The Examination of Reliability According to Classical Test and Generalizability on a Job Performance Scale

Atilla YELBOĞA, Ezel TAVŞANCIL***

Abstract

In this research, the classical test theory and generalizability theory analyses were carried out with the data obtained by a job performance scale for the years 2005 and 2006. The reliability coefficients obtained (estimated) from the classical test theory and generalizability theory analyses were compared. In classical test theory, test retest and Cronbach's alpha reliability coefficient, Kendall's concordance coefficient for interrater reliability; in generalizability theory, G and Phi coefficients with multivariate model were computed and the consistency of these coefficients were examined. As the result of the research, it is seen that for the same measurement condition, the reliability coefficients obtained from the classical test theory and from the multivariate model of generalizability theory generated harmonious results.

Key Words

Reliability, Generalizability Theory, Classical Test Theory.

* *Correspondence*: Atilla Yelboğa, PhD., Ankara University, Measurement and Evaluation Application and Research Center, Advisory Board Member, Ankara/Turkey.
E-mail: ayelboga@gmail.com
** Prof. Ezel Tavşancıl, Ankara University, Faculty of Educational Sciences, Department of Measurement and Evaluation, Cebeci, Ankara/Turkey.

Kuram ve Uygulamada Eğitim Bilimleri / Educational Sciences: Theory & Practice
10 (3) • Summer 2010 • 1847-1854

The essence of Spearman's model was that any observed score could be envisioned as the composite of two hypothetical components (a true score component and a random error component) (Crocker & Algina, 1986). The true score formula can be written as Observed score= True score + Random error. The relationship between true score and observed score as expressed in a mathematical model was explained by Spearman in what has become known as the true score model or classical test theory. Based on the positive or negative influences of the random error component, it is possible for a respondent's observed score to be higher or lower than the same respondent's true score (Anastasi, 1988; Murphy & Davidshoper, 1998; Rust & Golombok, 1992; Tavşancıl, 2002; Thorndike, 1988).

The classical test theory partitions observed score variance into only two components, true score variance and random score variance (Guilford, 1954; Gulliksen, 1967; Lord & Novick, 1968). Consequently, it is possible to examine only a single source of measurement error at any given time. This poses a serious dilemma for the researcher, since in reality several types of measurement errors can exist concurrently (Baykul, 2000; Erkuş, 2003). In classical test theory, it possible to consider estimates of measurement error due to inconsistency in forms (equivalence), observers (interrater agreement), sampling the item domain (internal consistency or split-half), or time (test-retest or stability). Only one measurement error influence can be considered in a given analysis.

Each type of reliability estimate can be used to determine the degree to which true scores deviate from observed scores. The problem, however, is that classical test theory is unable to examine inconsistencies in test forms, raters, items, or occasions simultaneously (Brennan, 2001; Shavelson & Webb, 1991).

Generalizability theory, which subsumes classical test theory as a special case, also extends classical test theory by recognizing and estimating the magnitude of the multiple sources of error (Brennan, 2001; Crocker & Algina, 1986). Both sources of error variance and interactions among these sources can be considered simultaneously in a single generalizability analysis (Cronbach, Rajaratnam, & Gleser, 1963). Classical test theory admits consideration of only one type of measurement error at a time and does not consider the possible, completely independent, or separate interaction effects of the measurement error variance.

In G theory, unlike classical test theory, a behavioral observation is considered only one sample of behavior from an infinite universe of admissible observations (Atılgan & Tezbaşaran, 2005). The universe, therefore, is comprised of all the potential measurements that would be a direct substitute for the observation under investigation. The particular sample or observed score is not the focus of the generalizability analysis; rather, the analysis focuses on how well a particular sample of behavior generalizes to the larger universe of admissible observations.

An individual observation or measurement is merely a sample estimate from an infinite and varied universe of possible measurements. Measurements could be taken on a multitude of occasions, using various items and forms of measuring instruments. Individual measurements are estimates of an individual's true scores and be parts of a universe of admissible observations. Forms, items, occasions, and raters are called facets, which can include any characteristic of a measurement procedure that is potential source of error. The levels within each facet are conditions that can be infinitely large (Shavelson & Webb, 1991). The object of measurement is usually persons, which is typically not considered a source of error, because people vary and their score differences are real, systematic, and of great interest to investigators. The objects of measurement are to create error variance and are, therefore, not considered a facet. Anything that generates systematic variance can be the object of measurement. Other possibilities include schools, businesses, work groups, or occasions.

G theory decomposes, estimates, and reveals all measurement errors, which are termed variance components (Brennan, 2001; Tobar, Stegner & Kane, 1999). A G coefficient is produced for each data set representing the universe score variance divided by the observed score variance. G coefficients range between zero and one. The data for G theory studies can be crosses or nested. In a fully crossed data set, the objects of measurement have scores on all of the same facets and conditions. The data are nested when is not the case that is, when each condition of a facet does not appear in combination with each condition of other facets. Facets can be fixed or random. A facet is random if its conditions can be exchanged with any other of the conditions from the same facet (Kieffer, 1999). Facets are considered fixed when their levels are not exchangeable and are of specific interest to the investigator, as in many experiments.

Generalizability or G-study results can also be used to conduct so-called D(esign) studies to address important "what if" questions about variation in measurement design. Sources of error can be pinpointed and protocol modifications can be specified that will result in the desired level of generalizability (Brennan, 2001; Shavelson & Webb, 2004).

Decisions made in the context of cutoff scores (absolute decisions), as against decisions only considering stability in a relative standing, can be considered. Classical test theory does not admit a distinction between reliability involving absolute decisions made in the context of cutoff scores as against reliability involving decisions only considering stability in relative standing or rankings.

In generalizability studies, the coefficients that address reliability in the context of relative decisions are called generalizability coefficients. The coefficients that address reliability in the context of absolute decisions are called phi coefficients. A relative G coefficient reflects the degree to which the objects of measurement maintain rank order across facets, regardless of possible changes in raw score elevations. It is analogous to the reliability coefficients of classical test theory. Absolute G coefficients are more stringent and reflect both the degree of consistency in the rank ordering of measurement objects and the consistency in the elevations of the raw scores. Absolute G coefficients are most useful when the actual values of the obtained scores are important or meaningful to the investigator. These typically involve performance measurements where there is a cutoff value that is deemed particularly meaningful. For technical reasons, the terms phi coefficients or index of dependability coefficients are used instead of absolute G coefficients (Shavelson & Webb, 1991). Only those variance components that reflect the rank ordering of persons are involved in the G estimates for relative decisions, whereas all of the variance components are involved in the G estimates for absolute decisions.

Generalizability theory focuses on the simultaneous influence of multiple sources of measurement error variance, and therefore more closely fits the interest of researchers. Regardless of the strengths, generalizability theory has not yet been widely applied to affects measures or specifically to measures performance. Therefore, in the present study, we investigate the psychometric properties of scores from job performance scale using both classical and generalizability test theories.

Method

Subjects

Data for the study were obtained from 176 people in a service sector company. All eligible persons in the surveyed company participated in the study. The sample consisted of managers, assistant managers, chiefs and specialists.

Instrumentation

The Job performance scale was selected for the present study based on more thoroughly explored psychometric properties (e.g., Yelboğa, 2003). It is a 32-item-4 dimension scale. Item responses are scored from 1 to 5, with higher scores indicating more success in job. Previous reliability studies employing classical test theory have reported alpha coefficient 0.94 (min. Cronbach's alpha 0.70 by Nunnally, 1978) and test-retest correlation 0.84 for job performance scale scores.

Procedure

Analysis was performed using SPSS for classical test theory analysis and a PC version of the GENOVA program which was especially developed for generalizability analysis by Brennan (2001). Generalizability analyses were conducted to partition systematic and measurement error sources of variance within the data. Persons were considered the objects of measurement. Error variance facets were time and items, as well as the interactions effects. Both generalizability and phi coefficients were computed. The variance partitions were then employed in decision, or D-study, analyses to explore the estimated effects on score integrity of selected changes in the measurement protocol.

Results

Table 1 presents the results from test-retest analysis on 2005 and 2006. Table 2 presents the results from cronbach alpha coefficient and Table 3 presents the results from Kendall's concordance coefficients. Table 4 and Table 5 present the results from generalizability analyses from 2005 and 2006, respectively. Table 4 and Table 5 results use the variance components and derive the coefficients for the different measurement protocols, including D-study estimates for measurement used in the present study. Both generalizability coefficients, associated with relative

decisions, and phi coefficients, associated with absolute decisions are presented. Table 6 and Table 7 present the reliability coefficients results from classical test and generalizability analyses from 2005 and 2006, respectively.

Discussion

The administration of the Job performance scale in 2005 and 2006, the study obtained by the classical test theory Cronbach's alpha coefficient, Kendall's W coefficient and G theory obtained by working with the reliability coefficient and coefficient of Φ compared levels;

a. Cronbach's alpha coefficient, $E\rho^2$ and Φ coefficients are consistent with each other in job performance subscale (dimensions).

b. Cronbach's alpha coefficient, $E\rho^2$ and Φ coefficients are consistent with each other in job performance scale.

c. Test retest reliability coefficient, $E\rho^2$ and Φ coefficients are consistent with each other in job performance subscale (dimensions).

d. Test retest reliability coefficient, $E\rho^2$ and Φ coefficients are consistent with each other in job performance scale.

e. Kendall W coefficient, $E\rho^2$ and Φ coefficients are consistent with each other in job performance scale.

Classical test theory is currently favored by researchers and the importance of a theory which is still continuing. However, in common with a single analysis and a single reliability coefficient in classical test theory cannot be obtained. Different reliability methods and the estimated reliability coefficients for the discrepancies are limitations of classical test theory.

G theory, the potential error sources in the event of more than one reliability analysis of the calculation done with a single ensure that, relative and absolute reliability assessments for a comprehensive theory is a factor that can cut. G theory, classical test theory methods that test retest reliability and internal consistency methods combine. This research is more than one measure of such potential sources of error for the case of G theory, constitute a good alternative to classical test theory.

This research requires a strong statistical foundation to have both psychometric work to bring ease of use in the theory of G in terms of

psychology and educational studies to determine the reliability of the classical test theory can be used in place of the findings (Kieffer, 1999) are supported. However, the group with the same data in two different theories of research, reliability estimates have revealed that (e.g., Crowley, Thompson & Worchol, 1994) should be taken into consideration.

This study considered two theories used in future research consider the following suggestions may be given to researchers;

1. This study used the scale of job performance, has high reliability coefficients with the classical test theory analysis have revealed. These findings were confirmed with G theory. Similar work, different scales have reliability coefficients can be done with the G theory produced consistency reliability coefficients will contribute to work psychometric testing.

2. Cut the level of reliability which techniques to use, substance related scale scores on the structure of assumptions, research conditions and vary depending on purpose, that is taken into consideration in this study, classical test theory and the G theory produced similar results were observed. Researchers to do similar study, both the number of different substances and evaluator with the ease of doing alternative analysis as well as achieve a single analysis because of the reliability G coefficient are encouraged to use the theory.

3. Reliability studies in classical test theory, can be done a single analysis with G theory. Considering time and cost status of the G theory is faster than classical test theory analysis can be said, and produced similar results. Related to this topic and research groups working with different scales is recommended to do.

4. In this study, the D studies in G theory was applied in the operation of different measurement scenarios. According to this scenario, the assessor by the number of different substances and different reliability coefficients has been cut. According to this scenario estimated accuracy of the results concerning the reliability of research will contribute to the field.

5. In the assessment of reliability in the context of this research, G theory model with multivariate analysis was (bxmxd). Similar studies with different models of the G theory examination will provide useful information in the field.

References/Kaynakça

Anastasi, A. (1988). *Psyhological testing.* New York: Macmillan.

Atılgan, H. ve Tezbaşaran, A. A. (2005). Genellenebilirlik kuramı alternatif karar çalışmaları ile senaryolar ve gerçek durumlar için elde edilen G ve Phi katsayılarının tutarlılığının incelenmesi. *Eğitim Araştırmaları Dergisi*, *18*, 236-252.

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme.* Ankara: Cem Web Ofset.

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

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

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

Crowley, S. L., Thompson, B., & Worchol F. (1994). Validity studies the children's depression inventory: A comparison of generalizability and classical test theory analyses. *Education and Psychological Measurement, 54*, 705-713

Erkuş, A. (2003). *Psikometri üzerine yazılar.* Ankara: Türk Psikologlar Derneği.

Guilford, J. P. (1954). *Psychometrics methods* (2nd ed). New York: McGraw-Hill.

Gulliksen, H. (1967). *Theory of mental tests* (6th ed). New York: John Willey and Sons.

Kieffer, K. M. (1999). Why generalizability theory is essential and classical test theory is often inadequate. In B. Thompson, (Ed.), *Advances in social science methodology* (Vol. 5, pp. 149-170). Stanford, CT: JAI.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* New Jersey: Addison-Wesley.

Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological testing* (4th ed). New Jersey: Prentice Hall.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed). New York: McGraw-Hill.

Rust, J., & Golombok S. (1992). *Modern psychometrics. The science of psychological assessment* (2nd ed). New York: Routledge.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A prime.* California: Sage.

Shavelson, R. J., & Webb, N. M. (2004). Generalizability theory. In Kemp-Leonard, K. (Ed.). *Encyclopedia social measurement* (pp. 99-105). Oxford, UK: Elsevier.

Tavşancıl, E. (2002). *Tutumların ölçülmesi ve SPSS ile veri analizi.* Ankara: Nobel.

Thorndike, R. L. (1988). Reliability. In J. P. Keeves (Ed). *Educational research methodology and measurement an international handbook* (pp.330-343). New York: Pergamon.

Tobar, D. A., Stegner, A. J., & Kane, M. T. (1999). The use of generalizability theory in examining the dependability of scores on the profile of mood states. *Measurement in Physical Education and Exercise Science*, *3*(3), 141-156.

Yelboğa, A. (2003). *İnsan kaynakları yönetiminde performans değerlendirilmesi için geliştirilen bir ölçeğin psikometrik özelliklerinin incelenmesi.* Yayımlanmamış yüksek lisans tezi, Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.