PSYCHOLOGY AND COGNITIVE SCIENCES
Openventio PUBLISHERS

ISSN 2380-727X Open Journal http://dx.doi.org/10.17140/PCSOJ-2-110

Review

*Corresponding author
Rashid Saif Almehrizi, PhD
Associate Professor
Educational Measurement and Statistics
Department of Psychology
Assessment and Technical Support Unit, Director, College of Education
Sultan Qaboos University
P.O. Box: 32, PC: 123 Al-Khoudh
Sultanate of Oman
Tel. +968 2414 1613
E-mail: mehrzi@squ.edu.om

Volume 2 : Issue 1
Article Ref. #: 1000PCSOJ2110

Article History
Received: March 25th, 2016
Accepted: April 28th, 2016
Published: May 2nd, 2016

Citation
Almehzizi RS. Expected agreement coefficient for norm-referenced tests with classical test theory. *Psychol Cogn Sci Open J*. 2016; 2(1): 11-14. doi: 10.17140/PCSOJ-2-110

Expected Agreement Coefficient for Norm-Referenced Tests With Classical Test Theory

Rashid S. Almehrizi, PhD*

Department of Psychology, College of Education, Sultan Qaboos University, Sultanate of Oman

ABSTRACT

There are two types of agreement coefficients for psychological test scores: norm-referenced and criterion-referenced agreement coefficients. These coefficients were derived within the framework of generalizability theory that is known for its theoretical and practical complexities. Under the framework of classical test theory, the paper derived the norm-referenced agreement coefficient. This derivation was based on the assumption of randomly equivalent test forms. The resulted expected agreement coefficient was different from its counterpart in generalizability theory. However, the estimators of this norm-referenced agreement coefficient were equal under the two frameworks to coefficient alpha reliability.

KEYWORDS: Norm-referenced; Agreement coefficient; Coefficient alpha; Classical test theory.

INTRODUCTION

Psychological tests can follow two frameworks for interpretation and uses of their results: Norm-referenced and criterion-referenced. With norm-referenced interpretation and uses, investigator's interest focuses on the relative ordering of examinees with respect to the performance for the norm group which the examinee is associated.[1] In generalizability theory framework, relative error scores variance is defined as the expected squared difference between an examinee's observed deviation score (from examinee's true score) and the associated group's observed deviation score. On the other hand, criterion-referenced interpretation suggests that the investigator's interest focuses on absolute interpretations of scores and absolute error scores variance.[1-3] Relative error scores variance is defined as the expected squared difference between an examinee's observed deviation score and the examinee's true score.[4]

Since the first distinction between norm-referenced and criterion-referenced interpretations of test results, many researchers including Glaser and Nitko[5] and Popham and Husek[6] argued that reliability coefficients in the classical test theory are appropriate for norm-referenced tests. These coefficients (such as KR-20[7] and coefficient alpha[8]) depend on the relative standing of an examinee on a norm group.[9-10]

Kane and Brennan[11] introduced a very useful general agreement function that is used to summarize different existing agreement coefficients for different uses and interpretations of test scores. Using this general agreement function, Kane and Brennan[10] defined the norm-referenced expected agreement coefficient for norm-referenced tests (called generalizability coefficient) with generalizability theory framework. Using the general linear model for design (all examinees take same set of items) in generalizability theory for examinee's observed score on each item, , on a sample of items, Brennan and Kane derived the agreement coefficient for norm-referenced interpretation and showed that the estimator of this coefficient is equal to coefficient alpha developed by Cronbach.[8]

The concept of expected agreement and its derivation method is very useful to un-

ISSN 2380-727X Open Journal http://dx.doi.org/10.17140/PCSOJ-2-110

derstand test results and enhance its interpretation and uses.[12] It helps to differentiate examinees' error scores and accordingly examinees' true scores and test score reliability. Brennan[1] explained that norm-referenced agreement coefficient is associated with relative error scores whereas criterion-referenced agreement coefficient is associated with absolute error scores. The two types of error scores differ in their definition and implication when estimating and interpreting test score reliability.

The current application and utilization of the expected agreement is limited to generalizability theory frameworks. However, generalizability theory involves both theoretical and practical complexities.[1] It is based on mixture of concepts of variance components in analysis of variance and concepts of classical test theory. Similarly, the estimation of the expected agreement coefficient requires estimation of mean squares.[1,13]

On the other hand, classical test theory is based on simpler concepts and estimation methods that are appreciated by many practitioners.[4] The advantages and application of expected agreement are not yet introduced within classical test theory. One possible reason behind delaying usages of expected agreement coefficient in classical test theory might be traced to its conventional definition of equivalent test forms.

The paper introduced the expected agreement for norm-referenced interpretations of test scores within classical test theory framework. The paper presents the context and assumptions of randomly equivalent test forms that are necessary to develop the expected agreement coefficients. The paper derived the expected agreement/reliability coefficient for norm-referenced tests utilizing the general agreement coefficients pioneered by Kane and Brennan.[11] Moreover, the estimator of this expected agreement coefficient was outlined.

METHOD

Procedure

The paper used the procedure outlined by Kane and Brennan[11] for deriving the expected agreement between two randomly selected instances of a testing procedure. The procedure assumes that the instances or tests are randomly selected from a universe of possible instances, which support the assumption that the expected distribution of outcomes for the population is believed to be the same for each administration of the testing procedure. The agreement function, $a(S_{pi},S_{pj})$ defines the degree of agreement between any two scores of an examinee on two testing procedures, S_{pi} and S_{pj}. This agreement function can take any form as long it satisfies three conditions:

(1) $a(S_{pi},S_{pj})\geq 0$,

(2) $a(S_{pi},S_{pj})=a(S_{pj},S_{pi})$, and

(3) $a(S_{pi},S_{pi})+a(S_{pj},S_{pj})\geq 2a(S_{pi},S_{pj})$.

Two general agreement indices of instances for the testing procedure are defined: One is corrected for chance while the other is not corrected. The index of agreement which is not corrected for chance is:

$$\theta = \frac{A}{A_m}$$

The term A is the expected agreement given by $A=E_{p,I,J}a(S_{pI},S_{pJ})$, where the expectation is taken over the population of examinees and over pairs of tests that are independently sampled from the universe of tests and administered to the same population of examinees. The term A_m is the expected agreement between the instance of the testing procedure and itself, $A_m=E_{p,I}a(S_{pI},S_{pI})$, where A_m represents the maximum value of A. A is equal to A_m when each examinee in the population has the same score on every test. Kane and Brennan noted that A_m corrects the problem of the dependence of on the scale of $a(S_{pI},S_{pJ})$.

The index of agreement which is corrected for chance is

$$\theta_c = \frac{A-A_c}{A_m-A_c}$$

where term A_c quantifies the agreement between the two instances of the testing procedure that is due solely to chance. It is defined as the expected agreement between the score, S_{pi}, for a random selected examinee p on one test and the score, S_{qj}, for another independently sampled examinee q on an independently another sample test. That is.,

$$A_c=E_{p,q,I,J}\, a(S_{pI},S_{qJ})=E_{p,I}\, a(S_{pI})E_{q,J}\, a(S_{qJ}).$$

Also, Kane and Brennan[11] define the expected disagreement or loss as the difference between the maximum expected agreement and the expected agreement,

$$\sigma^2(\epsilon) = L = A_m - A.$$

This expected loss gives the error score variance associated with the expected agreement function.

RESULTS

In order to derive the expected agreement coefficient within the context of classical test theory, we need to first introduce the concept of randomly equivalent test forms instead of the classical equivalent test forms. Randomly equivalent test forms is evident when the test developer is able to build a very large or infinite number of different test forms from a large pool of items measuring the psychological construct. Hence, test forms of equal size are considered randomly equivalent forms if each is sampled randomly and independently from the large pool of items. These test forms are not expected to have equal mean scores nor equal variance. However, examinees error scores from these randomly equivalent test forms are expected

ISSN 2380-727X Open Journal http://dx.doi.org/10.17140/PCSOJ-2-110

to be uncorrelated. Moreover, it is assumed that any test form is administered to a large sample of examinees that are randomly selected from the population of examinees.

In order to derive the expected agreement/reliability of test scores on test form (say form X), we need to hypothesize that this test form and another hypothesized form (say form Y) are randomly equivalent test forms with different items but equal in terms of size (form X with I items and form Y with J items). Let us refer form X as a reference test form and the other test form (form Y) as a hypothesized test form. These two forms are then administered to the same sample of examinees of size N.

For a norm-referenced test where the decision is based on the relative position of examinees to their peer examinees, the agreement function is defined as the expected product of relative distance of the observed average scores ($\bar{X}_p$ and $\bar{Y}_p$) on two randomly equivalent test forms from the associated mean score for items on each test form (T_I and T_J) over all examinees.

$$A(r) = E_{P,I,J}(\bar{X}_p - T_I)(\bar{Y}_p - T_J) = \frac{1}{n^2} E_{P,I,J} \sum_i \sum_j (X_{pi} - T_i)(Y_{pj} - T_j)$$
$$= E_{P,I,J} \overline{(X_{pi} - T_i)(Y_{pj} - T_j)}$$

where the expectation is over infinite randomly equivalent test forms of *X* and *Y*; each with equal number of items from the domain, over infinite randomly independent samples of *N* examinees from the population, and $E_{P,I,J}\overline{(X_{pi}\text{-}T_i)(Y_{pj}\text{-}T_j)}$ is the expected mean pair wise covariance of items on *X* with items on *Y* with relative to their individual item mean scores.

For the reference test form *X*, $E_{P,I}\overline{(X_{pi} - T_i)(X_{pi'} - T_{i'})}$ represent the expected mean pair wise covariance of distinct items on *X* ($i \neq i'$) with relative to their mean scores. Similarly, let $E_{P,J}\overline{(Y_{pj} - T_j)(Y_{pj'} - T_{j'})}$ have similar definition for items on test form *Y*. Because of randomly equivalent test forms,

$$E_{P,I,J}\overline{(X_{pi} - T_i)(Y_{pj} - T_j)} = E_{P,I}\overline{(X_{pi} - T_i)(X_{pi'} - T_{i'})} = E_{P,J}\overline{(Y_{pj} - T_j)(Y_{pj'} - T_{j'})}$$

Hence, the expected agreement function, *A(r)* , becomes

$$A(r) = E_{P,I}\overline{(X_{pi} - T_i)(X_{pi'} - T_{i'})} = \frac{1}{n(n-1)} E_{P,I} \sum\sum_{i \neq i'} \overline{(X_{pi} - T_i)(X_{pi'} - T_{i'})}$$

By simple algebra, *A(r)* becomes,

$$A(r) = \frac{1}{n(n-1)} \left[n^2 E_{P,I}(\bar{X}_p - T_I)^2 - E_I \sum_i E_p (X_{pi} - T_i) \right],$$

where $T_I = E_I \sum_i T_i$.

This expected agreement function gives the true score variance for norm-referenced tests, $\sigma^2(T_r)$. The maximum expected agreement for norm-referenced testing is,

$$A_m(r) = E_{P,I}(\bar{X}_p - T_I)(\bar{X}_p - T_I) = E_{P,I}(\bar{X}_p - T_I)^2$$

The expected agreement for norm-referenced testing due to chance is,

$$A_c(r) = E_{P,Q,I,J}(\bar{X}_p - T_I)(\bar{Y}_q - T_J) = E_{P,I}(\bar{X}_p - T_I)\, E_{Q,J}(\bar{Y}_q - T_J)$$
$$= E_{P,I}\left(\frac{1}{n}\sum_i X_{pi} - \frac{1}{n}\sum_i T_i\right) E_{Q,J}\left(\frac{1}{n}\sum_j Y_{qj} - \frac{1}{n}\sum_j T_j\right) = 0,$$

because $E_p(X_{pi}) = T_i$ and $E_Q(Y_{qj}) = T_j$. Hence, the norm-referenced agreement coefficient is,

$$\theta(r) = \theta_c(r) = \frac{\frac{1}{n(n-1)} E_{P,I} \sum\sum_{i \neq i'} (X_p - T_i)(X_{pi'} - T_{i'})}{E_{P,I}(\bar{X}_p - T_I)^2}.$$

$$\text{or } \theta(r) = \theta_c(r) = 1 - \frac{\frac{1}{n(n-1)} \left[E_I \sum_i Ep(X_{pi} - T_i)^2 - nE_{P,I}(\bar{X}_p - T_I)^2 \right]}{E_{P,I}(\bar{X}_p - T_I)^2}.$$

This coefficient can be also written as,

$$\theta(r) = \theta_c(r) = \frac{n}{n-1}\left(1 - \frac{E_I \sum_i E_p (X_{pi} - T_i)^2}{n^2 E_{p,I}(\bar{X}_p - T_I)^2}\right).$$

This result suggests that the correction for chance agreement has also no effect on the norm-referenced agreement.

The expected loss associated with the norm-referenced agreement coefficient is,

$$L(r) = A_m(r) - A(r) = \frac{1}{n(n-1)} \left[E_I \sum_i E_P (X_{pi} - T_i) - nE_{P,I}(\bar{X}_p - T_I)^2 \right]$$
$$= \frac{1}{n(n-1)} E_{P,I} \sum_i ((X_{pi} - T_i) - (\bar{X}_p - T_I))^2$$

which equals the appropriate error score variance for norm-reference d tests, $\sigma^2(\epsilon_r)$.

This error score variance is similar to the relative error score variance identified by Brennan and Kane[2] using Generalizability theory. This quantifies the expected squared difference between each examinee's observed deviation score from the test average score and the deviation of an examinee's true score from the test average score on the domain of items.

ESTIMATION

The components of all expressions of the expected agreement/reliability coefficients have the form of expected value of some terms over different random sets of items from the domain of items and over different random samples of examinees from the population of examinees. The sample counterparts of these terms can be used to estimate these expected values.

The expected norm-referenced agreement/reliability coefficients can be estimated by collecting data from adminis-

ISSN 2380-727X Open Journal http://dx.doi.org/10.17140/PCSOJ-2-110

tering one test form of n items to a representative sample of N examinees. If we substitute ($\bar{X}_p$), T_i and T_I by their sample counterparts, $\bar{x}_p = \frac{1}{n}\sum_i x_{pi}$, $\bar{x}_i = \frac{1}{N}\sum_p x_{pi}$, and $\bar{x} = \frac{1}{n}\sum_i \bar{x}_i = \frac{1}{N}\sum_p \bar{x}_p$ respectively, the estimator of the expected agreement coefficient for norm-referenced test is,

$$\hat{\theta}(r) = \frac{\frac{1}{n(n-1)}\sum\sum_{i\neq i'}\hat{\sigma}_{ii'}}{\hat{\sigma}^2(\bar{x}_p)} = 1 - \frac{\frac{1}{n(n-1)}\left[\sum_i \hat{\sigma}^2(x_{pi}) - n\hat{\sigma}^2(\bar{x}_p)\right]}{\hat{\sigma}^2(\bar{x}_p)}$$

$$= \frac{n}{n-1}\left(1 - \frac{\sum_i \hat{\sigma}^2(x_{pi})}{n^2\hat{\sigma}^2(\bar{x}_p)}\right)$$

The associated loss is,

$$\hat{L}(r) = \frac{1}{n(n-1)}\left[\sum_i \hat{\sigma}^2(x_{pi}) - n\hat{\sigma}^2(\bar{x}_p)\right],$$

Which gives the estimator of the relative error score variance for norm-referenced test
In these equations,

$$\hat{\sigma}_{ii'} = \frac{1}{N\text{-}1}\sum_p (x_{pi} - \bar{x}_i)(x_{pi'} - \bar{x}_{i'}),$$

$$\hat{\sigma}^2(x_{pi}) = \frac{1}{N\text{-}1}\sum_p (x_{pi} - \bar{x}_i)^2,$$

$$\hat{\sigma}^2(\bar{x}_{pi}) = \frac{1}{N\text{-}1}\sum_p (\bar{x}_p - \bar{x})^2.$$

DISCUSSION AND CONCLUSION

The paper derived the expected agreement coefficient for norm-referenced tests using classical test theory framework under the assumption of randomly equivalent test forms as replacement of the conventional equivalent test forms. The estimators of the resulted coefficient proved itself to be equal to coefficient alpha for Cronbach[8] that was derived under different assumption of essentially tau-equivalent test form.

This result supports what Glaser and Nitko[5] and Popham and Husek[6] argued that reliability coefficients in the classical test theory such as coefficient alpha and KR-20 are appropriate for norm-referenced tests. The error scores associated with coefficient alpha is the relative error score variance that is defined as the difference between individual examinee's performance and the performance of his/her peers who took the test.

The estimation of the expected agreement coefficient for norm-referenced tests can use either unbiased or biased estimators of its terms. It can be easily showed that if the biased estimators of the terms in the above equations are used, they would give identical estimates of the expected agreement coefficient for norm-reference tests. However, the estimation of the error score variances and the true score variance, however, are affected by whether the unbiased or biased sample variances are used (The unbiased estimators are preferred).

REFERENCES

1. Brennan RL. Generalizability theory and classical test theory. *Applied Measurement in Education*. 2010; 24(1): 1-21. doi: 10.1080/08957347.2011.532417

2. Brennan RL, Kane MT. An index of dependability for mastery tests. *J Educ Meas.* 1977; 14(3): 277-289. doi: 10.1111/j.1745-3984.1977.tb00045.x

3. Brennan RL, Kane MT. Signal/noise ratios for domain-referenced tests. *Psychometrika*. 1977; 42(4): 609-625. doi: 10.1007/BF02295983

4. Gao X, Brennan R, Guo F. Modeling measurement facets and assessing generalizability in a large-scale writing assessment. GMAC Research Report. 2015.

5. Glaser R, Nitko AJ. Measurement in learning and instruction. In: Thorndike RL, ed. *Educational measurement*. Washington DC, USA: American Council on Education; 1971.

6. Popham WJ, Husek TR. Implications of criterion-referenced measurement. *J Educ Meas*. 1969; 6(1): 1-9. doi: 10.1111/j.1745-3984.1969.tb00654.x

7. Kuder GF, Richardson MW. The theory of the estimation of test reliability. *Psychometrika.* 1937; 2(3): 151-160. doi: 10.1007/BF02288391

8. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika.* 1951; 16(3): 297-334. doi: 10.1007/BF02310555

9. Almehrizi RS. Coefficient alpha and reliability of scale scores. *Applied Psychological Measurement*. 2013; 37(6): 438-459. doi: 10.1177/0146621613484983

10. Cronbach LJ, Shavelson RJ. My Current thoughts on coefficient alpha and successor procedures. *Educ Psychol Meas*. 2004; 64(3): 391-418. doi: 10.1177/0013164404266386

11. Kane MT, Brennan RL. Agreement coefficients as indices of dependability for criterion-referenced tests. *Applied Psychological Measurement*. 1980; 4(1): 105-126. doi: 10.1177/014662168000400111

12. Almehrizi R. Normalization of mean squared differences to measure agreement for continuous data. *Stat Methods Med Res*. 2013. doi: 10.1177/0962280213507506

13. AlKharusi H. Generalizability theory: An analysis of variance approach to measurement problems in educational assessment. *J Studies Educ*. 2012; 2(1): 184-196. doi: 10.5296/jse.v2i1.1227