Güler, N., Eroğlu, Y. ve Akbaba, S. (2014). Reliability of criterion-dependent measurement tools according to Generalizability Theory: Application in the case of eating skills. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 14*(2), 217-232.

Geliş Tarihi: 11/03/2014 Kabul Tarihi: 08/11/2014

RELIABILITY OF CRITERION-DEPENDENT MEASUREMENT TOOLS ACCORDING TO GENERALIZABILITY THEORY: APPLICATION IN THE CASE OF EATING SKILLS

Neşe GÜLER[*]
Yüksel EROĞLU[**]
Sırrı AKBABA[***]

ABSTRACT

Applied behavioral analysis is one of the most frequently utilized to help students with mental disabilities develop skills of living independently. Applied behavior analysis is based on direct measurement with criterion-dependent measurement tools. As applications of them have played significant role to evaluate students with mental disabilities; the reliability of these measures has become increasingly important issue. For this reason, in this study, generalizability theory was used to estimate the reliability of them and the role of different sources of error in the variability of measurements in the case of measuring eating by the spoon skills. The results indicated measurement results varied depending on occasions, occasion by task and task by rater effects whereas effects of tasks and raters were negligible.

Keywords: Reliability; Generalizability Theory; Criterion-Dependent Measurement Tools

GENELLENEBİLİRLİK KURAMINA GÖRE ÖLÇÜT BAĞIMLI ÖLÇME ARAÇLARINDA GÜVENİRLİK: YEMEK YEME BECERİLERİ ÖRNEĞİNDE BİR UYGULAMA

ÖZ

Zihinsel engelli öğrencilere bağımsız yaşama becerileri kazandırmanın en etkili yollarından birisi uygulamalı davranış analizidir. Uygulamalı davranış analizinin temelini ölçüt bağımlı ölçme araçlarıyla yapılan doğrudan ölçümler oluşturmaktadır. Ölçüt bağımlı ölçme araçlarının uygulanması zihinsel engelli öğrencilerin değerlendirilmesinde önemli rol oynadığı için bu ölçme araçlarının güvenirlikleri gittikçe önemli bir konu olmaktadır. Bu nedenle bu araştırmada bu ölçme araçlarının güvenirliği ve çeşitli hata kaynaklarının ölçüm sonuçlarının değişkenliğinde oynadığı rol, kaşıkla yemek yeme becerilerinin ölçülmesi örneğinde genellenebilirlik kuramı aracılığıyla kestirilmeye çalışılmıştır. Araştırma sonuçları, birey, birey ve görev ortak etkisi ve görev ve puanlayıcı ortak etkisinin önemli bir değişkenlik kaynağı olduğunu buna karşılık görev ve puanlayıcı ana etkisinin önemsiz olduğunu göstermiştir.

Anahtar Sözcükler: Güvenirlik, Genellenebilirlik Kuramı, Ölçüt Bağımlı Ölçme Araçları

[*] Assoc.Prof.Dr., Sakarya University, Faculty of Education, , e-mail: gnguler@gmail.com

[**] Research Assistant, Uludag University, Faculty of Education, e-mail: yeroglu45@gmail.com

[***] Prof. Dr., Uludag University, Faculty of Education, e-mail: sakbaba@uludag.edu.tr

1. INTRODUCTION

One of the methods most influential in assuring students with mental disabilities to acquire the skills for independent living is the applied behaviour analysis. The analysis requires that the pre-behavioural and post-behavioural stimuli should be arranged systematically so as to achieve the desired change of behaviours in individuals with disabilities (Heward, 1996). On the basis of applied behaviour analysis lays the direct and continual measurement of behaviour. In direct and continual measurements, the learners' performance is continuously observed and evaluated in environments where the behaviour takes place (Özyürek, 1996), and the criterion-dependent tools of measurement are employed in those measurements (Varol, 1996).

The criterion-dependent measurement tools are composed of announcements, criteria and questions. The announcements section of the tool contains the stages of the skill under analysis as well as the criteria established. If a student's performance level is to be determined through one single application of the criterion-dependent measurement tool, it is 100% adopted as a criterion. The questions/tasks section of the tool, on the other hand, is arranged according to the method to be used in determining the level of performance. The criterion-dependent measurement tool is prepared by adding the main instructions and the independent column for skills to this section to determine the level of performance through the method of single opportunity; and by adding the columns of main instruction, independent and verbal clues, modelling and physical help through single opportunity method to determine the performance level through the method of multiple opportunities (Varol, 2004).

Single opportunity method is the direct observation and recording of what students can do by giving them only the main instructions (the instructions which enable individuals to actualize a skill when given to individuals having that skill). The aim in using this method is to determine how much of the skill a student can utilise independently. Multiple opportunity method, however, is offering the students new opportunities after giving them the main instructions so that they can fulfil the stages that they have difficulty in performing. The aim in using this method is to see what clues the students have employed in actualising each stage of the skill and whether or not they have performed the skills independently (Varol, 2004).

Criterion-dependent measurement tools are employed to determine students' starting level in a concept/skill in the teaching process, to record the progress that they make during teaching, and to determine the levels of achieving the teaching objectives at the end of teaching (Gürsel, 1993). In performance-based evaluation, the measurement of the stages of the skill desired to be performed via standard questions/tasks helps to exhibit the difficulties that students experience in performing the skill (Tindal, Yovanoff, & Geller, 2010). Teachers as well as psychological counselling and guidance experts evaluate students' behaviours generally through direct observations. The reliability of observation-based scoring is one of the most important issues in such research, as in all tools of measurement used in education (Goodwin & Goodwin, 1991). On the other hand, the greatest disadvantage of scoring based on observation is its subjectivity. Therefore, mostly the average of one rater's scorings at different times or of scores given by more than one rater at a time is used in order to attain higher objectivity and reliability in instances of observation-based scoring. The evaluation of one single student's behaviour in educational or clinical settings is also done similarly. As is specified by Educational

and Psychological Test Standards (American Association of Educational Research, American Association of Psychologists, 1999), "whatever the nature of measured behavior is, it is necessary to determine the reliability and validity of measurement results used in making certain educational decisions" (as cited in Lei, Smith, & Suen, 2007).

In domestic as well as foreign studies regarding special education, it is observed that only raters are taken into consideration as the source of variation in performance-based evaluation and that the methods related to classical test theory (fit indices, Pearson's correlation coefficients, t tests or variance analyses, intragroup variation coefficients) are employed in calculating the interraters reliability (Akköse, 2008; Akmanoğlu & Batu, 2004; Özkan & Gürsel, 2006; Parrott, Schuster, Collins, & Gassaway, 2000; Topsakal & Düzkantar, 2010).

The most serious restriction of the statistical methods used in determining reliability based on classical test theory is that they focus only on interrater inconsistencies as a source of error. Reckase (1995) stated that there might be many possible errors in observations made in natural environments. Possible sources of errors which might be encountered in the evaluation of behaviors are reported to be evaluators, items constituting a measurement tool, time, method, place and dimension (Hintze & Matthews, 2004; Lei et al., 2007; Volpe, McConaughy, & Hintze, 2009; Web & Shavelson, 2005). The concept of generalizability (Cronbach, Gleser, Nanda, & Ratjaratman, 1972) enables the prediction of an error of measurement over different sources of variability against the limitation of explaining the source of error in measurement with a single source of variability (for example, based on only source of rater variability). In this way, observed scores of individuals under measurement (measurement objects) can predict universe scores (real scores) as correctly as possible (Atılgan & Tezbaşaran, 2005; Güler & Gelbal, 2010).

The theory of generalizability (G) is a statistical method which enables us to determine the reliability of measurement results, and design, research into and conceptualize reliable observations, (Brennan, 2001; Cronbach et al., 1972). G theory aims to generalize measurement results obtained from a group of individuals - even from only a single individual (Lei et al., 2007) – obtained measurement results, certain number of items through which these results are obtained, much beyond raters or situations (Brennan, 1992; Shavelson & Webb, 1991). According to Shavelson and Webb (1991), G theory is a more broadened version of the classical test theory from four different perspectives: 1) Generalizability theory addresses multiple variance sources in a single analysis. 2) Enables the determination of the size of each source of variance. 3) Enables the calculation of two different reliability coefficients regarding both relative decisions based on individuals' performance levels and absolute decisions about individuals' performance levels. 4) A suitable theory which enables the arrangement of measurements, where error of measurement can be minimized, depending on a certain aim (D-studies). In brief, G theory is a suitable theory to predict the reliability of results obtained through measuring performance where different sources of error are likely.

In a specific situation, beyond measuring a performance, task, etc. by observing, all likely observation conditions and variability sources including acceptable whole of observations are called "universe" in G theory. Thus, G theory removes the traditional difference between reliability and validity by stating that reliable results can be reached when making precise predictions about universe (Güler, 2009). In G theory, items (tasks),

measurement tools, raters or different measurement times included in measurement process each is called source of variability (facet). And levels of variability sources are expressed as "condition" of variability sources. The condition of each variability source might have an infinite size. In measurement, if it is individuals, students, etc. that reveals variability, main attention is paid to condition called as "object of measurement", not as a source of variability, which constitutes real systematic variability (Kieffer, 1998; Musquash & O'Connor, 2006). However, it is not obligatory that measurement object should be composed of individuals all the time; sources of variability such as item, condition, etc. might be objects of measurement in accordance with the nature of a study as well (Brennan, 1992; Lei et al., 2007). While the variance related to object of measurement is required to be big, the variance value related to each source of variability is required to be as small as possible (Alharby, 2006). The mean of values which can be obtained from all possible measurement conditions of measurement object is called "universe score". Universe score reflects real change which a researcher is essentially interested in and interpreted in a way similar to real score variance in CTT (classical test theory) (Kieffer, 1998).

Sources of variability included in G theory can be taken as fixed or random. If conditions included in a source of variability have a characteristic of being able to be replaced by other possible conditions which are likely to be included in that source of variability, this source of variability is defined as random (Kieffer, 1998). For example, if tasks taking place in the measurement of a kind of performance have a characteristic of being able to be replaced again by other possible tasks which can take place in a measurement to be made in the same field, in this case, tasks within the scope of a study are taken "*randomly*". Studies made depending on sources of variability, where random conditions come into question, enables a researcher to be able to make a generalization for that source of variability to the universe including all conditions. However, if a researcher is interested only in certain conditions included in a study which he or she makes depending on source of variability and does not have an aim such as making a generalization to other conditions, in this case the source of variability under discussion is defined as *"fixed"* (Crocker & Algina, 1986). In studies including fixed sources of variability, it will not be appropriate for a researcher to make a generalization (Kieffer, 1998).

In generalizability theory, as different from classical test theory, there are two separate variances of error. In this way, as in the correlation coefficient obtained in CTT, besides the generalizability coefficient obtained for relative decisions, the calculation of reliability coefficient for absolute decisions not taken into consideration in CTT becomes possible as well. G coefficient calculated for relative decision is calculated not through the height of a raw score obtained by each student (object of measurement; not necessarily be a student or an individual all the time) from a source of variability but depending on its place in the ranking of other students' scores. This coefficient reliability is similar to the one in classical theory. However, the G coefficient calculated for absolute decision is a more strict value and puts forward both the degree of consistency of scores obtained by students in ranking and that of the consistency of raw scores. In performance measurements, where a point above a certain cutting point is important (for example, in qualifying examinations, specialty examinations, etc.), absolute G coefficient can be preferred (Brennan, 1992; Lee & Frisbie, 1999). In situations, where the place of obtained scores in ranking is important, it will be appropriate to use relative G coefficient. To remove confusion in G coefficients calculated for relative and absolute decisions, the

value calculated for relative decisions is called G coefficient and the value calculated for absolute decisions is called phi (Φ) coefficient or reliability (dependability) coefficient.

Both generalizability (G) coefficient and Φ coefficient take values between 0 and 1. Φ coefficient is a more rigid value when compared to G coefficient. The G coefficient obtained in designs with a single source of variability and completely crossed is interpreted similarly to Cronbach α coefficient included in CTT (Musquash & O'Connor, 2006; Sudweeks, Reeve, & Bradshaw, 2005).

Studies included in G theory can be defined as crossed designs or nested designs. If all levels of a source of variability in a study are present at all levels of other source of variability, this study design is called completely crossed design. For example, if all students in a classroom (b) answer all items in a test (m) and all items of all students are scored by the same raters (p), this design is expressed as completely crossed design. The crossed design is indicated with "x" symbol. The demonstration of the crossed design in the given example is in the form of "b x m x p". On the other hand, if a level of a source of variability is present only at one level of the other and not present at the others, it means this study includes nested design. For example, if, in a written examination, each student answers a different item (m) and each student's answer (b) is evaluated by a different rater (p), it means that this study employs a nested design. Nested design is shown with ":" symbol. The demonstration of the nested design in the given example is in the form of "b : m : p". However, in some studies, both crossed design and nested design are used together and this kind of designs is called mixed design (Brennan, 1992; Shavelson & Webb, 1991). Although G theory can be used in studies employing all these designs expressed here, in order to make predictions related to all sources of variability, in possible cases, the use of completely crossed designs provides an advantage in G theory studies (Kieffer, 1998).

There are two studies in the investigation of reliability in the generalizability theory: 1. Generalizability study (G-study) 2. Decision study (D-study). G study enables making predictions about all sources of variability at the same time and together through the method of ANOVA (Atılgan, 2005; Güler, 2009). Using results obtained from G-study, with D-study one tries to predict cases where error can be minimized for specific aims. And results obtained through D- study help a researcher to make predictions about what results can be reached when he or she changes the number of items, raters or observations (Volpe et al., 2009). D- study, in one sense, can make interpretations similar the aim of using Spearman Brown formula included in CTT (Musquash & O'Connor, 2006). With Spearman Brown formula, prediction of reliability becomes possible according to the change in the number of items included in the measurement tool through which measurement is made. However, in D-study, this prediction is not limited only to number of items but at the same time enables prediction of values which reliability, that is, generalizability and Φ coefficient can take in case of measurement made with a single study including all sources of variability levels. Thus, D-studies help predict most effective measurement cases and reliability (Lee & Fitzpatrick, 2003).

In educational studies where reliability of evaluation based on performance is examined through generalizability theory, it is observed that laboratory skills (Webb, Schlackman, & Sugrue, 2000), success at doing mathematical operations (Güler & Gelbal, 2010; Lane, Liu, Ankenmann, & Stone, 1996), being able to write a composition on a given topic (Baker, Abedi, Linn, & Niemi, 1996; Novak, Herman, & Gearhart, 1996) and skills of

reading fluently (Hintze & Petitte, 2001; Hintze, Owen, Shapiro, & Daly, 2000) are examined. In the field of special education were encountered only two studies where G theory was used in the evaluation based on the measurement of performance. The first of these studies examined the scores which pre-school children with insufficient linguistic and phonological skills took with respect to three basic skills belonging to language competence (Bruckner, Yoder, & McWilliam, 2006) and the other investigated into the reading skills of the students with understanding difficulty (Tindal et al., 2010). However, no studies employing G theory in the reliability of the measurement of basic skills of the students with mental disabilities have been encountered. For this reason, the aim of this study is to examine the effect of task, rater and time on the eating skill occupying a place in the education of having students with mental disabilities acquire self-care skills through generalizability theory.

2. METHODS

2.1. Participants and Implementation

The study is based on the observation of a student attending a private institution at meal times in natural environment. The student observed was registered to the institution with the diagnosis of mental retardation and has epilepsy. The student was observed one day a week (on Tuesdays) continuously for seven weeks. The observations made by a nurse and a psychological counselor who were institution personnel and knew the students closely started in March and ended in June, 2011. Prior to the study, two special education teachers were asked for their opinions about how to make observations. Observers made their evaluations independently from one another.

2.2. Measurement Tool

During observations, evaluations were prepared by Varol (2004) according to the multiple opportunity method and made under the heading of "Skill of Eating by Using the Spoon" by using skill analysis form. The section of the form related to "skill of eating by using the spoon" is composed of 14 items. All the items were evaluated by using a four-point rating including physical help (1), being a model (2), verbal cue (3) and independent (4). The analysis of the scores obtained according to G-theory was made by EDU-G and the scores of reliability coefficients of each rater according to classical test theory were calculated by SPSS.16 statistical package program.

3. RESULTS

As explained in the Introduction section, too, although mostly individuals or students are included as an object of measurement in generalizability theory, this might change depending on the study. In this study, too, there is a single student under measurement and the eating skill of this student is scored at different times. For this reason, the measurement object of this study is occasion. There are two facets in the study, namely steps of the skill (tasks) and raters (raters). The student's skill was scored throughout seven weeks with its all steps by both raters and, in this way, the study is completely composed of crossed design (O x T x R). According to this design, the results related to the components of variance, which were obtained through generalizability analysis are given below in Table 1.

Table 1.
Analysis of Variance Results and Variance Component Estimates for Occasions, Tasks, Raters and Interactions

Source of variance	SS	df	MS	Variance Component Estimates	Percentage of Total Variance Estimates
O	72.20408	6	12.03401	0.39495	28.1
T	42.77551	13	3.29042	0.01557	1.1
R	0.08163	1	0.08163	-0.02211	0.0
OT	83.65306	78	1.07248	0.36355	25.9
OR	1.48980	6	0.24830	-0.00693	0.0
TR	30.48980	13	2.34537	0.28571	20.3
OTR	26.93878	78	0.34537	0.34537	24.6
Total	257.63265	195			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 significance test results (Goodwin & Goodwin, 1991). Also, there are percentages of 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 occasions, tasks and raters. While occasions (object of measurement) account for the largest percentage of the variance (28.1%), the main effect of the task accounts for very small percentage of the variance (1.1%) and the main effect of the rater does not account for any variance. These obtained results exhibit a condition which is required in measurement ideally. The variance resulting from an object of measurement is required to be big, but values regarding other sources of variability are required to be as low as possible. This situation indicates that variability in measurement results does not depend on the rater or tasks. In short, here we mention the inter-rater consistency. On the other hand, when two-way interactions are examined, it is observed that occasion-by-task and task-by-rater account for 25.9% and 20.3% of the total variance respectively. As understood from here, the difficulty level of the steps of the skill show differences depending on time for the student and the scoring of the steps of the skill changes according to the rater as well. When the fact that the student under rating has epilepsy is considered, this situation is not surprising at all. Although the student is expected to improve the skill, which is aimed to be acquired, routinely within the course of time, an epileptic attack in this process might sometimes lead the skill to disappear completely and sometimes most of it to be lost. As another interaction, occasion-by-rater yielded negative variance component estimates. Negative variance values, as suggested by Cronbach et al. (1972), are taken as zero. As required by its definition, variance values cannot take negative values, but like the appearance of a value smaller than 1 in F statistics in ANOVA, variance might appear as negative because of sampling error (Goodwin & Goodwin, 1991). This situation is an indication of the fact that raters rating students at the same time and independently from one another do ratings which are totally consistent with one another. At last, the three way- interaction, occasions-by-tasks-by-raters, is also named as “residual” or “error” in the ANOVA model used here. If measurement results obtained in the study are reliable, this value belonging to the residual is expected to be as low as possible. According to Table 1, the three-way interaction accounted for 24.6% of the total variance. According

to G theory, this obtained variance value is required to be as small as possible. This value indicates that the change in the scores might have appeared depending on different sources of variability not included in the study. As a result, as also understood from Table 1, as an advantage of G theory, the researcher is able to see clearly what extent of the total variance appeared as a result of the interaction of which source or sources (Güler, 2009).

In G theory, the G coefficient which might correspond to the reliability coefficient in classical test theory is calculated. G and Φ coefficients included in the study and calculated over 14 tasks and 2 raters were found to be .91 and .89 respectively. Moreover, the calculation of G and Φ coefficients by using the values in Table 1 is shown in detail in Table 2.

Table 2.

Calculation of G coefficient

I. G-coefficient for 14 tasks and 2 raters (n_t:14, n_r:2)

$$E\hat{\rho}^2(T,R) = \frac{\hat{\sigma}_o^2}{\hat{\sigma}_o^2 + \frac{1}{n_t}\hat{\sigma}_{ot}^2 + \frac{1}{n_r}\hat{\sigma}_{or}^2 + \frac{1}{n_t n_r}\hat{\sigma}_{otr}^2}$$

$$= \frac{.40}{.40 + \frac{1}{14}.36 + \frac{1}{2}0.0 + \frac{1}{128}.35}$$

$$= \frac{.40}{.4385}$$

$$=.91$$

II. Φ -coefficient for 14 tasks and 2 raters (n_t:14, n_r:2)

$$\Phi(T,R) = \frac{\hat{\sigma}_o^2}{\hat{\sigma}_o^2 + \frac{1}{n_t}\hat{\sigma}_t^2 + \frac{1}{n_r}\hat{\sigma}_r^2 + \frac{1}{n_t}\hat{\sigma}_{ot}^2 + \frac{1}{n_r}\hat{\sigma}_{or}^2 + \frac{1}{n_t n_r}\hat{\sigma}_{tr}^2 + \frac{1}{n_t n_r}\hat{\sigma}_{otr}^2}$$

$$= \frac{.40}{.40 + \frac{1}{14}.02 + \frac{1}{2}0.0 + \frac{1}{14}.36 + \frac{1}{2}0.0 + \frac{1}{28}.29 + \frac{1}{28}.35}$$

$$= \frac{.40}{.4495}$$

$$=.89$$

As stated in Table 2, too, in G theory, by using the results obtained from G-study, the conditions where the error can be minimized for specific purposes through D-study are tried to be predicted. In D-study, in case of a decrease or an increase in the number of tasks or raters, the values which reliability, in other words, generalizability and Φ coefficient might take are predicted. The number of tasks included in the measurement tool used in this study is not certain. However, in case of an increase or a decrease in the number of raters, the extent of the change in the reliability has been investigated. The results of the D-study are given in Table 3.

Table 3.
G and Φ coefficients of D Studies (n_t:14)

	1 rater	**2 rater***	3 rater	4 rater	5 rater
G-coefficient	.89	**.91**	.92	.92	.93
Φ-coefficient	.85	**.89**	.90	.91	.92

(*Number of raters taking part in the study)

As seen in Table 3, increasing the number of the raters increases the value of reliability greatly. For this reason, increasing the number of the raters is not expected to make an important contribution to further studies. Cronbach α values calculated according to the classical test theory related to the scores of each rater included in the study were found to be .907 and .867, and, G-coefficient values calculated according to G theory over a single facet (o x t) according to the completely crossed design method were found to be .91 and .87 respectively. The G-coefficient calculated over 1 rater with D-study was found to be .89, which appears to be predicted as close to these values.

4. DISCUSSION

As seen in this study, too, in measurement situations where there are many sources such as task and rater which cause variability in measurement results, G theory provides detailed information through a single analysis. Especially in situations such as education and psychology where individuals' behaviors are evaluated through observation, for observation results to be objective, it is a frequently encountered situation that more than one rater takes part. In this kind of scorings, consistency between raters is of particular importance. G theory is a suitable method of determining reliability which can be used in situations where more than one rater makes scorings.

It is known that criterion-dependent measurement tools provides opportunities to determine initial levels of students with intellectual disability in terms of behavior whose education is to be performed, record developments in education process objectively and continuously develop the education program under implementation in the direction of pieces of feedback (Varol, 2004). In this study, the eating skills criterion-dependent measurement tool prepared in accordance with multiple-opportunities method was examined from the perspective of G theory, which enables the examination of many possible sources of error and inter-rater consistency. As a conclusion, this study made to reveal the reliability of the criterion-dependent measurement tool used in special education to make important decisions is expected to popularize the use of the measurement tool in question and light the way for scientific research studies to be made with using this measurement tool.

In this study, as a source of variability, only rater and task were used. It might as well be possible to assess the reliability of eating criterion dependent measurement tool through studies including different and more number of variability sources (environment, different forms, etc.). Moreover, it can be suggested that similar studies should be made on other criterion dependent measurement tools used with the aim of assessing different skills in special education.

REFERENCES

Akköse, M.C. (2008). *The Effectiveness of Simultaneous Prompting on Teaching Naming Kitchen Tools to Children with Developmental Disabilities: A Multiple Exemplar Approach Instruction*. Unpublished master's thesis, Anadolu University Institute of Educational Sciences, Eskişehir.

Akmanoğlu, N., & Batu, S. (2004). Teaching pointing to numerals to individuals with autism using simultaneous prompting. *Education and Training in Developmental Disabilities, 39*(4), 326-336.

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 thesis, The Pennsylvania State University, Pennsylvania.

Atılgan, H. (2005). Generalizability theory and a sample application for inter-reliability. *The Journal of Educational Sciences and Practice, 4,* 24-31.

Atılgan, H., & Tezbaşaran, A.A. (2005). An investigation on consistency of G and Phi coefficients obtained by generalizability theory alternative decisions study for scenarios and actual cases. *Eurasian Journal of Educational Research, 18*, 236-252.

Baker, E., Abedi, J., Linn, R., & Niemi, D. (1996). Dimensionality and generalizability of domain-independent performance assessments. *Journal of Educational Research, 89*(4), 197-205.

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

Brennan, R.L. (1992). *Elements of Generalizability Theory*. New York, NY: Springer-Verlog.

Bruckner, C.T., Yoder, P.J., & McWilliam, R.A. (2006). Generalizability and decision studies: An example using conversational language samples. *Journal of Early Intervention, 28,* 139-153.

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

Crocker, L. and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Orlando, FL: Holt, Rinehart and Winston.

Goodwin, L.D. & Goodwin, W.L. (1991). Research Notes: Using Generalizability Theory in Early Childhood Special Education. *Journal of Early Intervention, 15*, 193-204.

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*, 93-103.

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*, 989-1019.

Gürsel, O. (1993). *Zihinsel engelli çocukların doğal sayıları, gerçek nesneler kullanılarak eşleme resimleri işaret ederek gösterme, rakamlar gösterildiğinde söyleme becerilerinin gerçekleştirilmesinde basamaklı öğretim yöntemiyle sunulan bireyselleştirilmiş öğretim materyalinin etkililiği*. Eskişehir: Anadolu Üniversitesi Yayınları.

Heward, W.L. (1996). *Exceptional Children: An Introduction to Special Education*. USA: Prentice Hall.

Hintze, J.M., & Matthews, W.J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral assessment. *School Psychology Review*, *33*, 258-270.

Hintze, J.M. & Petitte, H.A.P. (2001). The generalizability of CBM oral reading fluency measures across general and special education. *Journal of Psychoeducational Assessment, 19*(2), 158-170.

Hintze, J.M., Owen, S.V., Shapiro, E.S., & Daly, E.J. (2000). Generalizability of oral reading fluency measures: Application of G-theory to curriculum-based measurement. *School Psychology Quarterly, 15*(1), 52-68.

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.

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

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.

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*, 433-439.

Musquash, C., & O'Connor, B.P. (2006). SPSS and SAS Programs for Generalizability Theory Analysis. *Behavior Research Methods, 38*(3), 542-547.

Novak, J.R., Herman, J.L., & Gearhart, M. (1996). Establishing validity for performace-based assessments: An illustration for collections of student writing. *The Journal of Educational Research, 89*(4), 220-233.

Özkan, Ş.Y., & Gürsel, O. (2006). The Effectiveness of Simultaneous Prompting on Teaching Photo Copy Skills to Students with Mental Disabilities. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Özel Eğitim Dergisi, 7*, 29-45.

Özyürek, M. (1996). *Sınıfta Davranış Yönetimi: Uygulamalı Davranış Analizi*. Ankara: Karatepe Yayınları.

Parrott, K.A., Schuster, J.W., Collins, B.C., & Gassaway, L.J. (2000). Simultaneous prompting and instructive feedback when teaching chained tasks. *Journal of Behavioral Education, 10,* 3-19.

Reckase, M.D. (1995). The Reliability of Ratings Versus Reliability Scores. *Educational Measurement: Issues and Practice, 14(*4), 31.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Newburry Park CA: Sage.

Sudweeks, R.R., Reeve, S., & Bradshaw, W.S. (2005). A comparison of generalizability theory and many facet measurement in analysis of college sophomore writing. *Assessing Writing, 9,* 236-261.

Tindal, G., Yovanoff, P., & Geller, J.P. (2010). Generalizability theory applied to reading assessments for students with significant cognitive disabilities. *The Journal of Special Education, 44*(1), 3-17.

Topsakal, M., & Düzkantar, A.U. (2010). The Effectiveness of Simultaneous Prompting in Teaching Car Washing to Children with Mental Retardation by Correcting Error. *Abant İzzet Baysal Üniversitesi Dergisi, 10*, 79-94.

Varol, N. (1996). Beceri Öğretimi Materyali Geliştirme ve Beceri Öğretiminde İpuçlarının Kullanımı. *Gazi Üniversitesi Eğitim Fakültesi Dergisi, 16*(1), 35-46.

Varol, N. (2004). *Öz Bakım Becerilerinin Öğretimi*. Ankara: Kök Yayınevi.

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*, 382-401.

Webb, N.M., & Shavelson, R.J. (2005). Generalizability theory: Overview. B.S. Everitt & D.C. Howell (Eds.), In *Encyclopedia of Statistics in Behavioral Science* (pp.717-719). Hoboken, NJ: Wiley.

Webb, N.M., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education, 13*, 277-301.

GENİŞ ÖZET

1. GİRİŞ

Zihinsel engelli öğrencilere bağımsız yaşama becerileri kazandırmada en etkili yöntemlerden birisi uygulamalı davranış analizidir. Bu analiz, engelli bireyde istenilen davranış değişikliğini sağlayabilmek için davranış öncesi ve sonrası uyaranların sistematik olarak düzenlenmesini gerektirir (Heward, 1996). Uygulamalı davranış analizinin temel noktasını, davranışın doğrudan ve sürekli ölçülmesi oluşturmaktadır. Doğrudan ve sürekli ölçümlerde öğrencinin performansı, davranışın oluştuğu ortamlarda doğrudan gözlenerek sürekli değerlendirilmektedir (Özyürek, 1996) ve bu ölçümlerde ölçüt bağımlı ölçme araçları kullanılmaktadır (Varol, 1996).

Ölçüt bağımlı ölçme araçları; bildirimler, ölçüt ve sorular bölümünden oluşur. Ölçüt aracının bildirimler bölümü, analizi yapılan becerinin basamaklarını ve belirlenen ölçütleri içerir. Eğer ölçüt bağımlı ölçme aracının öğrenciye bir kez uygulanması sonucunda öğrencinin performans düzeyi saptanacaksa, ölçüt olarak %100 benimsenmektedir. Ölçüt bağımlı ölçme aracının sorular/görevler bölümü ise performans düzeyinin belirlenmesinde kullanılacak yönteme göre düzenlenmektedir. Bu bölüme, becerinin ana yönergesi ve bağımsız sütunu eklenerek tek fırsat yöntemiyle performans düzeyi belirlemeye yönelik; ana yönerge, bağımsız, sözel ipucu, model olma ve fiziksel yardım sütunları eklenerek çoklu fırsat yöntemiyle performans düzeyi belirlemeye yönelik ölçüt bağımlı ölçme aracı hazırlanır (Varol, 1996).

Tek fırsat yöntemi, öğrenciye sadece ana yönerge (beceriye sahip bir kişiye verildiğinde, becerinin gerçekleştirilmesini sağlayan yönerge) verilerek yapabildiklerinin doğrudan gözlenmesi ve kaydedilmesidir. Bu yöntemin kullanılmasındaki amaç; öğrencinin, becerinin ne kadarını bağımsız olarak gerçekleştirdiğini saptamaktır. Çoklu fırsat yöntemi ise, öğrenciye ana yönergenin verilmesinden sonra öğrencinin yapmakta zorlandığı basamakları yerine getirmesi için yeni fırsatlar verilmesidir. Bu yöntemi kullanmanın amacı ise öğrencinin becerinin her bir basamağını hangi ipucunu kullanarak gerçekleştirdiğini ya da beceriyi bağımsız olarak gerçekleştirip gerçekleştirmediğini saptamaktır (Varol, 1996).

Eğitimde kullanılan tüm ölçme araçlarında olduğu gibi bu tür çalışmalarda da gözleme dayalı puanlamanın güvenirliği en önemli konulardan biridir (Goodwin ve Goodwin, 1991). Ancak gözlemle puanlama yapmanın en büyük dezavantajı ise sübjektifliğidir. Bu sebeplerdir ki, her bir öğrencinin davranışının gözlemlenerek öğrenciye verilen puanın daha objektif ve güvenilir olmasını sağlayabilmek için çoğunlukla ya bir puanlayıcının farklı zamanlarda yaptığı birden fazla puanlamanın ortalaması ya da aynı zamanda birden fazla puanlayıcı puanlarının ortalaması alınır. Aynı zamanda, eğitim ya da klinik ortamlarda tek bir öğrencinin davranışının değerlendirilmesi de benzer şekilde yapılmaktadır. Eğitim ve Psikolojik Test Standartlarının da (Amerikan Eğitim Araştırmaları Birliği, Amerikan Psikologlar Birliği, 1999) belirttiği üzere, "ölçülen davranışın niteliği ne olursa olsun, belirli eğitimsel kararların verilmesinde kullanılan ölçme sonuçlarının güvenirliği ve geçerliği belirlenmesi gerekir" (akt:Lei, Smith ve Suen, 2007).

Özel eğitime ilişkin yurtiçi ve yurtdışı araştırmalarda, performansın ölçülmesine dayalı değerlendirmede değişkenlik kaynağı olarak yalnızca puanlayıcıların dikkate alındığı ve

puanlayıcılar arası güvenirliğin hesaplanmasında Klasik Test Kuramı'na ilişkin yöntemlerden (Uyum indeksi, Pearson korelasyon katsayısı, t testi veya varyans analizi, grup içi değişkenlik katsayısı) yararlanıldığı görülmektedir (Akmanoğlu ve Batu, 2004; Özkan ve Gürsel, 2006; Topsakal ve Düzkantar, 2010; Parrott, Schuster, Collins ve Gassaway, 2000; Akköse, 2008). Klasik Test Kuramı'na dayalı güvenirliğin belirlenmesinde kullanılan istatiksel yöntemlerin en ciddi sınırlılığı hata kaynağı olarak yalnızca değerlendiriciler arası tutarsızlığa odaklanmalarıdır. Genellenebilirlik Kuramı (Cronbach, Gleser, Nanda&Ratjaratman, 1972) ölçmedeki hatanın kaynağını tek bir değişkenlik kaynağı ile açıklanmanın (örneğin; sadece puanlayıcı değişkenlik kaynağına bağlı olarak) sınırlılığına karşı ölçme hatasının farklı değişkenlik kaynaklarından kestirilmesine olanak vermektedir. Böylece ölçme konusu olan bireylerin (ölçme objelerinin) gözlenen puanları evren puanlarına (gerçek puanlara) olabildiğince doğru bir şekilde kestirilebilmektedir (Atılgan ve Tezbaşaran, 2005; Güler ve Gelbal, 2010).

Genellenebilirlik (G) Kuramı ölçme sonuçlarının güvenirliğinin belirlenmesini, güvenilir gözlemlerin tasarımını, araştırılmasını ve kavramsallaştırılmasını sağlayan istatistiksel bir kuramdır (Cronbach ve diğerleri, 1972; Brennan, 2001). G Kuramı bir grup bireyden - hatta bazen sadece tek bir bireyden (Lei, Smith ve Suen, 2007) - elde edilen ölçme sonuçlarının, bu sonuçların elde edildiği belirli sayıdaki maddeleri, puanlayıcıların ya da durumların çok daha ötesine genellenebilmesi amacını taşır (Brennan, 1992; Shavelson ve Webb, 1991). Shavelson ve Webb (1991)'e göre, G Kuramı dört farklı açıdan Klasik Test Kuramı'nın daha genişletilmiş bir halidir: 1. Genellenebilirlik Kuramı, çoklu varyans kaynaklarını tek bir analizde ele alır. 2. Her bir varyans kaynağının büyüklüğünün belirlenmesini sağlar. 3. Hem bireylerin performanslarına dayalı göreceli kararlar hem de bireylerin performanslarıyla ilgili mutlak kararlar alınmasına ilişkin iki farklı güvenirlik katsayısının hesaplanmasına olanak tanır. 4. Belirli bir amaca bağlı olarak, ölçme hatasının en aza indirgenebileceği ölçmelerin düzenlenmesine (D-çalışmaları) imkân tanır. Kısacası G Kuramı, farklı hata kaynaklarının olası olduğu performansın ölçülmesiyle elde edilen sonuçların güvenirliğinin kestirilmesine uygun bir kuramdır.

Performansa dayalı değerlendirmenin güvenirliğinin Genellenebilirlik Kuramı ile incelendiği eğitimle ilişkili araştırmalarda; laboratuvar becerilerinin (Webb, Schlackman, &Sugrue, 2000), matematiksel işlem yapabilmedeki başarının (Güler ve Gelbal, 2010; Lane, Liu, Ankenmann & Stone, 1996), verilen bir konuda kompozisyon yazabilmenin (Gierl, 1998; Novak, Herman, &Gearhart, 1996; Baker, Abedi, Linn&Niemi, 1995) ve akıcı şekilde okuma becerilerinin (Hintze&Petitte, 2001; Hintze, Owen, Shapiro, &Daly, 2000) incelendiği görülmektedir. Özel eğitim alanında performansın ölçülmesine dayalı değerlendirmede G Kuramı'nın kullanıldığı iki araştırmaya rastlanabilmiştir. Bu araştırmaların ilkinde, dilbilimsel ve fonolojik açıdan yetersizliği bulunan okulöncesi çocukların dil yeteneğine ilişkin üç temel becerilere ait puanları (Bruckner, Yoder ve McWilliam, 2006), diğerinde ise anlama güçlüğü çeken öğrencilerin okuma becerileri (Tindal, Yovanoff&Geller, 2010) incelenmiştir. Buna karşın, özel eğitime gereksinim duyan öğrencilerin temel becerilerin ölçülmesinin güvenirliğinde G Kuramı'nın kullanıldığı herhangi bir çalışmaya rastlanmamıştır. Bu nedenle bu araştırmanın amacı görev, puanlayıcı ve zamanın, zihinsel engelli öğrencilere öz-bakım becerileri kazandırma eğitiminde önemli bir yer tutan yemek yeme becerisi üzerindeki etkisini Genellenebilirlik Kuramı ile incelemektir.

2. YÖNTEM

Katılımcılar ve Uygulama

Araştırma, özel bir kuruma devam eden bir öğrencinin yemek saatlerinde doğal ortamda gözlenmesine dayanmaktadır. Gözlenen öğrenci, mentalretardasyon tanısıyla kuruma kaydolmuş ve epilepsi hastasıdır. Öğrenci, haftanın bir günü (salı günleri) sürekli olarak yedi hafta boyunca gözlenmiştir. Kurum personeli olan ve öğrencileri yakından tanıyan bir hemşire ve bir rehber öğretmen tarafından yapılan gözlemler, 2011 yılının mart ayında başlamış ve haziran ayında sona ermiştir. Araştırmaya başlanmadan önce gözlemlerin nasıl yapılması gerektiği konusunda iki özel eğitim öğretmeninin görüşü alınmıştır. Gözlemciler değerlendirmelerini birbirlerinden bağımsız olarak gerçekleştirmiştir.

Ölçme Aracı

Gözlemler sırasında değerlendirmeler, Varol (2004) tarafından çoklu fırsat yöntemine göre hazırlanmış "Kaşığı Kullanarak Yemek Yeme Becerisi" başlığı altında, beceri analizi formu kullanılarak yapılmıştır. Formun "kaşığı kullanarak yemek yeme becerisine" ilişkin bölümü 14 maddeden oluşmaktadır. Bütün maddeler fiziksel yardım (1), model olma (2), sözel ipucu (3) ve bağımsız (4) olmak üzere dörtlü bir derecelendirme kullanılarak değerlendirilmiştir. Elde edilen puanların G Kuramı'na göre analizi EDU-G, her bir puanlayıcının puanlarının Klasik Test Kuramı'na göre güvenirlik katsayıları SPSS.16 bilgisayar paket programıyla yapılmıştır.

Bulgular

Giriş bölümünde de açıklandığı üzere, Genellenebilirlik Kuramı'nda ölçme objesi olarak çoğunlukla bireyler ya da öğrenciler alınmakla birlikte, çalışmaya bağlı olarak bu durum değişebilmektedir. Bu çalışmada da ölçülen tek bir öğrenci bulunmakta ve bu öğrencinin yemek yeme becerisi farklı zamanlarda puanlanmaktadır. Bu sebeple, bu çalışmanın ölçme objesi "zaman (occasion)" olmaktadır. Becerinin basamakları (task) ve puanlayıcılar (rater) olmak üzere çalışmada iki yüzey bulunmaktadır. Öğrencinin becerisi, yedi haftanın tümünde tüm basamaklarıyla her iki puanlayıcı tarafından puanlanmış, böylelikle çalışma tümüyle çaprazlanmış desen (O x T x R) oluşturmaktadır.

Analiz sonuçlarına göre, ölçme objesi olan zaman değişkenliği açıklamada en yüksek orana sahipken (28.1%), değişkenliği açıklamada; görev ana etkisi oldukça düşük bir yüzdeye (1.1%) sahiptir ve puanlayıcı ana etkisi varyansı ise sıfırdır. Elde edilen bu sonuçlar, ölçmede ideal olarak istenen bir durumu sergilemektedir. Ölçme objesinden kaynaklı varyansın büyük olması; diğer değişkenlik kaynaklarına ilişkin değerlerin ise olabildiğince düşük olması istenir. Bu durum, ölçme sonuçlarındaki değişkenliğin puanlayıcı ya da görevlere bağlı olmadığını göstermektedir. Kısacası puanlayıcılar arası tutarlılık söz konusudur. Diğer taraftan ikili etkileşimlere bakıldığında; zaman-görev ve görev-puanlayıcı etkileşimleri sırasıyla, değişkenliğin %25.9'unu ve %20.3'ünü açıklamaktadır. Buradan anlaşılacağı üzere, beceri basamaklarının zorluk düzeyi öğrenci için zamana göre farklılık göstermekte ve beceri basamaklarının puanlanması da puanlayıcılara göre farklılaşmaktadır. Puanlanan bireyin epilepsi hastası olduğu

düşünüldüğünde, bu durum hiç de şaşırtıcı olmamaktadır. Kazandırılmaya çalışılan beceriyi öğrencinin rutin olarak zaman içinde ilerletmesi beklenirken, bu süreçte geçirmiş olduğu bir nöbet, becerinin bazen tamamen kaybolmasına bazen de çok büyük bir kısmının yitirilmesine sebep olabilmektedir. Bir diğer etkileşim olan, zaman-puanlayıcı etkileşiminin ise negatif varyans kestirimine sahip olduğu görülmektedir. Negatif varyans değerleri, Cronbach ve diğerleri (1972)'nin de önerdiği gibi sıfır olarak alınmaktadır.

En son olarak, zaman-görev-puanlayıcı etkileşimi –ANOVA modelinde artık ya da hata terimi olarak da isimlendirilir- yer almaktadır. Eğer çalışmada, ölçme sonuçları güvenilir ise artığa ait olan bu değerin olabildiğince küçük olması istenir. Elde edilen sonuçlara göre, bu etkileşimin toplam varyansın % 24.6'sını açıkladığı gözlenmektedir. G Kuramı'na göre, elde edilen bu varyans değerinin olabildiğince küçük olması istenir. Bu değer, puanlardaki değişimin çalışmada yer almayan farklı değişkenlik kaynaklarına bağlı ortaya çıkmış olabileceğinin sinyalini vermektedir. Sonuç olarak, G Kuramı'nın bir avantajı olarak araştırmacı, toplam varyansın ne kadarının hangi kaynak ya da kaynakların etkileşimi sonucu oluştuğunu açıkça görebilmektedir (Güler, 2009).

Çalışmada yer alan 14 görev ve 2 puanlayıcı üzerinden hesaplanan G ve Φ katsayısıları sırasıyla .91 ve .89 olarak bulunmuş ve her bir puanlayıcı puanlarına ilişkin Klasik Test Kuramı'na göre hesaplanan Cronbach α değerleri de .907 ve .867 olarak hesaplanmış, aynı sırayla G Kuramı'na göre tek yüzey üzerinden (o x t) tümüyle çapraz desene göre hesaplanan G katsayı değerleri de .91 ve .87 olarak hesaplanmıştır. D çalışmasıyla bir puanlayıcı üzerinden hesaplanan G katsayısının (.89) da gerçekteki bu değerlere yakın bir değer olarak kestirildiği görülmektedir.

3. SONUÇLAR

Çalışma sonuçlarından anlaşılacağı üzere, ölçme sonuçlarındaki değişkenliğe sebep olan; görev, puanlayıcı gibi pek çok kaynağın bulunduğu ölçme durumlarında G Kuramı tek bir analizle ayrıntılı bilgi sağlamaktadır. Özellikle eğitim ve psikoloji gibi bireylerin davranışlarının gözlenerek değerlendirildiği durumlarda; gözlem sonuçlarının objektif olması için birden fazla puanlayıcının yer alması sıklıkla rastlanan bir durumdur. Bu tür puanlamalarda puanlayıcılar arası tutarlılık da ayrı bir önem taşımaktadır. G Kuramı, birden fazla puanlayıcının puanlama yaptığı durumlarda kullanılabilecek uygun bir güvenirlik belirleme yöntemi olarak tercih edilebilir.