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. 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. Relative error scores variance is defined as the expected squared difference between an examinee’s observed deviation score and the examinee’s true score.

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

The concept of expected agreement and its derivation method is very useful to un-
understand test results and enhance its interpretation and uses. It helps to differentiate examinees’ error scores and accordingly examinees’ true scores and test score reliability. Brennan 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. 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.

On the other hand, classical test theory is based on simpler concepts and estimation methods that are appreciated by many practitioners. 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 interpretations of test scores within classical test theory. Similarly, the estimation of the expected agreement coefficient requires estimation of mean squares.

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:

\[ \theta = \frac{A}{A_n} \]

The term \( A \) is the expected agreement given by

\[ A = E_{p,I} a(S_{pI},S_{pI}) \]

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 \( A_n \) is the expected agreement between the instance of the testing procedure and itself, \( A_n = E_{p,I} a(S_{pI},S_{pJ}) \), where \( A_n \) represents the maximum value of \( A \). \( A \) is equal to \( A_n \) when each examinee in the population has the same score on every test. Kane and Brennan noted that \( A_n \) corrects the problem of the dependence of \( A \) on the scale of \( a(S_{pI},S_{pJ}) \).

The index of agreement which is corrected for chance is

\[ \theta_c = \frac{A - A_c}{A_n - A_c} \]

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, \( S_{pn} \), for a random selected examinee \( p \) on one test and the score, \( S_{qn} \), for another independently sampled examinee \( q \) on an independently another sample test. That is,

\[ A_c = E_{p,q,I,J} a(S_{pn},S_{qn}) \]

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

\[ \sigma^2(c) = L = A_n - A. \]

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
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 randomly equivalent test forms with different items but equal in terms of size (form X with I items and form Y with J items). This expected agreement for norm-referenced testing due to chance is,  

$$A_c(r) = E_{p\mid j,p\mid j}(\bar{X}_p - T_1) (\bar{Y}_q - T_2) = E_{p\mid j}(\bar{X}_p - T_1) E_{q\mid j}(\bar{Y}_q - T_2)$$  

and  

$$= E_{p\mid j}\left(\frac{1}{n} \sum_{i=1}^{n} X_{pi} - \frac{1}{n} \sum_{i=1}^{n} T_i\right) E_{q\mid j}\left(\frac{1}{n} \sum_{j=1}^{n} Y_{pj} - \frac{1}{n} \sum_{j=1}^{n} T_j\right) = 0,$$

because $E_p(X_{pi}) = T_1$ and $E_q(Y_{pj}) = T_2$. Hence, the norm-referenced agreement coefficient is,  

$$\theta(r) = \theta(r) = 0$$  

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 $(\bar{X}_p, \bar{Y}_q)$ on two randomly equivalent test forms from the associated mean score for items on each test form $(T_1, T_2)$ over all examinees.  

$$A(r) = E_{p\mid j}(X_{pi} - T_1)(Y_{pj} - T_2) = E_{p\mid j}(X_{pi} - T_1)(Y_{pj} - T_2)$$  

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

$$A(r) = E_{p\mid j}(X_{pi} - T_1)(X_{pi} - T_2) = E_{p\mid j}(Y_{pj} - T_1)(Y_{pj} - T_2)$$  

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

$$A(r) = \frac{1}{n(n-1)} \left[ E_{p\mid j}(\bar{X}_p - T_1)^2 - E_{p\mid j}(X_{pi} - T_2)^2 \right],$$  

where $T_1 = E_{p\mid j}(X_{pi} - T_1)$.

This expected agreement function gives the true score variance for norm-referenced tests, $\sigma(T_1)$. The maximum expected agreement for norm-referenced testing is,  

$$A_u(r) = E_{p\mid j}(\bar{X}_p - T_1)(\bar{X}_p - T_1) = E_{p\mid j}(\bar{X}_p - T_1)^2$$  

The expected agreement for norm-referenced testing due to relative error score variance identified by Brennan and Kane 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-
This result supports what Glaser and Nitko essentially tau-equivalent test form. In these equations, the estimators of the expected agreement coefficient for norm-referenced test is,

\[
\hat{\theta}(r) = \frac{1}{n \sigma_0^2} \sum_{i=1}^{n} \frac{\sigma^2(x_{pi}) - \sigma^2(\bar{x}_p)}{\sigma^2(\bar{x}_p)} = \frac{1}{n \sigma_0^2} \sum_{i=1}^{n} \frac{\sigma^2(x_{pi}) - \sigma^2(\bar{x}_p)}{\sigma^2(\bar{x}_p)}
\]

The associated loss is,

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

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

\[
\hat{\sigma}_e = \frac{1}{N-1} \sum_{i=1}^{N} (x_{pi} - \bar{x}_i)(x_{pi} - \bar{x}_i),
\]

\[
\hat{\sigma}(x_{pi}) = \frac{1}{N-1} \sum_{i=1}^{N} (x_{pi} - \bar{x}_i)^2,
\]

\[
\hat{\sigma}_e^2(\bar{x}_p) = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{x}_p - \bar{x}_i)^2.
\]

**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 result coefficient proved itself to be equal to coefficient alpha for Cronbach that was derived under different assumption of essentially tau-equivalent test form.

This result supports what Glaser and Nitko and Popham and Husek 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**


