Reliability Estimation for Criterion-Referenced Test

Chen Xizhen

Putian Institute

Fujian, China

Email address: xizhenchen@163.com

1. Introduction

In a criterion-referenced test, one’ s interest for test is different from a norm-referenced test, so the term reliability in criterion-referenced test is different compared with that in norm-referenced test. Then those formulas for estimating reliability that is suitable for norm-referenced test would not be suitable for criterion-referenced test. It is necessary to seek formulas for estimating reliability that is suitable for criterion-referenced test. In this paper two issues of reliability estimation will be discussed.

2. Reliability estimation using test scores as domain scores

When a test is constructed by some items of a course, examinee’ s score in this test is often compared with their domain score in the course. For example, a teacher wants to know the mastery degree of each student in this course. For obvious practical reasons, however, it is impossible to arrange all vocabulary in a test, students can only be tested on a limited number of words. Then teacher intends to use each examinee’ s observed score as an estimate of the student’ s domain score, the proportion of the words that the student master. The examiner wants to know the accuracy of estimate when we regard the observed score as domain score of each student. From the point of view of generalizability theory, when the teacher administers the course test and intends to use the observed score of the test to estimate the domain score of student, the teacher is conducting a D-study in which examinees are crossed with items. The universe of generalization consists of all items in this course, and domain score of a student is a mean score of all items scores in the universe of generalization. The domain score of a student is an “ideal” score of goal of measurement (here is student) in given universe of generalization. Since area of domain score in generalizability theory is wider than that of true score in classical test theory, it can be imagined that reliability when we intend to decision student’ s observed score in a test to domain score in the course is less than to decision the student’ s observed score to true score in the test. Therefore, we use following formula:

$$\Phi = s^2 / [s^2(p) + s^2(\Delta)]$$

as an estimation of reliability, where $s^2(p)$ is domain score variance of measuring goal, $s^2(\Delta)$ is absolute error variance. These variances can be estimated as follow.

3. Reliability estimation using test scores as mastery classification

In mastery classification, the domain scores are divided into K mutually exclusive mastery categories defined by K-1 cut scores such as excellent, good, middle, pass and not pass, that needs 4 cut score. But the most commonly cited example has one cut score and two categories: master and nonmaster. After a test, we set a cut score, so that examinees scoring at or above this cut score are classified as category that who masters the course, and those scoring below the cut score are classified as category that who nonmasteres the course. Now the question is: since there are variant differences between examinee’ s scores in a test and examinee’ s domain scores in the course. Then is it consistency between mastery classifications based on examinee’ s observed scores and cut score of the test and that based on domain scores and cut score of the course? How high the degree of consistency is? This is reliability problem of mastery classifications. How to set an index to measure the degree of consistency, it is obviously a very meaning problem. In below, condition in which only one cut score is considered and in which more cut score can solved similarly.

To solve this problem, we must estimate the probability of misclassifying (that is to say examinee is true master but is misclassified as nonmaster or examinee is not master but is misclassified as master based on the test score). If this probability of misclassifying Q is estimated, then the accuracy of decision P=1-Q. For estimating Q, it is necessary to set a cut score π_0 of examinee' s domain scores, and to consider that if examinee' s domain score at or above this cut score then this examinee is classified as mastery category, and if examinee' s domain score below this cut scorethen that is classified as nonmastery category. Since there is error of measurement, therefore, it is not every one whose domain scores larger than π_0 to pass the test, and it is possible that examinee whose domain score less than π_0 passes the test. Then it is necessary to set two scores π_1 and π_2, π_1 indicates the highest domain score that is considered in the nonmastery region, and π_2 indicates the lowest domain score that is considered in the mastery region, $\pi_1<\pi_2$ The range between π_1 and π_2 is seen as indifference zone. We are unsure whether an examinee whose domain scores in the range should be considered as a master or a nonmaster. We are concerned about the probability that an examinee with a domain score equal to π_1 will be misclassified as a master andthe probability that an examinee with a domain score equal to π_2 will be misclassified as nonmaster. If two false distinguish probability are computed, then the sum is the probability of misclassifying. Since an examinee' s domain score is the proportion of the domain of items that an examinee can answer correctly. Thus an examinee' s domain score can be interpreted as the probability that the examinee answers a randomly chosen item correctly. It is to say the probability distribution of examinees with domain score π will answer correctly K items out of n randomly chosen items obeys binomial distribution, therefore, the binomial distribution model can be used to calculate the probability of the examinee with domain score π will answer correctly K items out of n randomly chosen items.

The methods of estimating decision consistency can be obtained by improved estimating formula Φ. We add to a factor $(m-l)^2$ in the numerator and denominator of Φ, in which l is cut score, m is the row-score mean (m and l are proportion-correct score for the sample), the formula is obtained as follow:

$$\Phi(l)=\frac{\hat{s}^2(p)+(m-l)^2}{\hat{s}^2(p)+(m-l)^2+\hat{s}^2(I)+\hat{s}^2/n_i}$$

when cut score l equals to raw-score mean m, Φ equals to $\Phi(l)$, when difference between l and m increases, it results in an increase in decision consistency, This result is the same as the analysis above.

References

Chen Xizhen, How to use formulas for estimating reliability correctly, Acta Psychologica Cinica, Vol.23 No.1.39-47.

Chen Xizhen, Reliability coefficient and correlation ratio between the observed scores and latent trait, Acta Psychologica Cinica, Vol.25. No.4.395-399.

Wang Songgui, Linear Model Theory and Its Application, Anhwei Education Publishing House, 1987.

Robert L. Brennan, Elements of Generalizability Theory, The American College Testing Program, 1983.