EJAES

JOURNAL OF ADVANCED EDUCATION STUDIES

İleri Eğitim Çalışmaları Dergisi

6 (1):67-91, 2024

RELIABILITY OF INTERVIEWS USED FOR ADMISSION TO TEACHER EDUCATION PROGRAMS IN NORTHERN CYPRUS: GENERALIZABILITY THEORY

Hasan Özder[1]

Aygil Takır[2]

Geliş Tarihi/Received: 05.01.2024 Elektronik Yayın / Online Published: 15.06.2024

DOI: 10.48166/ejaes.1415158

ABSTRACT

The aim of this study was to investigate the reliability of raters' ratings using the same rating guide in two different interviews (Interview 1 and Interview 2) for university admission of prospective teachers. A total of fifty-eight prospective teachers and 10 raters participated in the study. The raters gave their ratings according to five dimensions, namely general culture (GC), language (L), self-image (SI), hobbies (H), and attitude towards the teaching profession (ATTP) in two interviews. The data were analyzed using the Mann-Whitney U-test and Generalizability Theory. The results of the study show that there is no significant difference between the raters' ratings in each interview. The G coefficients were unacceptable for the GC and SI dimensions in Interview 1. In addition, the G coefficients were relatively low for the L, H, and ATTP dimensions in Interview 1 and all dimensions in Interview 2.

Keywords: Admission to teacher education programs, interview, reliability, generalizability theory

[1] Ataturk Teachers Academy, hasan.ozder@aoa.edu.tr, ORCID: 0000-0003-1094-3590,

[2] Eastern Mediterranean University, aygil.takir@emu.edu.tr, ORCID:0000-0003-3042-7585

KUZEY KIBRIS'TA ÖĞRETMEN EĞİTİMİ PROGRAMLARINA KABULDE KULLANILAN MÜLAKATLARIN GÜVENİLİRLİĞİ: GENELLENEBİLİRLİK KURAMI

ÖZET

Bu çalışmanın amacı, aday öğretmenlerin üniversiteye kabulü için yapılan iki farklı mülakatta (Mülakat 1 ve Mülakat 2) aynı derecelendirme kılavuzunu kullanan puanlayıcıların puanlamalarının güvenilirliğini araştırmaktır. Çalışmaya toplam elli sekiz öğretmen adayı ve 10 puanlayıcı katılmıştır. Puanlayıcılar iki mülakatta genel kültür (GK), dil (D), öz-imaj (Öİ), hobiler (H) ve öğretmenlik mesleğine yönelik tutum (ÖMT) olmak üzere beş boyuta göre puanlama yapmışlardır. Veriler Mann-Whitney U-testi ve Genellenebilirlik Teorisi kullanılarak analiz edilmiştir. Çalışmanın sonuçları, her bir mülakatta puanlayıcıların puanlamaları arasında anlamlı bir fark olmadığını göstermektedir. Mülakat 1'de GK ve ÖI boyutları için G katsayıları kabul edilemez düzeydedir. Ayrıca, Mülakat 1'de L, H ve ÖMT boyutları ve Mülakat 2'de tüm boyutlar için G katsayıları nispeten düşüktür.

Anahtar Kelimeler: Öğretmen eğitimi programlarına kabul, mülakat, güvenilirlik, genellenebilirlik kuramı

1. INTRODUCTION

There are several studies aimed at examining the ideal/effective teacher. The results of these studies indicate that affective characteristics that facilitate teaching and learning are one of the key competencies of an ideal/effective teacher (Taneri, 2017). One of the most comprehensive specifications of affective qualities of ideal or effective teachers was developed by Stronge (2007). In his study, he emphasised the importance of caring teachers who can be defined as kind, gentle, and encouraging and make students feel capable and important. Arnon and Reichel (2007) also defined the qualities of ideal/effective teachers mainly based on affective qualities as having a sense of humour, being kind-hearted, fair, optimistic, emphatic, flexible, confident, caring and polite.

Teachers' attitude toward the teaching profession plays an important role in teacher effectiveness (Stronge, 2007). According to Schulte, Slate, and Onwuegbuzie (2008), an ideal/effective teacher is willing to collaborate with other teachers and share ideas and strategies. In addition, an ideal/effective teacher is committed to the profession and professional development. For this purpose, an effective teacher spends time learning new teaching strategies, new content and teacher-related tasks.

Some other studies in the literature show that one of the most important characteristics of ideal/effective teachers is personality traits (i.e. Lupascu, Pânisoarâ, & Pânisoarâ, 2014; Leger, 2014). On the other hand, many teacher education programs accept prospective teachers based only on academic criteria such as cognitive tests or examinations/standardized test scores (Haberman & Post, 1998). Research has shown that the assessment of prospective teachers on academic criteria alone does not affect teacher effectiveness (Bardach & Klassen, 2020) or has a mainly poor effect on the prediction of success in the classroom (Corcoran & O'Flaherty, 2018). The important characteristics that affect teachers' performance in the classroom include affective characteristics such as verbal expressiveness (Andrew, Cobb, & Giampietro, 2005), personality traits (Kim, Jörg, & Klassen, 2019) and leadership

skills (Byrnes et al., 2003). In other words, prospective teachers' inclinations, skills and attitudes influence what they will do in their classrooms in future (Taneri, 2017).

The decision to admit to a teacher education program is critical because the retention and graduation rates of teacher candidates in these programs are high, and thus they serve as unofficial gatekeepers to the profession (Thomson et al., 2011). Currently, some teacher education programs use a variety of admission variables to determine which applicants are admitted to the program. The most common variables include GPA, written responses, letters of reference, and work experience. In some countries, interviews are successfully used for the admission of prospective teachers to teacher training institutions. In the United States, some of the teacher education programs utilize individual and group interviews to determine which applicants are admitted into a program together with the other variables. Prospective teachers are also interviewed for admission to faculties of education in the United Kingdom (UK), where it is determined whether prospective teachers are physically and mentally suited to the teaching profession.

Individual interviews focus on the dimensions of qualities, for example, knowledge, personal qualifications, and verbal communication. Individual interviews can be a) loosely structured (a minimum of guidelines), b) moderately structured (panel interviews, predetermined scoring and questionnaires), c) highly structured (predefined questions and sample answers, panel interviews, training of interviewers and constant evaluation of the process) (Goho & Blackman, 2006).

Among these interview techniques, Pursell, Campion, and Gaylord (1980) recommended the use of the structured interview. According to them, structured interviews increase reliability and reduce subjectivity. Furthermore, a highly structured interview format has been suggested to function as a remedy for interviewer bias (Ebmeier & Ng, 2005). Structured interviews are more amenable to analysis than unstructured interviews, but the assessment that an interview uses should be developed through careful analysis of the research findings (Pellicer, 1981). Unstructured interviews are criticized for their low reliability (Petrarca & LeSage, 2014; Smith & Pratt, 1996) and invalid predictors of job performance and success (Mathis & Jackson, 2011).

Group interviews can also be used to facilitate the comparison of applicants for the same/similar positions and are a good tool to assess the common skills of applicants (Tran & Blackman, 2006). In educational research, group interviews have been used to select prospective teachers for teacher training programmes, showing that group interviews can be reliable and predictive of performance (Byrnes et al. 2003; Faulk, 2008; Shechtman, 1992). In their study, Byrnes et al. (2003) found that the results of group assessment interviews measuring psychomotor and affective behaviours of prospective teachers predicted classroom performance better than academic criteria. Faulk (2008) also conducted a study to investigate the ability of a group assessment procedure used as admission criteria into a teacher education program to predict future teaching success. Group interviews appear to be a useful tool for identifying prospective teachers who are more likely to succeed in the teaching profession. On the other hand, the available research is concerned with the validity of group interviews like individual interviews

discussed in the previous paragraphs (Huffcutt & Woehr, 1999; McDaniel, Whetzel, Schmidt, & Maurer, 1994).

As can be seen from the previous sections, one of the biggest problems in admitting prospective teachers to a teacher education programme is the reliability of interviews (Casey & Childs, 2007). In other words, the use of interviews for admission decisions is controversial, as there is no evidence of validity and reliability for most teacher education programmes. Researchers have suggested that reliability issues can be strengthened through rater training (Donnon & Paolucci, 2008; Jonsson & Svingby, 2007) and by increasing the number of raters per application (Byrnes et al., 2000; Caskey et al., 2001; Smith & Pratt, 1996). There are some methods and theories to assess the reliability of interviews through raters' scores (i.e. Fleiss' Kappa, Inter-rater Reliability, Rasch Model). Generalizability theory (G theory) is also used for this purpose (Wing & Chiu, 2001).

1.1 Theoretical Framework and Literature Review

G theory is a statistical theory for evaluating the reliability of behavioural measurements (Webb & Shavelson, 2018). G theory is an approach to estimating measurement accuracy in situations where measurements are subject to multiple sources of error (Cardinet, Johnson, & Pini, 2011). It is an approach that not only provides a means of estimating the reliability of measurements that have already been made but also allows information about error contributions to be used to improve measurement procedures in future applications. Substantially, it can be said that G theory estimates multiple sources of measurement error and allows decision-makers to design a measurement procedure that minimizes error.

In Classical Test Theory (CTT), observed score variance is simply the sum of true score variance and error variances. Reliability coefficient can be defined as the ratio of true score variance to observed score variance (Crocker & Algina, 1986). G theory liberalises CTT and ANOVA and extends traditional notions of reliability (Brennan, 2001).

G theory is a flexible and powerful psychometric perspective in at least two important ways. First, it expands the conceptualization of reliability to account for the possibility that multiple facets can systematically affect the quality of a measurement strategy. Second, it provides statistical tools for evaluating the effects of each facet of a measurement design and for planning measurement designs that maximize quality and efficiency (Furr, 2011: 123-124).

Cronbach et al. (1963) explained the concept of G theory as follows:

> A researcher asks about the accuracy or reliability of a measurement because he/she wants to generalize from the observations at hand to a class of observations to which it belongs...For example, when asking about the reliability of an essay grade, one wants to know how representative that grade is of the grades given to the same paper by other raters, or of the grades given to other papers by the same subject.

The term facet in the previous paragraph was originally introduced by Guttman and adopted by Cronbach and his collaborators in the early presentation of the developing G theory. Regardless of the

difference in terminology, direct parallels can be drawn between facets in G theory and factors in experimental design models (Briesch, Swaminathan, Welsh, M., & Chafouleas, 2014). A facet in G theory is equivalent to a factor in ANOVA (Atılgan & Tezbasaran, 2005; Cardinet, et al., 2011).

The most important aspect and distinctive feature of generalizability theory is its conceptual framework (Brennan, 2001). Two types of studies are conducted in the application of G theory: Generalizability (G) studies and Decision (D) studies (Li et al., 2015). The goal of a G study is to broadly define the population of admissible observations and thus estimate as many sources of variance as are potentially relevant to identify the main sources of measurement error. Specifically, a G study focuses on estimating the magnitude of measurement error attributable to different sources of variance. The information obtained from a G-study is then used as the basis for subsequent D studies, where the goal is to develop a measurement that minimizes error for a specific purpose (Briesch et al., 2014).

When planning a G theory, the first thing a researcher must do is identify the facets that play a role in the measurement process and the relationships between them. Facets can either be fully crossed or nested within other facets. Two facets are crossed when each level of one of the facets is combined with each level of the other facet in a data set.

A reliability coefficient (G) summarizes the results of a G study. It indicates the extent to which the measurement instrument or procedure used can reliably differentiate between the persons/objects involved. In other words, it tells whether the results obtained are satisfactorily reliable, regardless of the specific components that define the particular instrument or procedure. G coefficients take values between 0 (completely unreliable measurement) and 1 (perfectly reliable measurement) (Cardinet, et al., 2011).

The last phase, namely D study, aims to improve the procedure based on an analysis of its characteristics. For this purpose, an optimization design is defined. In this last step, the results of the G study are used, especially the estimated variance components for the main contributors to the measurement error (Briesch, et. al, 2014).

From the previous sections, it can be concluded that G theory provides a comprehensive conceptual framework and methodology for the simultaneous analysis of more than one facet of measurement in the study of assessment error and outcome reliability (Brennan, 2001). In other words, a major contribution of G theory is that it allows a decision-maker to identify the sources of measurement error and change the appropriate number of observations accordingly to obtain a certain level of generalizability (Marcoulides, 1993). Because of its advantages, G theory has been most widely applied in the fields of educational research and measurement theory (Cardinet, et al., 2011).

As discussed in previous paragraphs, G theory provides a natural framework for analyzing multiple sources of variation from complex measurement procedures. Admission interviews are one of these complex measurement procedures (Brennan, 1983). Therefore, G theory constituted the theoretical framework for this study.

1.2 Problem Statement

The teacher training institution where the study was carried out is one of the state higher education institutions that trains classroom and pre-school teachers in the TRNC and offers a four-year training period. There is only one department in the institution (Department of Teacher Education) and it includes classroom teacher education (CTEP) and pre-school teacher education (PTEP) programmes with courses developed in consultation with the Council of Higher Education (YOK) in Turkey.

To study at the teacher training institution is based on a written examination and interviews. The written examination consists of sub-tests in mathematics, Turkish, natural sciences, social sciences and English. In the assessment of the exam, one mistake leads to a deduction of 0.25 points. At the same time, a prospective teacher must give at least five correct answers in each sub-test to pass the exam and be ranked. Prospective teachers who pass the written test are invited to an interview and their attitude and behaviour towards teaching are assessed according to certain criteria. The governing board of the college appoints an interview committee consisting of five persons chaired by the president of the college. The members of the interview committee (raters) are recommended by the governing board to be impartial and experienced educators. The score for each prospective teacher is calculated by averaging all of the raters' scores, which are ranked from one to five on the scoring guide. There are five dimensions in the scoring guide: a) general culture (GC), b) language (L), c) self-image (SI), d) hobbies (H), and e) attitude toward the teaching profession (ATTP). Although there are dimensions in the interviews, spontaneous questions are also asked to the prospective teachers under each dimension according to the flow of the conversation.

A critical assessment of the reliability of these scores is important for future improvement of the interviews. In addition, the admission of prospective teachers is important for academic outcomes and student well-being. It contributes to the social and economic well-being of a nation because a teacher makes an economic contribution to a nation. The purpose of this study was to investigate the reliability of raters' ratings using the same rating guide in two different interviews (Interview 1 and Interview 2) for the admission of prospective teachers to the college. As explained above paragraphs, the teacher training institution conducts interviews for two different departments after the written exam. These two interviews were conducted according to the same evaluation guide and dimensions. The study aimed to determine the reliability of these two interviews. G theory provides a natural framework for analyzing multiple sources of variation from complex measurement procedures (Brennan, 2001). Using G theory, this study assessed the reliability of the interviews in the admission of prospective teachers for the two programmes (CTEP and PTEP) in college and determined the most appropriate tasks for assessment. In other words, G theory was conducted for both interviews and the results were analysed. In addition, G theory was used to determine the most reliable number of raters to use in the measurement.

2. METHODOLOGY

The reliability of the scores obtained from the interviews was determined in this study. All prospective teachers who wanted to study at this teacher training institution had to take the written test. Based on a fixed quota, those who passed the Written Test in the academic year 2022-2023 were invited for an interview, which was scored by five raters on five dimensions (general culture (GC), language (L), self-image (SI), hobbies (H), and attitude toward teaching profession (ATTP)). Prospective teachers who completed the interview are eligible to enrol for the college. In the current study, 28 prospective teachers who passed the written examination participated in Interview 1 to enrol for CTEP, and 31 prospective teachers participated in Interview 2 to enrol for PTEP. Five raters participated in each interview which was conducted at the same time. Different raters scored the prospective teachers on five dimensions by following the same evaluation guide in the interviews. The schematic view of the research is presented in Figure 1.

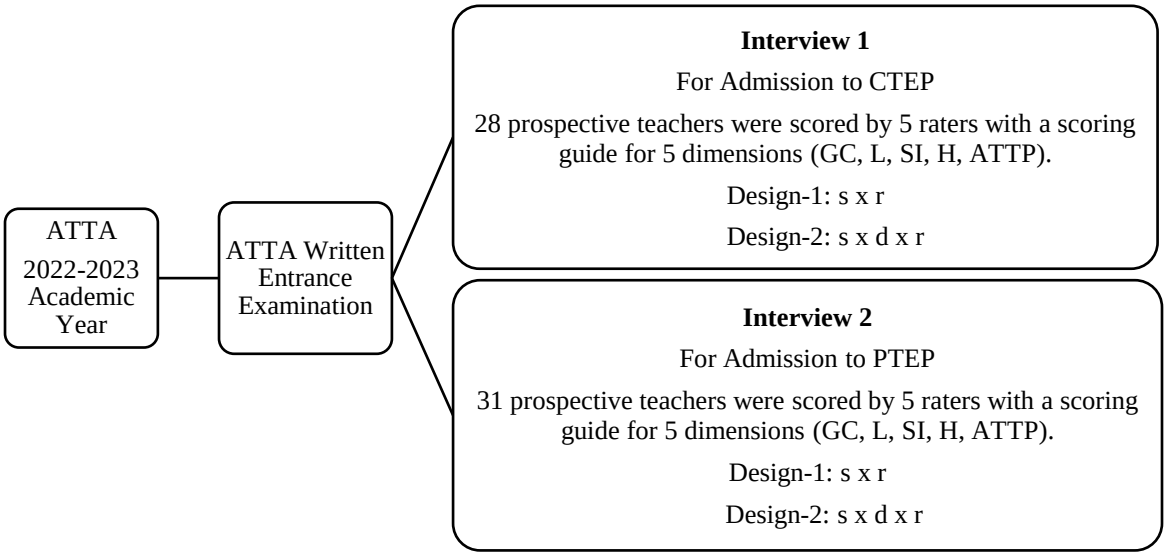

Figure 1. Schematic view of the research

Design-1: s x r

In a single facet pattern, there are three variability sources: (a) the measurement object, (b) the variability source (facet), and (c) the measurement object, the variability source, and residual or unidentified variance. In this design, the measuring object is the source of variability for prospective teachers (s) and raters (r), and there are three sources of variability: s, r and s × r, e. (Brennan, 2001; Shavelson & Webb, 1991).

Table 1. Single-facet crossed s x r design

	Raters (r)				
	1	2	3	4	5
Prospective teacher-1 (s)					
Prospective teacher -2 (s)					
Prospective teacher -3 (s)					
.					

As can be seen from Table 1, the single-facet crossed s x r design, which was arranged as Design-1, was carried out for the scores given by all 5 raters for each component of the prospective teachers in both Interview 1 and Interview 2. This study was conducted separately for the 5 dimensions in both Interview 1 and Interview 2.

In the Venn diagram shown in Figure 2, the three areas marked by the two circles represent the contribution of three "effects" to the variance of the total score: the prospective teacher effect, the rater effect and the prospective teacher-rater interaction effect. The prospective teacher-rater interaction effect is conflated with all unidentified sources of systematic variance plus the variance resulting from random fluctuations (e).

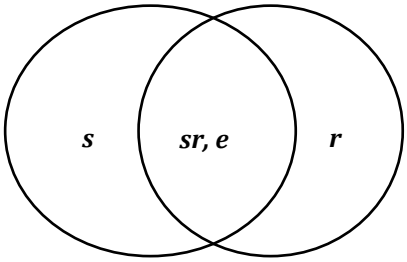

Figure 2. Variance partition diagram for the simplest model sr, where s and r represent prospective teachers and raters, respectively

Design-2: s x d x r

In this design dimensions (d) are crossed with raters (r) in the universe of admissible observations. Estimated variance components for this fully crossed design can be used to estimate results for any possible two-facet design (Brennan, 2001).

Table 2. Two-facet crossed s x d x r design

General Culture (d)					Language (d)					Self-image (d)					Hobbies (d)					Attitude toward Teaching Profession (d)				
Rater (r)					Rater (r)					Rater (r)					Rater (r)					Rater (r)				
1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5

As can be seen from Table 2, the scores of each prospective teacher for each dimension were included in the G study. There are 7 variability sources variance in this design represented in Figure 3.

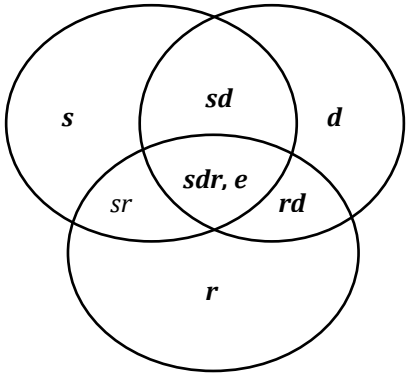

Figure 3. Variance partition diagram for the model s x d x r, where s, d, and r represent prospective teachers, dimensions and raters, respectively

In the Venn diagrams in Figure 3, each main effect is represented by a circle. The interaction effects are represented by the intersections of the circles. The total number of effects (seven) is the number of different areas in the Venn diagram. In each of these designs, other sources of residual error (e) are completely mixed with the effect containing all three indices.

Prospective teachers who passed the Written Entrance Examination and participated in an interview became participants in the study. For those who had opted for the Classroom Teacher Education Programme (CTEP) or the Pre-school Teacher Education Programme (PTEP), two separate interviews were conducted, referred to by the researchers as Interview 1 and Interview 2. A total of 59 prospective teachers and 10 raters were included in the study. The number of prospective teachers and raters in Interview 1 and Interview 2 has been listed below:

Interview 1 : 28 prospective teachers who passed the Written Entrance Examination and 5 raters.

Interview 2: 31 prospective teachers who passed the Written Entrance Examination and 5 raters.

The governing board of the college makes the final determination of the raters among the experienced faculty members.

The sample of the interview questions according to five dimensions is presented in Table 3.

Table 3. Sample of the interview questions

Dimensions	**Sample of Interview Questions**
General Culture (GC)	The prospective teachers were asked questions about - the Turkish Republic of Northern Cyprus (social, cultural, geographical, political and economic issues) - the current news in the media
Language (L)	The prospective teachers were asked to read a text of 200 words.
Self-image (SI)	The prospective teachers were asked to introduce themselves (Where do you live? What do your parents do? Do you have a brother or sister? Which high school did you graduate from? etc.)
Hobbies (H)	The prospective teachers were asked to talk about the activities they participate in, such as sports, culture, and music. Specify the social events they would like to attend in ATTA.
Attitude toward Teaching Profession (ATTP)	The prospective teachers were asked to explain why they wanted to become teachers.

The data for the study was collected through the raters' scores based on the evaluation guide for both interviews. The scoring guide presented in Table 4 was used for the study. The raters for the two interviews rated each prospective teacher on five dimensions. The raters completed the evaluation form based on questions they asked prospective teachers. Each rater rated independently and did not know the other raters' scores. The evaluation form is a 5-point Likert-type rubric for the 5 dimensions. After each prospective teacher was interviewed, the total scores of the five raters were added to determine the total score. The highest and lowest scores for a prospective teacher ranged from 25 to 125. For each dimension, the highest and lowest scores were 25 and 5, respectively. Both interviews were conducted during the same period, so the raters in both interviews were different. Although the raters used the same evaluation guide, the prospective teachers were asked different questions spontaneously, so different questions may have been asked in each interview. In addition, raters were not trained on how to ask questions or how to score prospective teachers' responses. There was a guideline, but the fact that the questions of the dimensions were not analytically defined allowed for a rough assessment.

Table 4. The scoring guide used by raters in the interviews

Dimensions / Prospective teachers	General Culture					Language					Self-image					Hobbies					Attitude toward Teaching Profession				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Prospective teachers-1																									
Prospective teachers-2																									
Prospective teachers-3																									
.																									

Significance tests were conducted to determine if there was a difference between the prospective teachers' scores in Interview 1 and Interview 2. After the normality test, it was found that the prospective teachers' scores in terms of all bouts and sum did not have a normal distribution. For this reason, the non-parametric test, namely Mann-Whitney U, was used to compare the results.

The data were analysed within the framework of G theory. EduG was used to analyse the data. EduG is a generalizability software package which is conceived specifically to exploit the symmetry property of G theory. It offers flexibility in the choice of the object of study and identification of instrumentation facets (Cardinet et al., 2011).

The analysis was conducted with variance components based on the G theory (Brennan, 2001; Güler, Uyanık, & Teker, 2012). Two different designs were used for the study. The reliability of the scores given by the 5 raters for each dimension of the prospective teachers was calculated separately. For this design, a factorial crossed design of prospective teachers (s) x raters (p) was used (Table 5).

Table 5. Average squares formula estimated for single-facet crossed s x r design

Source of variability	Variance component	Estimated Average of Squares
Prospective teacher (s)	σ_s^2	$\sigma_{sr,e}^2 + n_r + \sigma_s^2$
Rater (r)	σ_r^2	$\sigma_{sr,e}^2 + n_s + \sigma_r^2$
s x r	$\sigma_{sr,e}^2$	$\sigma_{sr,e}^2$

The reliability of the scores given by the raters to the prospective teachers in five dimensions was used in a two-facet crossed s x d x r design: prospective teacher (s) x dimension (d) x rater (p) (Table 6).

Table 6. Formulas of mean squares estimated for two-facet crossed s x d x r design

Source of variability	Variance component	Estimated Average of Squares
Prospective teachers (s)	σ_s^2	$\sigma_{sdr,e}^2 + n_r\sigma_{sd}^2 + n_d\sigma_{sr}^2 + n_d n_r\sigma_s^2$
Dimension (d)	σ_d^2	$\sigma_{sdr,e}^2 + n_s\sigma_{dr}^2 + n_r\sigma_{sd}^2 + n_s n_r\sigma_d^2$
Rater (r)	σ_r^2	$\sigma_{sdr,e}^2 + n_s\sigma_{dr}^2 + n_d\sigma_{sr}^2 + n_s n_d\sigma_r^2$
s x d	σ_{sd}^2	$\sigma_{sdr,e}^2 + n_r\sigma_{sd}^2$
s x r	σ_{sr}^2	$\sigma_{sdr,e}^2 + n_d\sigma_{sp}^2$
d x r	σ_{dr}^2	$\sigma_{sdr,e}^2 + n_s\sigma_{dp}^2$
s x d x r	$\sigma_{sdr,e}^2$	$\sigma_{sdr,e}^2$

3. FINDINGS

Table 7 summarises the significant test results of the scores of the prospective teachers in Interview 1 and Interview 2 for five dimensions, GC (p = .638), L (p = .832), SI (p = .191), H (p =.183), ATTP (p = .371) and total (p = .527). There was no significant difference between Interview 1 and Interview 2 raters' scores.

Table 7. Significance tests of the scores of the prospective teachers in Interview 1 and Interview 2 for five dimensions

Dimension	Interview 1		Interview 2		df	Mean Difference	U	P
	$\bar{X}$	SD	$\bar{X}$	SD				
GC	20.5	2.39	20.4	1.90	44	0.1	239	.638
L	21.6	3.40	21.9	2.21	44	0.3	250	.832
SI	22.3	2.53	23.2	1.85	44	0.9	210	.191
H	21.9	1.76	21.1	2.14	44	0.8	201	.183
ATTP	22.7	1.76	22.2	2.14	44	0.5	220	.371
Total	109.1	11.11	108.8	7.82	44	0.3	231	.527

Table 8 presents the variance variables and G-values of the dimensions of Interview 1 and Interview 2. As explained in the introduction part of the study, the G-coefficient of measurement indicates how well a measurement procedure differentiated between the objects of study, i.e. how well the procedure ranked the objects on a measurement scale (Cardinet, et al., 2010). The G coefficients were 0 for the GC dimension, 0.69 for the L dimension, 0.06 for the SI dimension, 0.50 for the H dimension, and 0.68 for the ATTP dimension in Interview 1 . The G coefficients were not acceptable for the GC and SI dimensions. On the other hand, the G-coefficients for the dimensions L, H and ATTP in Interview 1 were relatively low. The G-coefficients were 0.47 for the dimension GC, 0.46 for the dimension L, 0.56 for the dimension SI, 0.47 for the dimension H and 0.51 for the dimension ATTP in Interview 2. It can be said that the G-coefficients for all dimensions in Interview 2 were relatively low.

Table 8. Variance variables and G values of the dimensions of Interview 1 and Interview 2

		Interview 1					Interview 2				
Dimensions		SD	df	MS	%	G	SD	df	MS	%	G
	s	1.73	27	0.06	0.0		6.21	30	0.20	13.8	
GC	r	22.82	4	5.70	63.6	0.0	2.12	4	0.53	9.6	0.47
	s x r	12.37	108	0.11	36.4		13.07	120	0.10	76.6	
	s	18.14	27	0.67	20.7		11.39	30	0.37	7.2	
L	r	16.81	4	4.20	32.0	0.69	30.86	4	7.71	50.1	0.46
	s x r	22.74	108	0.21	47.3		24.73	120	0.20	42.7	
	s	2.74	27	0.10	0.5		5.54	30	0.18	16.2	
SI	r	18.47	4	4.61	62.5	0.06	1.03	4	0.25	4.7	0.56
	s x r	10.32	108	0.09	37.0		10.96	120	0.09	79.1	
	s	6.40	27	0.23	8.3		6.21	30	0.20	13.8	
H	r	16.38	4	4.09	50.0	0.50	2.12	4	0.53	9.6	0.47
	s x r	12.81	108	0.11	41.7		13.07	120	0.10	76.6	
	s	12.80	27	0.47	11.3		5.54	30	0.18	16.2	
ATTP	r	40.52	4	10.13	62.3	0.68	1.03	4	0.25	4.7	0.51
	s x r	16.60	108	0.15	26.3		10.96	120	0.09	79.1	

Table 8 showed that the main effect of rater (r) explained 63.6% of the total variance for the GC dimension in Interview 1. The estimated variance component for prospective teachers was 0.06 which accounts for 0% of the total variance in this dimension. The estimated variance component for prospective teachers by raters (sxr) was 0.11 which accounts for 36.4% of the total variance component in Interview 1 . On the other hand, the highest contribution to measurement error in this dimension was the s x r (0.10) accounting for 76.6% of the total variance in Interview 2. This showed that a proportion of the variance was due to the interaction of prospective teachers by raters and another systematic or unsystematic source of variance that was not measured in the study. The second largest source of variation to measurement error was due to differences among prospective teachers with a variance component of 0.20 accounting for 13.8% of the total variance in Interview 2. This indicates that the assessment process more or less determines the differences among prospective teachers. Similar interpretations can be made for other dimensions of Interview 1 and Interview 2.

Table 8 also showed that the highest contributions to measurement error were in the GC, SI, H, and ATTP dimensions, i.e., s x r, in interview 2. Thus, it can be concluded that some of the variance was due to the interaction of the prospective teachers by the raters and other systematic/unsystematic sources of variance that were not measured in the study for Interview 2. On the other hand, the main effect of rater (r) explained most of the total variance for all dimensions except L in Interview 1. This result can be interpreted to mean that the raters unanimously revealed the differences in four dimensions (GC, SI, H, and ATTP) for prospective teachers.

Table 9 shows the variance components identified by the G-study for the full factorial design for prospective teachers, 5 dimensions and 5 raters. The symbol "s" stands for the prospective teachers, the symbol "d" stands for the dimension and the symbol "r" stands for the sources of variability of the raters in the table.

Table 9. Variance components and the percentages of explanation of the total variance estimated as a result of the s x d x r design

	Interview 1				Interview 2			
	SD	df	MS	%	SD	df	MS	%
s	11.14	27	0.41	1.2	28.14	30	0.93	10.5
d	17.56	4	4.39	6.5	9.10	4	2.27	3.1
r	105.30	4	26.32	46.0	21.29	4	5.32	11.3
s x d	30.67	108	0.28	7.4	13.21	120	0.11	2.9
s x r	15.73	108	0.14	0.5	38.22	120	0.31	21.5
d x r	9.72	16	0.60	4.2	18.57	16	1.16	15.6
s x d x r	58.84	432	0.13	34.1	37.50	480	0.07	35.0

Interview 1 G= 0.29, Interview 2 G= 0.63

Table 9 also showed that raters account for the largest proportion (46% and 11.3%) of variance among the main effects in Interview 1. This result was evident in the measurements as the differences between raters were revealed. The joint effect of s x d explained 7.4% of the total variance. It can be interpreted that the interaction between prospective teachers and dimension (s x d) was an indicator of change in prospective teachers' performance on each dimension. Since the interaction effect s x r explained 0.5% of the total variance, it was assumed that raters did not rate prospective teachers differently. In other words, it can be said that the raters' ratings did not differ among the prospective teachers. The interaction effect d x r explained 4.2% of the total variance, i.e., it can be said that the raters did not rate the dimensions differently and gave similar scores to the prospective teachers. The s x d x r together explain 34.1% of the total variance. These results indicated that prospective teacher (s) x dimension (d) x rater (r) effect and/or random errors can be large. It can be said that the Interview 1 may contain different variance sources including the prospective teachers, dimensions, raters and random errors that were not measured in this study. This value of variance is desired to be as small as possible (Güler et al., 2012).

Table 9 showed that the variance components of the prospective teachers in interview 2 were low (10.5%), the differences between the prospective teachers cannot be sufficiently revealed and the observed values are not sufficient to represent the actual values. The interaction effect s x r explains 21.5% of the total variance. This result can be interpreted to mean that the raters rate the prospective teachers differently, in other words, the raters' ratings differ from one prospective teacher to another. Moreover, it can be said that some raters rate some prospective teachers more strictly or more generously than other raters. The joint effect of s x d explains 2.9% of the total variance. This result shows us that the performance of prospective teachers varies slightly from dimension to dimension. Since the interaction effect d x r explains 15.6% of the total variance, it can be assumed that the raters did not rate the dimensions differently. The largest variance value was determined for the interaction effect sxdxr (35%), as shown in Table 9. So, like Interview 1 , Interview 2 can also contain different sources of variance. This means that Interview 2 may contain random errors that were not measured in the study along with the prospective teachers, dimensions and raters.

The reliability coefficients of the five dimensions were 0.29 for Interview 1 and 0.63 for Interview 2. According to Crocker and Algina (1986), reliability coefficients vary between 0 and 1, and coefficients of 0.70 and above are acceptable. It can be concluded that both reliability coefficients obtained with the s x d x r design were unacceptable for both Interview 1 and Interview 2.

4. DISCUSSION

The purpose of this study was to investigate the reliability of the raters' scores using the same evaluation guide in two different interviews (Interview 1 and Interview 2) for the admission of prospective teachers to a teacher training institution. For this purpose, G theory was used and two designs were formed, namely, sxr and s x d x r. These two designs were discussed in terms of variance values and the reliability of the test for main and joint effects. Due to the G theory's power and usefulness, it was preferred to CTT in examining measurement error and reliability for two interviews.

At the beginning of the data analysis, it was investigated whether there was a significant difference between the two interview results. The result showed that there was no significant difference between the interview results. These results can be interpreted to mean that the raters gave similar scores to the prospective teachers in both Interview 1 and Interview 2.

In the s x r design created for the Interview 1 scores, the highest G coefficient was found in the L dimension (0.69), while the lowest G coefficient was in the GC dimension (0.0). In Interview 2, the G coefficient of the score in the SI dimension was the highest (0.56), whereas the G coefficient in the L dimension was the lowest. Furthermore, for Interview 2, the highest variance component was below the interaction effect s x r in all dimensions except L. This can be interpreted that the random errors might be large in these dimensions. As mentioned in the introduction, structured interviews increase reliability and reduce subjectivity (Pursell et al., 1980). On the other hand, unstructured interviews are criticised for their low reliability (Petrarca & LeSage, 2014; Smith & Pratt, 1996). The results of this study could be influenced by spontaneous conversations during the interviews and the raters could influence each other. Therefore, it can be said that the interview environment influences the questions asked by the raters and their ratings. It is strongly recommended that the interviews be conducted structurally, i.e. the interview questions should be prepared in advance and asked the prospective teachers in the same order. In addition, the evaluation guide used by the raters could be changed into an analytical evaluation guide. This helps to increase the consistency of the assessments. Raters could be trained in advance on how to use the analytical assessment guide.

For both interviews, the estimated variance component "prospective teachers" had a small effect (%1.2 for Interview 1 and %10.5 for Interview 2) explaining the total variance. According to the literature, the measurement item (prospective teachers) should have a significant effect in explaining the total variance (Taşdelen-Teker et al., 2016). Therefore, the prospective teachers did not differ significantly in their performance on the individual dimensions. This result of the study contradicts some studies in the literature (e.g. Gürlen, Boztunç-Öztürk, & Eminoğlu, 2019).

The results showed that the estimated variance component for the main effect "dimension" explained 6.5% for Interview 1 and 3.1% for Interview 2 of the total variance. The main effect of dimension had the smallest contribution to the total variance in Interview 2, suggesting that the dimension effect did not have a strong influence on the total variance. In short, the scoring of the prospective teachers did not too much differ according to the dimension.

It was found that the variance component of the main effect "rater" was higher in the sxdxr design in Interview 1 . The proportion of the definition of the estimated variance for prospective teachers and the rater main effect in the total variance showed that the group of prospective students was not homogeneous in terms of their performance in Interview 1 . Therefore, there was an effect due to the difference between raters. Some studies in the literature (e.g., Yılmaz & Tavsancıl, 2014; Yılmaz & Başbaşa, 2015) reported that the main effect of rater variability in explaining the overall variance was relatively small and concluded that raters were consistent in rating students. In this regard, the results of this study were consistent with the literature regarding the results of Interview 2, while they contradicted those of Interview 1. In addition, the results of the study showed that the percentage of the residual component in the variance for Interview 1 was 34.1%. Thus, it can be said that 34.1% of the total variance was due to unexplained systematic or unsystematic errors. On the other hand, the results of the analyses conducted for Interview 2 showed that the variance proportion of the residual component with the largest variance was 35%. For Interview 2, 35% of the total variance was due to unexplained systematic or unsystematic errors.

The results of the study show that the estimated variance component for the joint effect of prospective teacher-rater (sxr) explains 0.5% of the total variance in Interview 1 and %21.5 in Interview 2. While the variance component for the joint effect of prospective teacher-rater (sxr) was highest in Interview 2, it was lowest in Interview 1 . Therefore, prospective teachers' ratings by different raters differed in Interview 2. The prospective teacher dimension explained 7.4% of the total variance of the estimated variance component for the joint effect in Interview 1 and 2.9% of the total variance of the estimated variance for the joint effect in Interview 2. These results showed that prospective teachers did not differ by dimensions. Looking at the results of the estimated variance for the dimension rater (dxr) joint effect, it explains 4.2% of the total variance in Interview 1 and %15.6 for Interview 2. According to these results, it can be said that the scores obtained by the rater differed slightly according to the dimensions. In many of the research studies, expressions like criterion, task and item were used instead of dimensions which are consistent with the purpose of the study. The results of some of these studies are consistent with the present study, which showed that the scores awarded by raters according to the criteria/tasks/items differed slightly (Gürlen et al., 2019). On the other hand, some of the studies (i.e. Uzun, Aktaş, Aşiret & Yorulmaz, 2018; Yılmaz & Gelbal, 2011) found that the variance of the task component was too large, which contradicts the results of this study.

The unexplained variance was too large in the sxdxr designs for both Interview 1 and Interview 2 (%34.1 and %35, respectively). This result of the study showed that the different sources of variability

(e.g. gender and reading skills of the prospective teachers) should be taken into account in the designs. G theory analysis revealed a low G-coefficient and reliability index for Interview 1 and Interview 2 (0.29 and 0.63 respectively). The reason for this could be the unexplained sources of variance.

Several suggestions are made for future research regarding the findings and limitations of this study:

a) This study can be replicated with other raters and sources of variance. In other words, G theory can be conducted to evaluate other sources of error and their interactions and to conduct a comprehensive reliability analysis.
b) Comparative studies of the interview scores of prospective teachers using G theory can be carried out to gain a clear understanding of the reliability of the interviews.
c) Finally, it should be emphasised that there are some important theoretical and statistical issues that clearly need to be addressed in more detail in G Gtheory. As a result, there are potential areas of application in which G theory has not yet been fully exploited to date.

REFERENCES

Andrew, M. D., Cobb, C. D., & Giampietro, P. J. (2005). Verbal ability and teacher effectiveness. Journal of teacher education, 56(4), 343-354. https://doi.org/10.1177/0022487105279928

Arnon, S. and Reichel, N. (2007). Who is the ideal teacher? Am I? Similarity and differences in perception of students of education regarding the qualities of a good teacher and their qualities as teachers. *Teachers and Teaching: Theory and Practice, 13* (5), 441-446. https://doi.org/10.1080/13540600701561653

Atılgan, H. (2008). Using generalizability theory to assess the score reliability of the special ability selection examinations for music education programmes in higher education. *International Journal of Research & Method in Education, 31*(1), 63-76. https://doi.org/10.1080/17437270801919925

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,* 28-40.

Bardach, L. & Klassen, R. M. (2020). Smart teachers, successful students? A systematic review of the literature on teachers' cognitive abilities and teacher effectiveness. *Educational Research Review, 30,* 100312. Advance Online Publication. https://doi.org/10.1016/j.edurev.2020.100312

Brennan, R. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing.

Brennan, R. L. (2001) *Generalizability theory*. New York: Springer-Verlag.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52*(1), 13-35. https://doi.org/10.1016/j.jsp.2013.11.008

Byrnes, D., Kiger, G., & Shechtman, Z. (2003). Evaluating the use of group interviews to select students for teacher-education programs. *Journal of Teacher Education, 54*(2), 163-172. https://doi.org/10.1177/0022487102250310

Cardinet, J., Johnson, S. & Pini, G. (2011). *Applying generalizability theory using EduG*, New York: Taylor and Francis.

Casey, C., & Childs, R. (2007). Teacher education program admission criteria and what beginning teachers need to know to be successful teachers. *Canadian Journal of Educational Administration and Policy*, (67).

Caskey, M. M., Peterson, K. D., & Temple, J. B. (2001). Complex admission selection procedures for a graduate preservice teacher education program. *Teacher Education Quarterly, 28*(4), 7-21.

Corcoran, R. P., & O'Flaherty, J. (2018). Factors that predict pre-service teachers' teaching performance. *Journal of Education for Teaching, 44*(2), 175-193. https://doi.org/10.1080/02607476.2018.1433463

Crocker, L. M., & Algina, L. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Cronbach, L.J., Rajarathnam, N. & Gleser, G.C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163.

Donnon, T., & Paolucci, E. O. (2008). A generalizability study of the medical judgment vignettes interview to assess students' noncognitive attributes for medical school. *BMC Medical Education, 8*(1), 1-7. https://doi.org/10.1186/1472-6920-8-58

Ebmeier, H., & Ng, J. (2005). Development and field test of an employment selection instrument for teachers in urban school districts. *Journal of Personnel Evaluation in Education, 18*(3), 201e218. https://doi.org/10.1007/s11092-006-9021-4

Faulk, L. G. (2008). Predicting on-the-job teacher success based on a group assessment procedure used for admission to teacher education (Doctoral dissertation). Available from ProQuest Dissertations and Thesis database (UMI No. 3297518).

Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology*, London: Sage.

Goho, J. & Blackman, A. (2006). The effectiveness of academic admission interviews: an exploratory meta-analysis. *Medical Teacher, 28*(4), 335-340. https://doi.org/10.1080/01421590600603418

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

Gürlen, E., Boztunç-Öztürk, N., & Eminoğlu, E. (2019). Investigation of the reliability of teachers, self and peer assessments at the primary school level with generalizability theory. *Journal of Measurement and Evaluation in Education and Psychology, 10(*4), 406-421. https://doi.org/10.21031/epod.583891

Haberman, M. & Post, L. (1998).Teachers for multicultural schools: the power of selection. *Theory into Practice, 37*(2), 96-104. https://doi.org/10.1080/00405849809543792

Huffcutt, A., & Woehr, D. (1999). Further analysis of employment interview validity: A quantitative evaluation of interviewer-related structuring methods. *Journal of Organizational Behavior, 20*, 549–560.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Kim, L. E., Jörg, V., & Klassen, R. M. (2019). A meta-analysis of the effects of teacher personality on teacher effectiveness and burnout. *Educational psychology review, 31,* 163-195. https://doi.org/10.1007/s10648-018-9458-2

Leger, K. E. (2014). Defining teaching excellence: A phenomenological study of 2013 highly effective louisiana value-added model teachers with perfect evaluation scores. (Unpublished doctoral dissertation). Faculty of the College of Graduate Studies, Lamar University.

Lupascu, A. R., Pânisoarâ, G., & Pânisoarâ, I-O. (2014). Characteristics of effective teacher. *Procedia. Social and Behavioral Sciences, 127,* 534-538. https://doi.org/10.1016/j.sbspro.2014.03.305

Marcoulides, G. A. (1993). Maximizing Power in Generalizability Studies under Budget Constraints, *Journal of Educational Statistics, 18*(2), 197-206.

Mathis, R. L., Jackson, J. H., Valentine, S. R., & Meglich, P. (2016). *Human resource management*. Cengage Learning.

McDaniel, M. A., Whetzel, D. J., Schmidt, F. T., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616. https://doi.org/10.1037/0021-9010.79.4.599

Petrarca, D., & LeSage, A. (2014). Should it stay or should it go? Re-considering the Pre-service Teacher Education Admissions Interview. ICET 2014, 252.

Pellicer, L. O. (1981). Improved teacher selection with the structured interview. *Educational Leadership, 38*(6), 492-94.

Pursell, E. D., Campion, M. A., & Gaylord, S. R. (1980). Structured interviewing: Avoiding selection problems. *Personnel Journal, 59,* 907–912.

Schulte, D. P., Slate, J. R., & Onwuegbuzie, A. J. (2008). Effective high school teachers: A mixed investigation. *International Journal of Educational Research, 47*(1), 351-361. https://10.1016/j.ijer.2008.12.001

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

Shechtman, Z. (1992). Interrater reliability of a single group assessment procedure administered in several educational settings. *Journal of Personnel Evaluation in Education, 6(*1), 31-39.

Smith, H. A., & Pratt, D. (1996). The use of biodata in admissions to teacher education. *Journal of Teacher Education, 47*(1), 43-52. https://doi.org/10.1177/0022487196047001008

Stronge, J. H. (2007). *Qualities of effective teachers*. Alexandria, VA: Association for Supervision and Curriculum Development.

Taneri, P. O. (2017). The viewpoints of instructors about the effects of teacher education programs on prospective teachers' affective characteristics. *Eurasian Journal of Educational Research, 17* (70), 105-120. http://dx.doi.org/10.14689/ejer.2017.70.6

Taşdelen Teker, G., Şahin, M. G., & Baytemir, K. (2016). Using generalizability theory to investigate the reliability of peer assessment applications. *Journal of Human Sciences, 13*(3), 5574-5586. http://dx.doi.org/10.14687/jhs.v13i3.4155

Thomson, D., Cummings, E., Ferguson., A. K., Miyuki Moizumi, E., Sher, Y., Wang, X., & Childs, R. A. (2011). A role for research in initial teacher education admissions: A case study from

one Canadian university. *Canadian Journal of Educational Administration and Policy, 121*, 1–23.

Tran, T., & Blackman, M. C. (2006). The dynamics and validity of the group selection interview. *The Journal of Social Psychology, 146*, 183–201. https://10.3200/SOCP.146.2.183-201

Webb, N. M., & Shavelson, R. J. (2005). *Generalizability theory: overview*. Wiley StatsRef: Statistics Reference Online.

Wing Tat-Chiu, C. (2001). Scoring performance assessments based on judgements generalizability theory. Springer Science+Business Media, LLC

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

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

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

Uzun, N. B., Aktaş, M., Aşiret, S., & Yorulmaz, S. (2018). Using generalizability theory to assess the score reliability of communication skills of dentistry students. *Asian Journal of Education and Training, 4*(2), 85–90. https://doi.org/10.20448/journal.522.2018.42.85.90

GENİŞLETİLMİŞ TÜRKÇE ÖZET

KUZEY KIBRIS'TA ÖĞRETMEN EĞİTİMİ PROGRAMLARINA KABULDE KULLANILAN MÜLAKATLARIN GÜVENİLİRLİĞİ: GENELLENEBİLİRLİK KURAMI

GİRİŞ

Literatürdeki bazı çalışmalar ideal/etkili öğretmenlerin en önemli özelliklerinden birinin kişilik özellikleri olduğunu göstermektedir (örn. Lupascu, Pânisoarâ ve Pânisoarâ, 2014; Leger, 2014). Öte yandan, birçok öğretmen eğitimi programı, öğretmen adaylarını yalnızca bilişsel testler veya sınavlar/standartlaştırılmış test puanları gibi akademik kriterlere göre kabul etmektedir (Haberman ve Post, 1998). Araştırmalar, öğretmen adaylarının yalnızca akademik kriterlere göre değerlendirilmesinin öğretmen etkililiğini yordamadığını (Bardach ve Klassen, 2020) ya da zayıf bir şekilde yordadığını göstermektedir (Corcoran ve O'Flaherty, 2018).

Öğretmen eğitimi programlarına hangi adayların kabul edileceğini belirlemek için farklı ülkelerde çeşitli değişkenler kullanılmaktadır. En yaygın değişkenler arasında not ortalaması, yazılı sınavlar, referans mektupları ve iş deneyimi yer almaktadır. İngiltere gibi bazı ülkelerde ise öğretmen adaylarının öğretmen yetiştiren kurumlara kabulünde mülakatlar kullanılmaktadır.

Bireysel mülakatlar, bilgi, kişisel nitelikler ve sözlü iletişim gibi nitelik boyutlarına odaklanır. Bireysel mülakatlar a) *yapılandırılmamış* (asgari yönergeler), b) *orta düzeyde yapılandırılmış* (panel mülakatleri, önceden belirlenmiş puanlama ve anketler), c) *yüksek düzeyde yapılandırılmış* (önceden tanımlanmış sorular ve örnek cevaplar, panel mülakatleri, mülakatcilerin eğitimi ve sürecin sürekli değerlendirilmesi) şeklinde olabilir (Goho ve Blackman, 2006). Bu teknikler arasından, Pursell, Campion ve Gaylord (1980) güvenilirliği artırması ve öznelliği azaltması bakımından yapılandırılmış mülakatların kullanılmasını tavsiye etmektedir. Ayrıca, yapılandırılmış bir mülakat formatının mülakatci yanlılığına karşı bir çare olarak işlev gördüğü öne sürülmüştür (Ebmeier ve Ng, 2005). Yapılandırılmamış mülakatlar düşük güvenilirlikleri (Petrarca & LeSage, 2014; Smith & Pratt, 1996) ve iş performansı/başarısının geçersiz yordayıcıları olmaları nedeniyle eleştirilmektedir (Mathis & Jackson, 2011).

Öğretmen adaylarının bir öğretmen eğitimi programına kabul edilmesindeki en büyük sorunlardan biri mülakatların güvenilirliğidir (Casey & Childs, 2007). Çoğu öğretmen eğitimi programı için geçerlilik ve güvenilirlik kanıtı bulunmadığından, kabul kararlarında mülakatların kullanımı tartışmalıdır. Araştırmacılar, güvenilirlik sorunlarının *puanlayıcı eğitimi* (Donnon & Paolucci, 2008; Jonsson & Svingby, 2007) ve *başvuru başına puanlayıcı sayısının artırılmasıyla* (Byrnes et al., 2000; Caskey et al., 2001;Smith & Pratt, 1996) güçlendirilebileceğini öne sürmüştür. Mülakatların güvenilirliğini puanlayıcıların puanları üzerinden değerlendirmek için bazı yöntem ve teoriler

bulunmaktadır (örn. Fleiss' Kappa, Puanlayıcılar Arası Güvenilirlik, Rasch Modeli). Genellenebilirlik Kuramı (G Kuramı) da bu amaçla kullanılmaktadır (Wing & Chiu, 2001).

G Kuramı, davranışsal ölçümlerin güvenilirliğini değerlendirmek için kullanılan istatistiksel bir kuramdır (Webb & Shavelson, 2018). G Kuramı, ölçümlerin birden fazla hata kaynağına tabi olduğu durumlarda ölçüm doğruluğunu tahmin etmeye yönelik bir yaklaşımdır (Cardinet, Johnson ve Pini, 2011). Sadece halihazırda yapılmış olan ölçümlerin güvenilirliğini tahmin etmek için bir araç sağlamakla kalmayan, aynı zamanda hata katkıları hakkındaki bilgilerin gelecekteki uygulamalarda ölçüm prosedürlerini iyileştirmek için kullanılmasına izin veren bir yaklaşımdır.

Bu çalışmada araştırmanın gerçekleştirildiği kurum, KKTC'de sınıf ve okul öncesi öğretmeni yetiştiren 4 yıllık bir yükseköğretim kurumudur. Kurumda tek bir bölüm (Öğretmen Eğitimi Bölümü) bulunmakta ve Türkiye'deki Yükseköğretim Kurulu (YÖK) ile istişare halinde geliştirilen derslerle Sınıf Öğretmenliği (SÖEP) ve Okul Öncesi Öğretmenliği (OÖEP) programlarını içermektedir. Kurumda eğitim almak için yazılı sınav ve mülakatlar yapılmaktadır. Yazılı sınav matematik, Türkçe, fen bilimleri, sosyal bilimler ve İngilizce alt testlerinden oluşmaktadır. Sınavın değerlendirilmesinde bir hata 0.25 puanlık bir kesintiye yol açmaktadır. Yazılı sınavı geçen öğretmen adayları mülakata davet edilerek öğretmenliğe yönelik tutum ve davranışları belirli kriterlere göre değerlendirilmektedir. Yönetim kurulu, okul müdürünün başkanlığında beş kişiden oluşan bir mülakat komitesi atamaktadır. Mülakat komitesinin üyeleri (değerlendiriciler) yönetim kurulu tarafından tarafsız ve deneyimli eğitimcilerden seçilir. Her bir öğretmen adayının puanı, puanlama rehberinde birden beşe kadar sıralanan tüm puanlayıcıların puanlarının ortalaması alınarak hesaplanır. Puanlama kılavuzunda beş boyut bulunmaktadır: a) genel kültür (GK), b) dil (D), c) öz imaj (Öİ), d) hobiler (H) ve e) öğretmenlik mesleğine yönelik tutum (ÖMT). Mülakatlarda boyutlar bulunmakla birlikte, konuşmanın akışına göre her bir boyut altında aday öğretmenlere spontane sorular da sorulmaktadır.

Problem Durumu

Bu çalışmanın amacı, aday öğretmenlerin üniversiteye kabulü için yapılan iki farklı mülakatta (Mülakat 1 ve Mülakat 2) aynı derecelendirme kılavuzunu kullanan puanlayıcıların puanlamalarının güvenilirliğini araştırmaktır. Yukarıdaki paragraflarda açıklandığı gibi, öğretmen yetiştiren okul yazılı sınavdan sonra iki farklı bölüm için mülakat yapmaktadır. Bu iki mülakat aynı değerlendirme rehberi ve boyutlarına göre gerçekleştirilmektedir. Bu çalışmada, G Kuramını kullanılarak, iki program (CTEP ve PTEP) için öğretmen adaylarının kabulünde mülakatların güvenilirliğini değerlendirmiş ve sonuçlar analiz edilmiştir. Ayrıca, ölçümde kullanılacak en güvenilir puanlayıcı sayısını belirlemek için de G Kuramı kullanılmıştır.

YÖNTEM

Bu çalışmanın katılımcılarını, Yazılı Giriş Sınavını geçen ve mülakata katılmaya hak kazanan öğretmen adayları oluşturmaktadır. Sınıf Öğretmenliği Eğitim Programı (SÖEP) veya Okul Öncesi Öğretmenliği Eğitim Programı'nı (OÖEP) tercih edenler öğretmen adayları için Mülakat 1 ve Mülakat 2 olarak

adlandırılan iki ayrı mülakat gerçekleştirilmiştir. Çalışmaya toplam 59 öğretmen adayı ve 10 puanlayıcı dahil edilmiştir.

Çalışmanın verileri, her iki mülakat için değerlendirme rehberine dayalı olarak puanlayıcıların puanları aracılığıyla toplanmıştır. İki görüşme için puanlayıcılar her bir öğretmen adayını beş boyut üzerinden puanlamıştır. Puanlayıcılar değerlendirme formunu öğretmen adaylarına sordukları sorulara dayanarak doldurmuşlardır. Her bir puanlayıcı bağımsız olarak puanlama yapmış ve diğer puanlayıcıların puanlarını bilmemiştir. Değerlendirme formu, 5 boyut için 5'li Likert tipi bir dereceli puanlama anahtarıdır. Her bir aday öğretmenle görüşüldükten sonra, beş puanlayıcının toplam puanları toplanarak toplam puan belirlenmiştir. Bir aday öğretmen için en yüksek ve en düşük puanlar 25 ile 125 arasında değişmektedir. Her bir boyut için en yüksek ve en düşük puanlar sırasıyla 25 ve 5'tir. Her iki mülakat de aynı dönemde gerçekleştirilmiştir, dolayısıyla her iki mülakattaki puanlayıcılar farklıdır. Puanlayıcılar aynı değerlendirme rehberini kullanmış olsalar da aday öğretmenlere spontane olarak farklı sorular sorulmuştur, bu nedenle her mülakatde farklı sorular sorulmuş olabilir. Buna ek olarak, puanlayıcılar nasıl soru soracakları ya da aday öğretmenlerin yanıtlarını nasıl puanlayacakları konusunda eğitilmemiştir.

Öğretmen adaylarının Mülakat 1 ve Mülakat 2'den aldıkları puanlar arasında fark olup olmadığını belirlemek için anlamlılık testleri yapılmıştır. Normallik testinin ardından, öğretmen adaylarının alt ve toplam puanlarının normal dağılıma sahip olmadığı görülmüştür. Bu nedenle sonuçları karşılaştırmak için non-parametrik test olan Mann-Whitney U kullanılmıştır. Veriler G kuramı çerçevesinde analiz edilmiş ve EduG programı kullanılmıştır.

BULGULAR

Öğretmen adaylarının Mülakat 1 ve Mülakat 2'de beş boyut için aldıkları puanların anlamlı test sonuçları şu şekildedir: GK (p =.638), D (p =.832), Öİ (p =.191), H (p =.183), ÖMT (p = .371) ve toplam (p =.527). Mülakat 1 ve Mülakat 2 puanlayıcılarının puanları arasında anlamlı bir fark bulunmamıştır.

Mülakat 1 'de G katsayıları GK boyutu için 0, D boyutu için 0.69, Öİ boyutu için 0.06, H boyutu için 0.50 ve ÖMT boyutu için 0.68'dir. G katsayıları GK ve Öİ boyutları için kabul edilebilir değildir. Öte yandan, Mülakat 1 'de D, H ve ÖMT boyutları için G katsayıları nispeten düşüktür. Mülakat 2 'de G katsayıları GK boyutu için 0.47, D boyutu için 0.46, Öİ boyutu için 0.56, H boyutu için 0.47 ve ÖMT boyutu için 0.51'dir. Mülakat 2'deki tüm boyutlar için G katsayılarının nispeten düşük olduğu söylenebilir.

Puanlayıcı ana etkisi (r) Mülakat 1'deki Kİ boyutuna ilişkin toplam varyansın %63.6'sını açıklamaktadır. Öğretmen adayları için tahmin edilen varyans bileşeni 0.06'dır ve bu boyuttaki toplam varyansın %0'ına karşılık gelmektedir. Aday öğretmenler için puanlayıcılar tarafından tahmin edilen varyans bileşeni (sxr) 0.11 olup Mülakat 1 'deki toplam varyans bileşeninin %36.4'üne karşılık gelmektedir. Öte yandan, bu boyutta ölçme hatasına en yüksek katkı, Mülakat 2 'deki toplam varyansın %76.6'sını oluşturan s x r (0.10) olmuştur. Bu durum, varyansın bir kısmının aday öğretmenlerin puanlayıcılarla etkileşiminden ve çalışmada ölçülmeyen başka bir sistematik ya da sistematik olmayan

varyans kaynağından kaynaklandığını göstermektedir. Ölçüm hatasına yol açan ikinci en büyük varyasyon kaynağı, Mülakat 2 'deki toplam varyansın %13.8'ini oluşturan 0.20'lik varyans bileşeniyle aday öğretmenler arasındaki farklılıklardan kaynaklanmaktadır. Bu durum, değerlendirme sürecinin öğretmen adayları arasındaki farklılıkları az ya da çok belirlediğini göstermektedir. Mülakat 1 ve Mülakat 2 'nin diğer boyutları için de benzer yorumlar yapılabilir.

Mülakat 2 'de ölçme hatasına en yüksek katkının GK, Öİ, H ve ÖMT boyutlarında olduğu görülmektedir. Varyansın bir kısmının aday öğretmenlerin puanlayıcılarla etkileşiminden ve Mülakat 2 için çalışmada ölçülmeyen diğer sistematik/sistematik olmayan varyans kaynaklarından kaynaklandığı sonucuna varılabilir. Öte yandan, puanlayıcı ana etkisi (r) Mülakat 1'de D hariç tüm boyutlar için toplam varyansın çoğunu açıklamıştır. Bu sonuç, puanlayıcıların aday öğretmenler için dört boyuttaki (GK, Öİ, H ve ÖMT) farklılıkları ortaya çıkardığı şeklinde yorumlanabilir.

Beş boyutun güvenilirlik katsayıları Mülakat 1 için 0.29 ve Mülakat 2 için 0.63'tür. Crocker ve Algina'ya (1986) göre güvenilirlik katsayıları 0.70 ve üzeri katsayılar kabul edilebilirdir.

TARTIŞMA, SONUÇ ve ÖNERİLER

Bu çalışmanın sonuçları mülakatlar sırasındaki spontane konuşmalardan puanlayıcılar birbirlerini etkileyebildiğini göstermektedir. Giriş bölümünde de belirtildiği üzere, yapılandırılmış mülakatlar güvenilirliği artırmakta ve öznelliği azaltmaktadır (Pursell vd., 1980). Öte yandan, yapılandırılmamış mülakatlar düşük güvenilirlikleri nedeniyle eleştirilmektedir (Petrarca & LeSage, 2014; Smith & Pratt, 1996). Dolayısıyla, mülakat ortamının puanlayıcıların sordukları soruları ve yaptıkları puanlamaları etkilediği söylenebilir. Mülakatların yapısal olarak gerçekleştirilmesi, yani mülakat sorularının önceden hazırlanması ve aday öğretmenlere aynı sırayla sorulması tavsiye edilmektedir. Buna ek olarak, analitik değerlendirme rehberinin nasıl kullanılacağı konusunda önceden eğitilebilir.

Her iki mülakat için de "öğretmen adayları" tahmini varyans bileşeni toplam varyansı açıklamada küçük bir etkiye sahiptir. Literatüre göre, ölçüm maddesinin (öğretmen adayları) toplam varyansı açıklamada önemli bir etkiye sahip olması gerekmektedir (Taşdelen-Teker vd., 2016). Dolayısıyla, öğretmen adaylarının bireysel boyutlardaki performansları anlamlı bir şekilde farklılaşmamıştır. Araştırmanın bu sonucu literatürdeki bazı çalışmalarla çelişmektedir (örn. Gürlen, Boztunç-Öztürk ve Eminoğlu, 2019).

Araştırma sonuçları, farklı değişkenlik kaynaklarının (örneğin, öğretmen adaylarının cinsiyeti) dikkate alınması gerektiğini göstermiştir. G Kuramı analizi, Mülakat 1 ve Mülakat 2 için düşük bir G katsayısı ve güvenilirlik endeksi ortaya koymuştur. Bunun nedeni açıklanamayan varyans kaynakları olabilir.

Bu çalışma diğer puanlayıcılar ve varyans kaynakları ile tekrarlanabilir. Başka bir deyişle, diğer hata kaynaklarını ve bunların etkileşimlerini değerlendirmek ve kapsamlı bir güvenilirlik analizi yapmak için G Kuramı uygulanabilir.