Eğitim ve Bilim
2012, Cilt 37, Sayı 163

Education and Science
2012, Vol. 37, No 163

Dependability of Job Performance Ratings According to Generalizability Theory

Genellenebilirlik Kuramı'na Göre İş Performansı Ölçeklerinde Güvenilirlik

Atilla YELBOĞA[*]

Ankara University

Abstract

This article introduces the application of Generalizability Theory in assessing the reliability of job performance ratings. Generalizability Theory is frequently used in educational sciences and psychology. Basically, Generalizability Theory can be used to assess reliability in the presence of multiple sources of error. Also, it can be used to assess reliability in the presence of different types of sources of error. In this study, the application of Generalizability Theory in measurement involving multiple raters is considered in particular. Generalizability Theory seems like an ideal theory for examining multiple sources of error in job performance measurement. With this study, principles of Generalizability Theory are used in determining measurement errors that occur while evaluating job performance.

Keywords: Dependability, generalizability theory, job performance.

Öz

Bu makale, iş performansı ölçeklerinde güvenirliğin değerlendirilmesinde Genellenebilirlik Kuramı'nın kullanılmasına bir giriş niteliğindedir. Genellenebilirlik Kuramı, eğitim bilimleri ve psikolojide sıklıkla kullanılan bir kuramdır. Temelde, Genellenebilirlik Kuramı birden fazla hata kaynağını aynı anda göz önüne alarak güvenirliği değerlendirir. Özellikle de birden fazla değerlendiricinin bulunduğu ölçme durumlarında kullanımı tercih edilmektedir. İş performansı ölçümünde birden fazla hata kaynağının değerlendirilmesi için Genellenebilirlik Kuramı uygun bir yöntemdir. Bu çalışmada, iş performansında ölçme hatalarının değerlendirilmesi için Genellenebilirlik Kuramı'nın ilkeleri uygulanarak sonuçları tartışılmıştır.

Anahtar Sözcükler: Güvenilirlik, Genellenebilirlik Kuramı, iş performansı.

Introduction

Job performance is the most important dependent variable in industrial-organizational psychology (Schmidt & Hunter, 1992). A general definition of the construct of job performance reflects behaviors (both visually observable and non-observable) that can be evaluated (Viswesvaran, Schmidt, & Ones, 1996).

Individual job performance can be measured utilizing different methods. However, these methods can be classified into two broad categories: 1) organizational records, and 2) subjective evaluations. Organizational records are considered to be more "objective", in contrast to the subjective evaluations that depend on a human judgment (Viswesvaran & Ones, 2005).

Performance ratings have traditionally played a central role in the measurement of job performance in industrial-organizational psychology (Viswesvaran, Ones, & Schmidt, 2002). Several measures of job performance have been used over the years as criterion measures (cf.

[*] PhD.,Atilla YELBOĞA, Ankara University, Measurement and Evaluation Application and Research Center, Advisory Board Member, e-mail: ayelboga@gmail.com

Portions of this research were presented at XII. European Congress of Psychology, Istanbul, TURKEY, July 4-8,2011.

Campbell, Gasser, & Oswald, 1996; Cleveland, Murphy, & Williams, 1989). Reliability of criteria has been included as an important consideration by all authors writing about job performance measurement (Viswesvaran, Schmidt, & Ones, 1996).

Of the different ways to measure job performance, performance ratings are the most prevalent. Ratings are subjective evaluations that can be obtained from supervisors, peers, subordinates, self, or customers, with supervisors being the most commonly used source and peers constituting the second most commonly used source (Cascio, 1991; Cleveland, Murphy, & Williams, 1989; Viswesvaran, Schmidt, & Ones, 1996).

Comparing the different types of reliability estimates (coefficient of equivalence, coefficient of stability, etc.) for each dimension of job performance is also valuable. Reliability of a measure is defined as the ratio of the true to observed variance (Nunnally, 1978). Different types of reliability coefficients assign different sources of variance to measurement error. In general, the most frequently used reliability coefficients associated with criterion ratings can be broadly classified into two categories: interrater and intrarater. In the context of performance measurement, interrater reliability assesses the extent to which different raters agree on the performance of different individuals. As such, individual raters' idiosyncratic perceptions of job performance are considered to be part of measurement error. Intrarater reliability, on the other hand, assigns any specific error unique to the individual rater to true variance. That is, each rater's idiosyncratic perceptions of job performance are relegated to the true variance component. Both coefficient alpha and the coefficient of stability (rate-rerate reliability with the same rater) are forms of intrarater reliability. Intrarater reliability is most frequently indexed by coefficient alpha computed on ratings from a single rater on the basis of the correlations or covariance's among different rating items or dimensions. Coefficient alpha assesses the extent to which the different items used to measure a criterion are indeed assessing the same criterion. Rate-rerate reliability computed using data from the same rater at two points in time assesses the extent to which there is consistency in performance appraisal ratings of a given rater over time. Both of these indices of intrarater reliability, coefficient alpha and coefficient of stability (over short period of times when it is assumed that true performance does not change), estimate what the correlation would be if the same rater rerated the same employees (Cronbach, 1951).

The choice of methods for estimating reliability is especially important when correcting for measurement error in ratings of job performance. These ratings are usually obtained from a single supervisor, who uses a multi-item performance appraisal form (Murphy & Cleveland, 1995). Two methods are widely used to estimate the reliability of performance ratings. First, measures of internal consistency (e.g., coefficient alpha) can be used to estimate intrarater reliability. As will be noted below, the use of internal consistency measures to estimate the amount of measurement error in ratings is most appropriate if the term "measurement error" is used to refer to the rater's inconsistency in evaluating different facets of a subordinate's job performance. Second, measures of agreement between raters can be used to estimate interrater reliability. The use of interrater agreement measures to estimate the amount of measurement error in ratings is most appropriate if the term "measurement error" is used to refer to disagreements between similarly situated raters about individuals' levels of job performance (Murphy & DeShon, 2000a).

Some researchers (e.g. Schmidt & Hunter, 1996; Viswesvaran, Schmidt, & Ones, 1996) have argued that correlations between ratings provided by multiple raters provide the "correct" estimates of the reliability of job performance ratings. On the other hand, Murphy & DeShon (2000a) showed that correlations between performance ratings obtained from separate supervisors cannot be interpreted as reliability coefficients, and that these correlations reflect a number of factors other than the influence of "true performance" and "random measurement error" on ratings.

Murphy & DeShon (2000a) questioned the use of interrater correlations as estimates of the reliability of job performance ratings and suggested that analyses based on Generalizability Theory would be more useful and informative. The main point of their critique was that raters can rarely be thought of as parallel tests, and that the correlations between ratings obtained in organizational settings reflected sources of variance other than "true score" and "error" as defined in the classic theory of reliability. They went on to note that alternatives to the parallel-test model that once dominated psychometric thinking were well known, and demonstrated their application to performance rating. Although these alternative approaches can be more difficult to implement than simpler methods based on interrater correlation, a large body of research demonstrates that methods based on the parallel test model are not appropriate in this context, and that more modern methods of estimating the psychometric characteristics of performance ratings are needed. Schmidt, Visweavaran & Ones (2000) criticized (Murphy & DeShon, 2000a)'s paper and labeled their conclusions as "radical." In particular, Schmidt, Visweavaran & Ones (2000) argue that: (a) the classic theory of reliability and the parallel test model derived from that theory are the appropriate for understanding performance ratings, and to suggest otherwise is nihilistic, (b) measurement models have little to do with substantive models of the processes that generate scores on a test or measure, and (c) reliability and validity are distinct concepts that should not be confused.

Classic reliability theory can be traced back to Spearman, and reached perhaps its highest point of development in Lord & Novick's (1968) classic text. Lord & Novick (1968) show that if you start with the definition that observed scores (X) equals true scores (T) plus error (e) and define "e" as a variable that is normally distributed with a mean of zero, uncorrelated with T and uncorrelated with "e"s obtained from other measures, it is possible to derive virtually the entire theory of psychometrics as it was presented in textbooks up to that time. Most important, this definition led directly to the formula for the correction for attenuation, which allows you to estimate the correlation among true scores. The definition of error as a variable that is uncorrelated with either T or with other "es" is central to this theory; unless this assumption is met, the correction for attenuation will not provide a valid estimate of the true score correlation.

Error in performance measurement has been shown to originate from multiple sources. Classical test theory approaches to determining score reliability, however, are not capable of identifying and untangling this profusion of error. Classical reliability was not conceptualized to do this; it accounts for only one error source, the consistency with raters evaluates a set of performances. Other potential sources remain but as undifferentiated error. A more advanced method is needed, one capable of accommodating multiple source of error and of placing findings into theoretical contexts beyond the local panels of raters found individual studies (Bergee, 2007).

Murphy & DeShon (2000b) suggested that the parallel test model was not useful for evaluating the psychometric characteristics of ratings, and suggested that Generalizability Theory provided a tighter foundation for such analyses. Their recommendations were based in part on decades of research on performance appraisal that demonstrates that raters are not parallel tests in any sense of the word.

Theoretical Framework of Generalizability Theory

Generalizability Theory (GT) is based on analysis of variance and provides a framework for examining the dependability (i.e,, reliability) of behavioral measurements (Cronbach, Gleser, Nanda & Rajaratnam, 1972; Cronbach, Rajaratnam & Gleser, 1963). In GT, a distinction is usually made between generalizability (G) studies and decision (D) studies (Shavelson & Webb, 1991). The purpose of a G-study is to simultaneously estimate multiple sources of variance (e.g., variance due to ratees, raters, items) within a single, multifaceted experiment (Atılgan & Tezbaşaran, 2005; Kan, 2007; Yelboğa & Tavşancıl, 2010). As such, in contrast to classical test theory (which only partitions observed variance into true and random error variance), GT provides more accurate estimates of the dependability of observations (Brennan, 1992; Brennan, 2000). The purpose of the D-study is to use the estimated variance components from the G-study to project reliability estimates under any number of measurement conditions modeled in the G-study (e.g., differing numbers of raters and items). Projecting reliabilities under different measurement conditions is useful because one may readily observe how to improve the dependability (i.e., reliability) of the observations. Few studies have used GT to analyze the dependability of performance ratings. Similar to the current study, these studies generally have investigated variance components associated with the rater, the task, and these components' interactions with the ratee (Clauser, Clyman, & Swanson, 1999). For example, Kraiger & Teachout (1990) investigated the generalizability of performance ratings (made for research purposes) of Air Force jet mechanics across several rating forms and rater sources (i.e., self, peer, and supervisors). Within-rater source analyses revealed that the most variance was attributed to an undifferentiated residual term, followed by the ratee term (i.e., object of measurement or true score variance), with the remaining terms (i.e., those associated with the items, forms, and all interactions) accounting for negligible amounts of variance. Note that these within-source analyses do not include a rater term because only one rater per source was available and, as such, the estimates in Kraiger & Teachout (1990) are analogous to intrarater, rather than interrater, reliabilities. Webb, Shavelson, Kim, & Chen (1989) investigated the generalizability of job performance ratings of Navy machinist mates across tasks (e.g., job knowledge tests, hands-on performance tests) and rater sources (i.e., self, peer, and supervisor). Ratings in this study were collected as part of a pilot testing phase of a data collection project. For all rater sources, results indicated that the person effect (i.e., object of measurement) accounted for the largest amount of variance and that the second largest effect was the residual variance component. For supervisor and self-ratings, only one rater was available and, as such, the effects attributable to the rater main effect and rater interaction effects could not be computed. However, multiple peer raters were available, and results indicated that the rater-by-ratee interaction effect accounted for a substantial amount (i.e., 23%) of the variance in observed ratings. Greguras & Robie (1998) analyzed the generalizability of developmental ratings made by supervisors, peers, and subordinates. Within source analyses revealed, across sources, that the largest amount of variance was attributable to an undifferentiated error term, followed by a combined rater main effect and rater-by-ratee interaction effect, followed by the ratee effect (i.e., true score variance). Results further indicated negligible amounts of variance being accounted for by the item effect or the item-by-person interaction effect.

G theory seems an ideal utility for examining multiple sources of error in job performance measurement. With this study, we applied G theory principles to the determination of measurement error in evaluations of job performance.

Method

Participants and Administration

Data were collected from white collar personnel in a financial company in which a multisource rating instrument was administered to rate job performance in 2005. Our original data set contained 170 ratees. These 170 ratees were rated on by 3 different managers. Each manager (rater) evaluated independently.

Analyses

We estimated generalizability (G) coefficients (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991), which have the advantage (over reliability coefficients based in classical test theory) of simultaneously recognizing multiple sources of measurement error. According to Shavelson and Webb (1991, p. 93), relative G coefficients, which we use, are "analogous to the reliability coefficient in classical theory, namely, true-score variance divided by expected observed score variance (i.e., an intraclass correlation coefficient)." Note that internal consistency, interrater, and test-retest reliability estimates can all be thought of as special cases of G coefficients that recognize a single source of measurement error in each case.

The first step in estimating G coefficients is to conduct a generalizability (G) study to estimate the variance component associated with each factor. Variance components are estimated using analysis of variance models. It is important to note that, as a general rule, the variance component for a factor is not equivalent to the estimated mean square for that factor because the expected value of the mean square is typically a combination of variance components. Thus, one must correctly define the expected mean square to accurately compute variance components (Shavelson & Webb, 1991).

G study designs can have either crossed or nested variables. In a completely crossed design, each variable is fully represented in all other variables. In the first of the G studies in the present investigation, persons were completely crossed with the task and rater facets, which were completely crossed with one another.

The second step is a decision (D) study, which uses the variance components estimated in the G study to decide what type of measurement design is necessary to achieve an acceptably high G coefficient. In such D studies, an investigator generalizes based on specified measurement procedures. One kind of decision is relative; a second type of decision is absolute. D studies pose "what if" questions by estimating reliability under different hypothetical scenarios, using variance components established in G studies to make these estimations. D studies establish both relative (δ) error variance and absolute (Δ) error variance (table 2). From these error variances, two kinds of coefficients are determined. Essentially an enhanced intraclass correlation coefficient, the generalizability coefficient ($E\varrho^2$,table 2), the ratio of persons variance, $\sigma^2(\tau)$, to itself plus relative error variance, $\sigma^2(\delta)$, is analogous to the reliability coefficient in classical test theory. A second coefficient, the index of dependability (Φ, table 2), is the ratio of persons variance to itself plus absolute error variance, $\sigma^2(\Delta)$. This latter coefficient, which takes into account all other sources of variance outside of the persons main effect, is not possible in classical test theory determinations of score reliability.

Measure

A multi-rater job performance instrument was developed by Yelboğa (2003) and were used in this study. The instrument comprised 4 scales (e.g., job knowledge). First scale containing 9 items, second and third scales containing 10 items respectively and fourth scale containing 3 items, for a total of 32 items. All items used a 5-point Likert-type scale ranging from 1 = insufficient to 5 = to an excellent. Scale names are managerial sufficiency, job information sufficiency, behavioral sufficiency and self developing sufficiency respectively.

Results

G Study Analyses

Table 1 displays the results of the G study. The table reports outcomes by effect (α); each effect's *df,* sums of squares, and mean square; and the estimated variance component for each effect, $\sigma^2(\alpha)$, obtained from mean squares via an algorithm Brennan (2001) illustrated.

Table 1.
Variance Components

Effect (α)	*df(α)*	*SS(α)*	*MS(α)*	$\sigma^2(\alpha)$
person (p)	175	1275,458	7,288	0,067
task (t)	31	80,410	2,594	0,001
rater (r)	2	37,587	18,793	0,003
pt (p x t)	5425	736,475	0,136	0,000
pr (p x r)	350	293,955	0,840	0,022
tr (t x r)	62	126,409	2,039	0,011
ptr (p x t x r)	19850	1638,049	0,151	0,151

Estimated variance components are presented for all main effects (*p, t, r*) and interactions among the three fully crossed variables. Accordingly Table 1, there was no variability in the person and task by person effects.

The most obvious variability was found in the p x t x r effects and person effects. These persons clearly were at different levels of performance; therefore, the person variability was anticipated. Rater variability, which ideally should have been zero, was quite low. The raters' rank ordering of the persons clearly varied. The three way interaction's variance component was quite high. This shows that there is high error variance.

D Study Analyses

Table 2 presents findings for the D study. The final two rows, Eϱ^2, the reliability like generalizability coefficient and Φ, the index of dependability are most important. Within the universe of generalizability established in this investigation, estimated reliability (Eϱ^2) was 0,89. For different hypothetical scenario listed in Table 2.

Table 2.
Random Effects p x t x r Decision Study Design for Performance Appraisal Data

$\sigma^2(\alpha)$	n_r	2	2	2	3	3	3	4	4	4
	n_t	28	32	36	28	32	36	28	32	36
$\sigma^2(\tau)$		0,067	0,067	0,067	0,067	0,067	0,067	0067	0,067	0,067
$\sigma^2(\delta)$		0,013	0,013	0,013	0,009	0,009	0,009	0,007	0,007	0,006
$\sigma^2(\Delta)$		0.015	0,015	0,014	0,010	0,010	0,010	0,008	0,007	0,007
Eϱ^2		0,83	0,84	0,84	0,88	0,89	0,89	0,91	0,91	0,91
Φ		0,82	0,82	0,82	0,87	0,87	0,88	0,90	0,90	0,90

Note: α: effect; n_r and $n_{t:}$: modifications of rater and task sample size respectively; τ:object of measurement (person); $\sigma^2(\delta)$: relative error variance; $\sigma^2(\Delta)$: absolute error variance; Eρ^2 : generalizability coefficient; Φ: index of dependability.

The more stringent index of dependability, Φ was 0,87. In the same way for different hypothetical scenario listed in Table 2. In contrast to (Eϱ^2), Φ increased steadily until about 4th hypothetical rater.

Discussion

Job performance measures play a crucial role in research and practice. Ratings (especially supervisory) are an important method of job performance measurement in organizations. Many decisions are made on the basis of ratings. As such, the reliability of ratings is an important concern in organizational science. Depending on the objective of the researcher, different reliability estimates need to be assessed.

Interrater correlations do not provide reasonable estimates of the reliability of job performance ratings and the reliability of ratings should not be evaluated using the parallel test model. For this reason Generalizability Theory can be used to assess the job performance ratings. Some researchers (e.g. Morris & Lobsenz, 2003; Murphy & DeShon, 2000a; Kieffer, 1999) argue that classical measurement theory is limited and that generalizability coefficients are appropriate. Accordingly Viswesvaran & Ones (2005), this argument is logically flawed. Both classical measurement theory and Generalizability Theory can be used to assess the different sources of error if the appropriate data are collected. The authors argue that both can yield the same information, provided that the appropriate data are collected and analyzed.

Generalizability analyses can be very useful for sorting out ratee effects, rater effects interactions, and so forth, and their implications for various generalizations one might want make about rating.

Generalizability Theory is considered a modern measurement theory, in contrast to the more classical approaches developed in the early 20th century. Generalizability Theory is especially well suited to evaluating ratings of human performance (Nunnally & Bernstein, 1994). In other words, Generalizability Theory seems an ideal utility for examining multiple sources of error in job performance measurement.

References

Atılgan, H., & Tezbaşaran, A. A. (2005). An investigation on consistency G and Phi coefficients obtained by generalizability theory alternative decision study for scenarios and actual cases. *Eurasian Journal of Educational Research*, *18*, 28-40.

Bergee, M. J. (2007). Performer, rater, occasion and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education , 55(4)*, 344-358.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: American College Testing.

Brennan, R. L. (2000). Performance assessment from the perspectives of generalizability theory. *Applied Psychological Measurement*, *24*, 339-353.

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

Campbell, J. P., Gasser, M. B., & Oswald, F. L. (1996). The substantive nature of job performance variability. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations*, 258-299. San Francisco: Jossey-Bass.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Clauser, B. E., Clyman, S. G., & Swanson, D. B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement*, *36*, 29-45.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology*, *74*, 130-135.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

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

Cronbach, L., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements.* New York: Wiley.

Greguras, G. J., & Robie, C. (1998). A new look at within source interrater reliability of 360 degree feedback ratings. *Journal of Applied Psychology*, *83*, 960-968.

Kan, A. (2007). Effect of using a scoring guide on essay scores: Generalizability theory. *Perceptual and Motor Skills*, *105*, 891-905

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

Kraiger, K., & Teachout, M. S. (1990). Generalizability theory as construct related evidence of the validity of job performance ratings. *Human Performance* , *3*, 19-35.

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

Morris, S. B., & Lobsenz, R. (2003). Evaluating personnel selection systems. In J. E. Edwards, J. C. Scott, & N. S. Raju (Eds.), *The human resources program-evaluation handbook* (pp. 109-129). Thousand Oaks, CA: Sage.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational and goal based perspectives*. Thousand Oaks, CA: Sage.

Murphy, K. R., & DeShon, R. (2000a). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873-900.

Murphy, K. R., & DeShon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology*, *53*, 913-924.

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

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

Schmidt, F. L., & Hunter, J. E. (1992). Causal modeling of processes determining job performance. *Current Directions in Psychological Science*, *1*, 89-92.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199-223.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*, 901-912.

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

Viswesvaran, C., & Ones, D. S. (2005). Job performance: Assessment issues in personnel selection. In A. Evers, N. Anderson, & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 354-375). United Kingdom: Blackwell.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (2002). The moderating influence of job performance dimensions on convergence of supervisory and peer ratings of job performance: uconfounding construct-level convergence and rating difficulty. *Journal of Applied Psychology*, *87(2)*, 345-354.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81(5)*, 557-574.

Webb, N. M., Shavelson, R. J., Kim, K. S., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinist mates. *Journal of Military Psychology*, *1*, 91-110.

Yelboğa, A. (2003). *The examination of the psychometric qualities of the scale developed for the performance evaluation in the management of human resources*, Unpublished master thesis. Ankara University.

Yelboğa, A. ve Tavşancıl, E. (2010). The examination of reliability according to classical test and generalizability on a job performance scale. *Educational Science: Theory & Practice*. 10(*3*), 1825-1854.