Sample Size for Estimation of G and Phi Coefficients in Generalizability Theory

Hakan ATILGAN[*]

Suggested Citation:

Atilgan, H. (2013). Sample size for estimation of g and phi coefficients in generalizability theory. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, 51, 215-228.

Abstract

Problem Statement: Reliability, which refers to the degree to which measurement results are free from measurement errors, as well as its estimation, is an important issue in psychometrics. Several methods for estimating reliability have been suggested by various theories in the field of psychometrics. One of these theories is the generalizability theory. In generalizability theory, two distinct reliability coefficients are estimated: the generalizability coefficient (G coefficient) for relative evaluation, and the index of dependability (Phi coefficient) for absolute decisions. Like in all methods of reliability estimation, G and Phi coefficients are estimated based on a data set obtained from a sample as a result of administering the instrument. Therefore, it has been a critical issue to determine what sample size is necessary in order to reliably estimate the population's characteristics.

Purpose of Study: The purpose of this study is to determine the adequate sample size required to ensure that the G and Phi coefficients obtained from a sample can estimate the G and Phi coefficients for the population in an unbiased way.

Methods: A total of 480691 students who took Form A of the SBS test for the 6th grade in 2008 were considered as the population of the study. Using a bootstrap method, a total of 1200 students were selected from this population, randomly falling into 12 subgroups consisting of different sample sizes (n=30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000), with each sample size having 100 replications. Since the test battery contained five subtests with distinct contents and numbers of items, and all items were replied to by all participants, a $p^{\bullet} x\ i^{\circ}$ multivariate G

[*] Dr., Ege University, Faculty of Education, Department of Educational Measurement and Evaluation, hakan.atilgan@ege.edu.tr

theory design was used. G and Phi reliability coefficients were estimated both for the population and each of the 12 distinct samples of different sizes. The relative root mean square error (R-RMSE) index was used as the error index to analyze the consistency of the G and Phi coefficients with the G and Phi parameters estimated for the population.

Findings and Results: It was found that the G and Phi coefficients estimated for a sample size of 30 tended to be less than the G and Phi parameters, and the R-RMSE value was greater than .01. When the sample size was 50 or more, R-RMSE values were less than .01. Thus it can be said that G and Phi coefficients are robust estimators of G and Phi parameters. Moreover, it was concluded that where the sample size is 400 or greater, R-RMSE values become stable. It was seen that a sample size of 400 is a more exact and robust estimator of G and Phi parameters, and increasing the sample size over 400 does not make a significant contribution to the unbiased estimation of G and Phi parameters.

Conclusions and Recommendations: A sample size of 30 does not provide an adequately unbiased estimation of G and Phi coefficients. It can be recommended that sample sizes of 50 to 300 are adequate for a robust estimation of G and Phi coefficients; however, a more exact and robust estimation requires a sample size of 400. In future research, the sample size for facets using different designs of G theory can be studied.

Keywords: Generalizability theory, sample size, generalizability coefficient, Phi coefficient

Due to the measurement errors present in educational and psychological measurements, accurate scores cannot be obtained. When measuring a variable it is desirable to obtain measurement scores as close to the real measure as possible. Therefore, reliability, which refers to the degree to which measurement results are free from measurement errors, as well as its estimation, play central roles in psychometrics. Several methods for estimating reliability have been suggested by various theories in the field of psychometrics. These methods of estimating reliability are statistics estimated based on a data set obtained from a sample as a result of administering an instrument. Therefore, it is critical to determine the sample size in order to estimate the reliability of the population. An adequate sample size must be used in order to accurately estimate reliability while ensuring economy in administering the instrument.

There are many suggestions in the psychometric literature about adequate sample sizes required to estimate reliability. Kline (1986), for example, reports that samples in reliability analysis must be contain 200 or more data points. On the other hand, Nunnally and Bernstein (1994) stress that a large sample size should be used to minimize sample errors, thus estimating reliability confidents accurately, and suggest that sample sizes should be 300 or more. However, Segall (1994) states that a sample size of 300 is small for reliability estimation. Charter (1999, 2003) recommends a sample size of 400 to estimate the population reliability precisely. On the other hand, Yurdugül (2008) reported that if the first eigenvalue of a measurement cluster is greater than six, a sample size of 30 is adequate; if the eigenvalue is between 3 and 6, a sample size of 100 is adequate; and if the eigenvalue is less than 3, a sample size of 300 or more is adequate. No upper limit is recommended for the sample size in reliability literature, but the ideal sample size has been discussed. In addition, Felt and Ankenmann (1998, 1999) stated that it is ill-advised to employ a sample size of less than 30, and Charter (2008) also suggested that levels below this threshold are unwise. One reliability estimation theory is Generalizability Theory, which was developed by Lee J. Cronbach et al. in 1972 based on the shortcomings of classical test theories (Crocker & Algina, 1986; Shavelson & Webb, 1991; Nunnally & Bernstein, 1994; Brennan, 2001a).

Generalizability Theory (G Theory) enables the assessment of reliability in behavioural measurements, and the design, research, and conceptualization of reliable observations (Shavelson & Webb, 1991; Brennan, 2001a). G Theory was first put forward by Cronbach et al. as a reaction to the shortcomings of the still popular real score model of classical reliability theory. Classical reliability theory considers the errors inherent in the measurement results to be errors coming from a single source. On the contrary, G theory considers the errors coming from all potential error sources together, as well as their interaction effects (Breannan, 2011). The purpose of G theory is to generalize the observed scores of measured subjects to the population scores accurately by defining and interpreting the measurement results and distinguishing different sources of variance. G theory assumes that the reliability of an observation depends on the studied population (Crocker & Algina, 1986; Shavelson & Webb 1991; Brennan, 2001a)

G theory takes into consideration two means of estimating reliability in education and psychology: relative and absolute evaluation. Therefore, in G theory two distinct reliability coefficients are estimated: a generalizability coefficient (G coefficient) for relative evaluations, and an index of dependability (Phi coefficient) for absolute decisions (Crocker & Algina, 1986; Shavelson & Webb, 1991; Brennan, 2001a).

Used for relative evaluations and symbolized by $\mathrm{E}\rho^2$, the G coefficient is defined as equal to the proportion of universe score variance [$\sigma^2(p)$] to the sum of the same variance and relative error variance [$\sigma^2(\delta)$]:

$$E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\delta)} \qquad (1)$$

Used for absolute evaluations and symbolized by Φ, the Phi coefficient is defined as equal to the proportion of universe score variance [$\sigma^2(p)$] to the sum of the same variance and absolute error variance [$\sigma^2(\Delta)$]:

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\Delta)} \tag{2}$$

(Shavelson & Webb 1991; Brennan, 2001a)

Cronbach et al. (1972) warn that variance components used in estimating the G and Phi coefficients can be unstable, depending on the sample size. Smith (1978) stresses that using a small sample size does not provide a sound ground in estimating the G and Phi coefficients, and if a sample size is small the G and Phi coefficients will not be stable. The issue of adequate sample size in estimating the G and Phi coefficients needs to be studied within generalizability theory (Shumate, Surles, Johnson & Penny, 2007). This study investigated the adequate sample size that will ensure that the G and Phi coefficients obtained from the sample can estimate the G and Phi coefficients for the population in an unbiased way.

Method

The Instrument and Data Collection

The results of the SBS test for the 6th grade held by the Ministry of National Education (MoNE) in 2008 were used in the study. This test consisted of five subtests with 80 multiple-choice items (4 choices per item). The Turkish subtest consisted of 19 items, and the Math, Science, and Social Studies subtests comprised 16 items each. The foreign language subtest contained 13 items. The answers (A, B, C, and D) given by 480691 students on "Form A" of the test were converted into a 1-0 matrix according to the answer keys of the relevant subtests, and this matrix was used in the study.

Population and samples

The study population consisted of a total of 480691 students who took Form A of the SBS test for the 6th grade in 2008. Using a bootstrap method, a total of 1200 students were selected from this population, randomly falling into 12 subgroups consisting of different sample sizes (n=30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000), with each sample size having 100 replications. Analyses were carried out on 1200 samples selected in this manner.

Data Analysis

The test battery used in the study consists of five subtests: In this test, there is a different set of items nested within each of the levels of the fixed facets, such as the Turkish, Math, Science, Social Studies, and Foreign Language. Brennan (2001) states that a model consisting of such subtests and items is called a "table of specifications." In this case, if all students taking the test (p) answer all of the items (i) in each of the subtests (s) ("x" is crosed with and ":" nested within), the model is defined as $p\ x\ (i{:}s)$. If, in this case, the number of items in each subtest is equal, the test follows a balanced design. However, when the number of items in each subtest is unequal it becomes an unbalanced design as a mixed model. Brennan (2001, p. 86) suggests that "unbalanced designs with mixed models are best treated using multivariate generalizability theory." Since the test used in this study consisted of five subtests with different numbers of items and all items are replied by all students, $p^{\bullet}\ x\ i^{c}$ multivariate G theory design was used (Brennan, 2001a; 2001b). In the design, a superscript filed circle • shows that the facet is crossed with the fixed multivariate variables and a superscript empty circle° shows that the facet is nested within fixed multivariate variables. In this design, the analyses were done using the PC version of mGENOVA 2.1. G and Phi coefficients were first calculated for the population. Next, G and phi coefficients were estimated for 1200 students in 12 subgroups consisting of different sample sizes (n=30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000), with each having 100 replications. Then, consistency between the G and Phi coefficients estimated from each of the sample sizes consisting of 100 samples (n=30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000) and the G and Phi parameters calculated for the population were analyzed. The relative root mean square error (R-RMSE) index was used as the error index for G and Phi coefficients.

$$R-RMSE\left(E\overset{2}{\rho_i}\right)=\sqrt{\frac{1}{M}\sum_{J=1}^{M}\frac{(E\overset{2}{\rho_{ij}}-E\rho^2)^2}{E\rho^2}} \tag{3}$$

$$R-RMSE\left(\Phi_i\right)=\sqrt{\frac{1}{M}\sum_{J=1}^{M}\frac{(\Phi_{ij}-\Phi)^2}{\Phi}} \tag{4}$$

In this equation $E\rho^2$ represents the G coefficient of a population, and Φ represents the Phi coefficient of the population. $E\overset{2}{\rho_{ij}}$ and Φ_{ij} respectively represent the G and Phi coefficients estimated from the jth sample for i sample size. The M in the equation represents the number of replications selected for each of the sample sizes using the simple random sampling method. In this study, 100 samples of M=100 were selected for each of the sample sizes (i=30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000) using the simple random sampling method.

For sample size studies based on simulation, R-RMSE is taken into consideration (Yurdugül, 2008), and an R-RMSE value closer to zero indicates robust estimation of a parameter (Yurdugül, 2009). If the estimated G and Phi values are equal to the G and Phi parameters, this indicates excellent consistency and the R-RMSE value is zero. As the R-RMSE values calculated as an error index get closer to zero, the G and Phi coefficients estimated from the samples can be said to be more robust estimators of the population G and Phi parameters. In this study it was assumed that when R-RMSE values are less than .01, the estimated G and Phi coefficients are robust estimators of the real G and Phi parameters.

Findings and Results

G and Phi parameters were calculated for the data set obtained from the population of 480691 using $p^{\bullet}\ x\ i^{\circ}$ multivariate G theory design. The G parameter value was calculated as .95774 and the Phi parameter value was .95397. Next, a total of 1200 students were selected from this population using a random sampling method creating 12 subgroups consisting of different sample sizes (n=30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000), each consisting of 100 samples. G and Phi coefficients were estimated for each of the samples selected. Graphics (see Figure 1) were produced in order to show how the G and Phi coefficients estimated from the samples changed according to sample sizes.

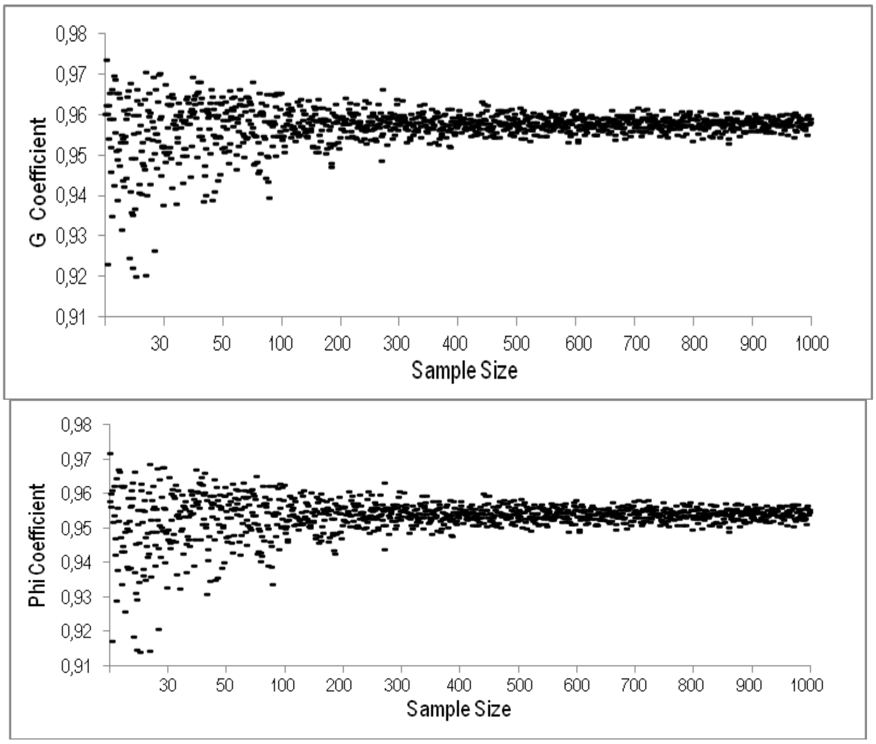

Figure 1. G and Phi coefficients estimated from different sample sizes

Figure 1 shows that G and Phi coefficients estimated from different sample sizes get closer to each other and form a narrowing cone as sample size increases. It was found that the G and Phi coefficients estimated for sample sizes of 30 tended to be lower than the G and Phi parameters ($E\rho^2$=.95774 and Φ=.95397). However, it can be said that when sample size was increased to 50, 100, 200, or 300, the consistency of both estimated G and Phi coefficients increases relatively and gets closer to the parameter values. It is seen in Figure 1 that when the sample size is 400, 500, 600, 700, 800, 900, or 1000, the estimated G and Phi coefficients are more stable, and when sample size is increased over 400 the consistency of the estimated G and Phi coefficients does not increase significantly.

The relative root mean square error (R-RMSE) index was used as the error index to analyze the consistency of the G and Phi coefficients estimated for 12 different sample sizes (100 samples per size) selected from the population with the G and Phi parameters estimated for the population. The R-RMSE values of the G and Phi coefficients estimated for each sample size (n=30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000) are shown in Table 1.

Table 1

R-RMSE Values of G and Phi Coefficients Estimated According to Each Sample Size

	R-RMSE			R-RMSE	
Sample Size (n)	G coefficient $(E\hat{\rho}^2)$	Phi coefficient Φ	Sample Size (n)	G coefficient $(E\hat{\rho}^2)$	Phi coefficient (Φ)
30	.01334	.01437	500	.00201	.00219
50	.00758	.00842	600	.00188	.00206
100	.00606	.00673	700	.00171	.00189
200	.00376	.00422	800	.00170	.00186
300	.00286	.00316	900	.00166	.00180
400	.00234	.00259	1000	.00140	.00152

The minimum and maximum R-RMSE values of the G coefficients estimated for each of the sample sizes ranged between .00140 and .01334. The minimum and maximum R-RMSE values of the Phi coefficients estimated for each of the sample sizes ranged between .00152 and .01437. The R-RMSE values given in Table 1 were found to be greater than .01 for both G and Phi coefficients when the sample size is 30. Thus it can be said that a sample size of 30 is too small to estimate the G and Phi coefficients; an adequately unbiased estimation is not possible when the sample size is 30. When the sample size is 50 or greater (n=50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000), the R-RMSE values were found to be less than .01, which suggests that G and Phi coefficients estimated from these sample sizes are robust estimators of G and Phi parameters.

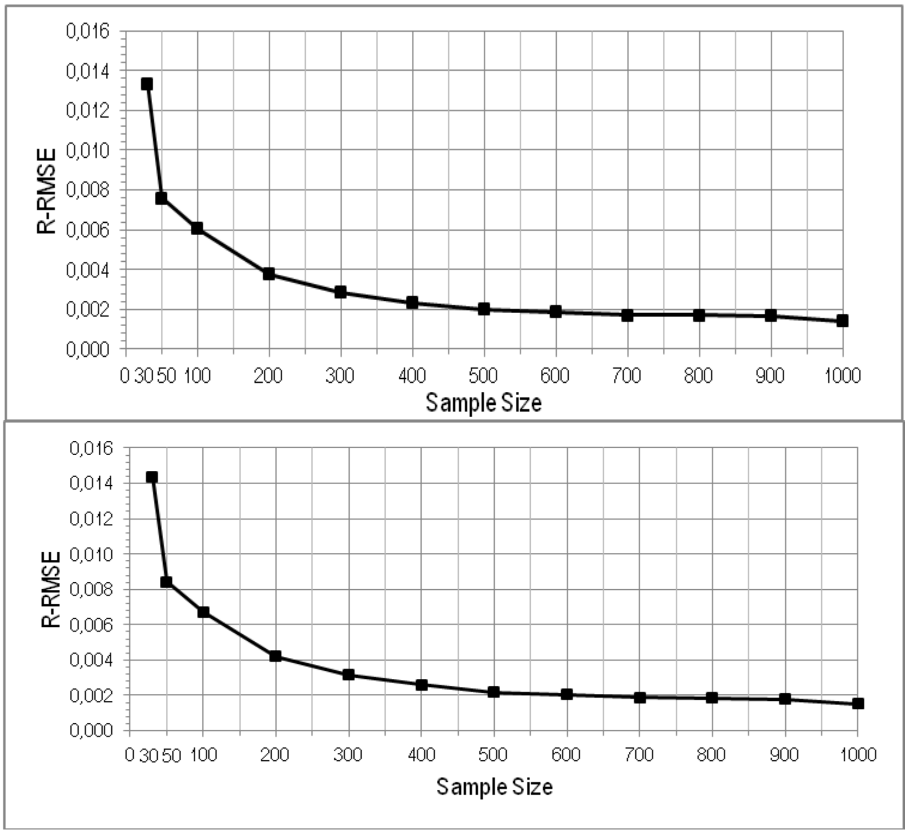

Figure 2. Change in R-RMSE values for G and Phi coefficients estimated, by sample size

Figure 2 shows how R-RMSE values of estimated G and Phi coefficients change according to sample size. As mentioned earlier, when the sample size is 50 or more, R-RMSE values drop below .01. It can be said that when the sample size is 50 or more, G and Phi coefficients can be estimated in an unbiased way, and these sample sizes are robust estimators of G and Phi parameters. On the other hand, as shown in Figure 2, when the sample size is 400 or more (n=400, 500, 600, 700, 800, 900, 1000), R-RMSE values become stable and do not change significantly. This suggests that increasing the sample size over 400 does not significantly improve the unbiased estimation of G and Phi parameters.

Conclusions and Recommendations

In this study the data set contains 480691 students, who took a test containing subtests with dichotomous (1-0) scaling, was taken as the population of the study.. Using a bootstrap method, a total of 1200 students were selected from this population, randomly falling into 12 subgroups consisting of different sample sizes (n=30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000), with each sample size being replicated 100 times. Since the test battery contained five subtests with distinct contents and numbers of items, and all items were replied to by all participants, $p^{\bullet}\ x\ i^{c}$ multivariate G theory design was used. Based on this design, the G and Phi coefficients estimated from different sample sizes were compared to the G and Phi parameters estimated for the population. As a result, it was concluded that when the sample size is 30, estimated G and Phi coefficients are not robust estimators of G and Phi parameters, and the R-RMSE value is greater than.01. This result supports that of Felt and Ankenmann (1998, 1999), who suggested that it is ill-advised to estimate reliability if the sample size is less than 30, as well as the conclusion reached by Charter (2008), who suggested that it is not wise to allow sample sizes below 30. It also verifies the warning by Cronbach et al. (1972) that variance components used in estimating the G and Phi coefficients can be unstable depending on the sample size, as well as the suggestion by Smith (1978) that when a small sample size is used G and Phi coefficients will not be stable. It is seen in this study that the R-RMSE values calculated for the G and Phi coefficients estimated for sample sizes of 50, 100, 200, and 300 are less than .01; i.e., they are robust estimators of G and Phi parameters. This finding is consistent with the findings of previous research, including that of Kline (1986), who recommended a sample size of 200 for reliability studies; Yurdugül (2008), who recommended sample size of 300 or more when the first eigenvalue is less than 3; and Nunnally and Bernstein (1994), who recommended a sample size of 300 or more. However, the findings of this study do not support Segall's (1994) suggestion that a sample size of 300 is small for reliability estimation. As a matter of fact, this study found that a sample size of 300 is enough to make an adequately unbiased reliability estimation (G and Phi). On the other hand, it was observed that when the sample size is 400 or more (n=400, 500, 600, 700, 800, 900, 1000), the estimated G and Phi coefficients move considerably closer to the G and Phi parameters, and become stable in sample sizes over 400. It was also found that R-RMSE values calculated for sample sizes of 400 and over (n=400, 500, 600, 700, 800, 900, 1000) were quite small, around .002, which suggests that these sample sizes offer more exact and robust estimation of G and Phi coefficients. At the same time, when the sample size is 400 or more (n=400, 500, 600, 700, 800, 900, 1000), estimated G and Phi coefficients become more stable. This shows that a sample size of 400 is a more exact and robust estimator of G and Phi parameters, supporting Charter (1999), who recommends a sample size of 400 to estimate the population reliability precisely. Moreover, it was seen that increasing the sample size over 400 does not make a significant contribution to the unbiased estimation of G and Phi parameters.

As a result, it was found that when the sample size is as small as 30, the G and Phi coefficients cannot be estimated in a stable way. On the other hand, it was concluded that when the sample size is 50, 100, 200, or 300, G and Phi coefficients can be estimated in an adequately unbiased way. Given a sample size of 400, the estimations given by the G and Phi coefficients are more exact and robust. Nevertheless, it can be said that increasing the sample size over 400 does not make a significant contribution to the unbiased estimation of G and Phi coefficients. A sample size of 50 to 300 can be thought adequate for the robust estimation of G and Phi coefficients; however, a more exact and robust estimation requires a sample size of 400.

In this study, $p^{\bullet} \ x \ i^{\circ}$ multivariate G theory design was used to estimate the G and Phi coefficients for a sample of people. By its nature, G theory estimates single G and Phi coefficients by evaluating different error sources together. Therefore, the sample size for different facets, including different items, time, scorers, etc., can be studied for different designs of G theory in future researches.

References

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

Brennan, J. R. (2001b). *Manual for mGENOVA.* City, IA: Iowa Testing Program, University of Iowa.

Brennan, L. R. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24:1-21.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, 21(4), 559-566.

Charter, R. A. (2003). Study sample are too small to produce sufficiently precise reliability coefficients. *The Journal of General Psychology*, 130(2), 117-129.

Charter, R. A. (2008). Statistical approaches to achieving sufficiently high test score reliability for research purposes. *The Journal of General Psychology*, 135(3), 241-251.

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

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for score and profiles*. New York: Wideliy.

Felt, L. S., & Ankenmann, R. D. (1998). Appropriate sample size for comparing alpha reliabilities. *Applied Psychological Measurement,* 22, 170-178.

Felt, L. S., & Ankenmann, R. D. (1999). Determining sample size for a test of the equality of alpha coefficients when the number of part-tests is small. *Psychological Methods*, 4, 366-377.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. New York: Methuen.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd ed).*New York: McGraw-Hill.

Segall, D. O. (1994). The reliability of linearly equated tests. *Psychometrika*, 59, 361-375.

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

Shumate, S. R., Surles, J., Johnson, R. L: & Penny, J. (2007). The effects of number of scale point and non-normality on the generalizability coefficient: A monte carlo study. *Applied Measurement in Education,* 20 (4), 357-376.

Smith, P. (1978). Sampling errors of variance components in small multifacet generalizability studies. *Journal of Educational Statistics*, 3, 319-346.

Yurdugül, H. (2008). Minimum sample size for Cronbach's coefficient alpha: A monte-carlo study. *H. U. Journal of Education*, 35: 397-405.

Yurdugül, H. (2009). The comparison of four different coefficient alphas from a psychometric point of view. *H. U. Journal of Education*, 36, 327-339.

Genellenebilirlik Kuramında G ve Phi Katsayılarının Kestirilmesi için Örneklem Büyüklüğü

Atıf:

Atilgan, H. (2013). Sample size for estimation of g and phi coefficients in generalizability theory. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, 51, 215-228.

(Özet)

Problem Durumu

Eğitimde ve psikolojide ölçme sonuçlarına karışan ölçme hataları nedeniyle yapılan ölçme ile gerçek puana ulaşılamaz. Yapılan ölçmeler ile ölçülen özelliğin gerçek puanına olabildiğince yakın ölçme sonuçları elde edilmek istenir. Bu nedenle; ölçme sonuçlarının ölçme hatalarından ne derece arınık olduğu anlamına gelen güvenirlik kavramı ve güvenirliğin tahmin edilmesi psikometri alanında önemli bir yer tutmaktadır. Öyle ki psikometri alanında geliştirilen kuramlar ile pek çok güvenirlik tahmin metodu önerilmiştir. Güvenirlik tahmin metodu öneren kuramlardan biri de Genellenebilirlik Kuramıdır. Genellenebilirlik kuramıyla bağıl değerlendirmeler için Genellenebilirlik (G) katsayısı ve mutlak değerlendirmeler için güvenirlik (Phi) katsayısı olmak üzere iki farklı güvenirlik katsayısı hesaplanır. Tüm güvenirlik kestirme metotlarında olduğu gibi Genellenebilirlik kuramında da G ve Phi katsayıları ölçme aracının bir birey örneklemine uygulanması ile elde edilecek örneklem puan dağılımından hesaplanan bir istatistiktir. Bu nedenle popülasyon güvenirliğinin tahmin edilmesi için örneklem büyüklüğünün ne olması gerektiği önemli bir soru olagelmiştir. Genel olarak güvenirlik kestirme çalışmalarında örneklem büyüklüğünün ne olması gerektiği konusunda psikometri literatürde farklı öneriler bulunmaktadır.

Araştırmanın Amacı

Genellenebilirlik kuramında G ve Phi katsayılarının hesaplanmasında kullanılan varyans bileşenlerinin örneklem büyüklüğüne bağlı olarak değişiklik gösterebilir. G ve Phi katsayılarının kestirilmesi için örneklem büyüklüğünün yeterli olması durumunda G ve Phi katsayıları doğru olarak kestirilemez. Bu nedenle G ve Phi katsayılarının kestirilmesi için uygun örneklem büyüklüğünün ne olması gerektiği genellenebilirlik kuramında çalışılması gereken bir alandır. Bu çalışmada, örneklemden elde edilen G ve Phi katsayılarının evren G ve Phi katsayılarını yansız olarak kestirebilmesi için örneklem büyüklüğünün ne olması gerektiği araştırılmıştır.

Araştırmanın Yöntemi

2008 yılında yapılan 6. Sınıf Seviye Belirleme Sınavı (SBS) testi "A" formunu alan 480691 kişi evren olarak kabul edilmiştir. Evren olarak kabul edilen bu veri setinden bootstrap metoduyla 12 farklı örneklem büyüklüğünde (n=30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000) ve her bir örneklem büyüklüğü için 100 tane olmak üzere toplam 1200 örneklem basit seçkisiz olarak çekilmiştir. Verilerin elde edildiği testte, madde sayıları eşit olmayan farklı kapsamda beş alt test bulunduğundan ve tüm maddeleri tüm bireyler yanıtladığından $p^{\bullet} x\ i^{c}$ çok değişkenli G kuramı deseni kullanılmıştır. Evren için ve bu evrenden 12 farklı örneklem büyüklüğünde çekilen örneklemler için G ve Phi katsayıları hesaplanmıştır. G ve Phi katsayılarının evren için hesaplanan G ve Phi parametreleri ile tutarlılıkları incelemek için hata indeksi olarak göreli hata kareler ortalaması karekökü (R-RMSE) kullanılmıştır. Hata indeksi olarak elde edilen R-RMSE değerleri sıfıra yaklaştıkça örneklemlerden kestirilen G ve Phi katsayılarının G ve Phi parametrelerinin sağlam kestiricisi olduğu söylenebilir. Bu çalışmada R-RMSE değerlerinin 0,01'den küçük olması durumunda kestirilen G ve Phi katsayılarının G ve Phi parametrelerinin sağlam kestiricisi olduğu kabul edilmiştir.

Araştırmanın Bulguları

Örneklem büyüklüğü 30 için kestirilen G ve Phi katsayılarının G ve Phi parametrelerinden küçük çıkma eğiliminde olduğu ve R-RMSE değerinin 0,01'den büyük çıktığı görülmüştür. Bununla birlikte örneklem büyüklükleri 50, 100, 200 ve 300 olarak arttığında, hem kestirilen G hem de kestirilen Phi katsayılarının göreli olarak tutarlılıklarının arttığı ve parametre değerlerine giderek yaklaştığı söylenebilir. Örneklem büyüklüğü 50 ve üstünde olduğunda R-RMSE değerleri 0,01'den küçük bulunduğundan G ve Phi katsayılarının G ve Phi parametrelerinin sağlam kestiricisi olduğu söylenebilir. Bununla birlikte, örneklem büyüklüğü 400, 500, 600, 700, 800, 900 ve 1000 olduğunda kestirilen G ve Phi katsayılarının daha kararlı davrandıkları, fakat örneklem büyüklüğünün 400'den sonra artırılması durumunda kestirilen G ve Phi katsayılarının tutarlılığının göreli olarak fazlaca değiştirmediği sonucuna ulaşılmıştır. Örneklem büyüklüğü 400 olduğunda G ve Phi parametrelerinin daha kesin ve daha sağlam kestirildiği, örneklem büyüklüğünün 400'den sonra artırılmanın G ve Phi parametrelerinin yansız kestirilmesinde önemli bir katkı sağlamadığını görülmüştür.

Araştırmanın Sonuçları ve Öneriler

G ve Phi katsayılarının kestirilmesi için örneklem büyüklüğünün 30 gibi küçük bir örneklem olması durumunda G ve Phi katsayılarının istikrarlı olarak kestirilemediği görülmüştür. Diğer yandan örneklem büyüklüğünün 50, 100, 200 ve 300 olması durumunda G ve Phi katsayılarının yeterince yansız olarak kestirilebileceği, ancak 400 örneklem büyüklüğünde ise G ve Phi katsayılarının daha kesin ve daha sağlam olduğu sonucuna varılmıştır. Diğer yandan örneklem büyüklüğünün 400'den sonra artırılmasının G ve Phi katsayılarının yansız olarak kestirilmesine katkı sağlamadığı söylenebilir. G ve Phi katsayılarının sağlam kestirilmesi için örneklem büyüklüğünün 50 ile 300 arasında olması, ancak daha kesin ve daha sağlam kestirme için örneklem büyüklüğünün 400 olması önerilebilir.

Bu çalışmada G ve Phi katsayılarının kestirilmesinde kişi örneklemi üzerinde, $p^{\bullet} \; x \; i^{\circ}$ multivariate G kuramı deseni ile çalışılmıştır. G kuramı özelliği gereği farklı hata kaynaklarını birlikte değerlendirerek tek bir G ve Phi katsayılarını kestiren bir kuramdır. Bu nedenle; madde, zaman, puanlayıcı vb. farklı hata kaynaklarının yer aldığı G kuramının farklı desenlerinde bu hata kaynakları için örneklem büyüklükleri çalışılabilir.

Anahtar Sözcükler: Genellenebilirlik Kuramı, Örneklem Büyüklüğü, Genellenebilirlik Katsayısı, Phi Katsayısı