Clarifying the concepts of reliability, validity and generalizability

Maria Valaste[1] and Lauri Tarkkonen[2]

[1] University of Helsinki, Finland
e-mail: maria.valaste@helsinki.fi

[2] University of Helsinki, Finland
e-mail: lauri.tarkkonen@helsinki.fi

Abstract

The uncertainty of parameters comes from two sources: sampling and measuring the study units. The purpose of this study is to investigate three concepts: reliability, validity and generalizability. Reliability gives the accuracy of a measurement. Validity relates to the truthfulness of a measurement. Generalizability theory defines "reliability-like" coefficient, called generalizability coefficient, which frequently is associated with reliability. Conclusion of this study is, generalizability coefficient should actually be related to the validity.

1 Introduction

When estimating parameters from some data with statistical methods, it is important to understand the uncertainty of parameters. The uncertainty comes from two sources: sampling and measuring the study units. Often the data is a (random) sample from a population. The first error then comes from collecting the data and generalizing the results to a population level. Another source of error is present when measuring the study units. When assessing the quality of the collected and measured data set, we end up questions: Are we measuring the right thing? How accurate our measurements are? The former question leads us to the concept of *validity* which is the most important property of measurement. The latter question is related to the concept of *reliability*.
At the beginning of the 20th century the concept of correlation had been discussed among statisticians. For historical reasons there were two separate traditions in studying correlational relationships. The psychometric tradition was concerned with a Pearsonian correlational analysis. The experimental tradition, started by Fisher (1925), was more concentrated on analysis of variance. During the century there were occasional attempts to synthesize these two traditions (Stanley 1971).

In 1904 Charles Spearman introduced the reliability coefficient based on correlation. Cronbach's alpha (Cronbach 1951)—the most widely applied estimator of reliability—is essentially based on the work of Kuder and Richardson in the 1930s. Cronbach *et al.* (1963) developed a new approach, called Generalizability Theory. It is based on ANOVA models and it was developed to liberalize the restrictive assumptions of Classical Test Theory (CTT).

2 Review of the central concepts

2.1 Reliability and Validity

Let x be the observed score. The true score model of Classical Test Theory (CTT) is

$$x = \tau + \varepsilon$$

where τ is the (unknown) true score and ε is the (random) measurement error. Assume $\mathrm{E}(\varepsilon) = 0$, $\mathrm{cov}(\tau, \varepsilon) = 0$.

Definition 1 (Lord & Novick 1968) *Reliability is a ratio of the true score variance to the observed variance. It is denoted by*

$$\rho_{x\tau} = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\varepsilon^2}.$$

Definition 2 (Lord & Novick 1968) *The validity coefficient of a measurement x with respect to a second measurement y is defined as the absolute value of the correlation coefficient*

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

2.2 Generalizability Theory

Generalizability theory (Cronbach *et al.* 1963, 1972; Brennan 2001) investigates and desigs reliable observations. Unlike in CTT where each test score has a single true score, single reliability coefficient and belongs to one family of parallel observations, in generalizability theory the error can be due to multiple sources.

"Reliability-like" coefficient, called generalizability coefficient, is based on the stepped-up intraclass correlation coefficient. Proper variance components are estimated by using

an ANOVA framework. Universe score variance $\sigma^2(\tau)$ is the estimated variance across the objects of measurement in the sample ("like" CTT true variance). Relative error variance $\sigma^2(\gamma)$ is the difference between observed deviation score and universe deviation score ("like" CTT error variance).

Definition 3 *Generalizability coefficient is ratio of universe score variance to expected score variance. It is denoted by*

$$\mathrm{E}\rho^2 = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\gamma^2}$$

where $\mathrm{E}(\cdot)$ *is the expectation.*

Perhaps one of the reasons why generalizability is frequently associated with reliability is the similarity of the equations of the generalizability coefficient and the reliability.

2.3 Example

ANOVA is a special case of the linear model. Consider a linear model

$$y = \alpha x + \delta \tag{1}$$

where y is the response variable, x is the predictor, α is the intercept and δ is the model error. We will include the random measurement error in the linear model by using the true score model $x = \tau + \varepsilon$. Hence the linear model

$$y = \alpha(\tau + \varepsilon) + \delta \tag{2}$$

contains now both error terms ε and δ explicitly. Fig. 1 demonstrates the extended model (2). A straight line is fitted. We concentrate on a particular data point (x, y). Like in Eq. (2) the predictor x is partitioned into true score τ and measurement error ε. Note these two terms are not in the same dimension because $cov(\tau, \varepsilon) = 0$. In model (1) the only error is the prediction error $\hat{\delta}$. If measurement error is taken into account (model (2)) then the dashed line in Fig. 1 is bias compared to the model (1).

3 Conclusions

The equations of the reliability and the generalizability coefficient are misleadingly similar. Generalizability is frequently associated with reliability but actually it should

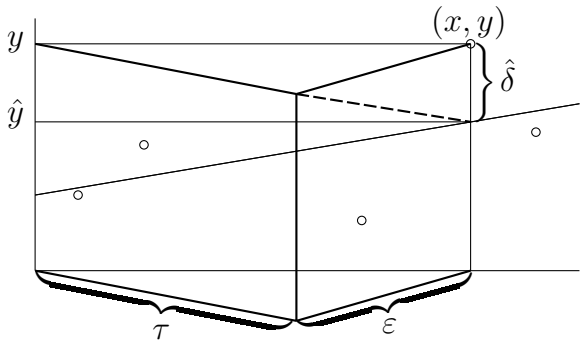

Figure 1: Measurement error and prediction error in a simple linear model (Valaste *et al.* in press).

be related to the validity (Valaste *et al.* in press). Generalizability theory is based on ANOVA. In ANOVA framework the x terms are fixed factors and the errors arise from the design. Thus the model does not contain random measurement error. Instead of reliability is based on pure random measurement error. Measurement framework approach (Tarkkonen 1987; Tarkkonen & Vehkalahti 2005; Vehkalahti *et al.* 2007) allows to model additional sources of error as multidimensional true scores and gives a general estimate of the measurement reliability.

References

Brennan, R. L. (2001). *Generalizability Theory.* Springer, New York.

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

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

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

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver and Boyd, London.

Kane, M. T. (1982). Sampling model for validity. *Applied Psychological Measurement*, **6**, 125–160.

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley, London.

Stanley, J. C. (1971). Reliability. In Thorndike, R. L., editor, *Educational measurement.* American Council on Education, Washington, D.C., second edition.

Tarkkonen, L. (1987). On Reliability of Composite Scales, no. 7 in Statistical Studies, Finnish Statistical Society, Helsinki, Finland.

Tarkkonen, L. & Vehkalahti, K. (2005). Measurement errors in multivariate measurement scales. *Journal of Multivariate Analysis*, **96**, 172–189.

Valaste, M., Vehkalahti, K. & Tarkkonen, L. (in press). Generalizability: Reliability or Validity?. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics*. Tokyo: Universal Academic Press.

Vehkalahti, K., Puntanen, S. & Tarkkonen, L. (2007). Effects of measurement errors in predictor selection of linear regression model. *Computational Statistics & Data Analysis*, **52**, 1183–1195.