Educational Measurement*, 40(1), 17-28.

Lei, P., Smith, M., & Suen, H. K. (2007). The use of generalizability theory to estimate data reliability in single subject observational research. *Psychology in Schools*, 44(5), 433-439.

Lucas, T., Brian, L., Arnetz, J. & Arnetz, B. (2010). Do ratings of african-american cultural competency reflect characteristics of providers or perceivers? Initial demonstration of a generalizability theory approach. *Psychology, Health & Medicine*, 15(4), 445-453.

Mercer, S. H., Dufrene, B. A., Martell, K. Z., Harpole, L. L., Mitchell, R. R. & Blaze, J. J. (2012). Generalizability theory analysis of CBM maze reliability in third- through fifth- grade students. *Assessments for Effective Intervention*, 20(10), 1-8.

Mushquash, C. & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analysis. *Behavior Research Methods*, 38 (3), 542-547.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory*: *A Primer.* USA: Sage Publications.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many facet measurement in an analysis of college sophomore writing. *Assessing Writing*, 9, 239-261.

Volpe, R. J., McConaughy, S. H., & Hintze, J. M. (2009). Generalizability of classroom behavior problem and on-task scores from the direct observation form. *School Psychology Review*, 38,3.

Yelboğa, A. & Tavşancıl, E. (2010). The examination of reliability according to classical test and generalizability on a job performance scale. *Educational Sciences: Theory & Practices*, 10(3), 1847-1854.

Yin, Y. & Shavelson, R. J. (2008). Application of generalizability theory to concept map assessment research. *Applied Measurement in Education*, 21, 273-291.

This page intentionally left blank.