THE EFFECT OF STANDARD ERROR OF MEASUREMENT TO APPROPRIATENESS OF EDUCATIONAL EVALUATION

Şeref TAN
Celal Bayar University, Turkey

Introduction

One of the things that teachers must do is grading students' performance to determine academic achievement of students in a class. In order to grade students' achievement, teachers need to have some measurement results. In educational measures, such as academic achievement, personality, attitude, what we measure is usually some abstract variables. Because of that we cannot define perfectly what we measure. Even we define it in some ways, we cannot measure it without measurement errors. Just preparing or selecting a test, then applying it will not give us sufficiently reliable test scores. It is well known that when we measure an abstract variable, we are going to have some amount of errors in our scores. This amount of errors that we have in our scores should be taken account in grading students' academic achievements. It could be concluded that the measurement error of scores should be used in scoring procedures. In this article, first, the measurement error of classical test theory is introduced very briefly then scoring procedure using measurement error is explained.

Meaning of the Standard Error of Measurement (S_E)

In classical test theory, a measurement error of scores is an estimation of unity of error scores in random distribution of error scores. For instance, when the standard error of measurement is 2,00 points, it means that about 34% of students who were taken the exam, have 0 to 2 points of random error in their scores. Similarly, 34% of students have –2 to 0 points random error in their scores. According to normal distribution, about 13,5% of students have 2 to 4 points of random error in their scores, and because the normal distribution is a symmetric distribution 13,5% of students have –4 to –2 points of random error in their scores.

It is obvious that when measurement error is 2,00 points, about;

68% of students error score will lie from –2 to +2 points,

95% of students error score will lie from –4 to +4 points,

99% of students error score will lie from –6 to +6 points in unit normal distribution.

Standard error of measurement is an index for error scores of a test. As it is given in Lord (1968) that a person's observed score (X) is a combination of true score (T) and error score (E), X=T+E.

When some assumptions are met (such as $\rho_{TE} = \rho_{EE} = 0$), observed score variance equal to; sum of the true score variance and error score variance, we can have an equation to estimate the standard error of measurement by using statistics (**Lord & Novick, 1968**) given below;

$$s_E = s_X \sqrt{(1 - r_X)}$$

S_X : standard deviation of scores

r_X : estimated reliability of scores

S_E. estimated standard error of scores.

Note that this equation is calculated by using raw scores. For example, if you applied it to scores gained by a multiple-choice test, include 20 items, and items are scored by "0-1" scoring procedure, each item gets only 0 or 1 point. So, maximum score of this test is 20 points and minimum score is 0. All calculations should be done on raw score scale in estimating the standard error of measurement. Then, if wished the scores could be altered to a "0 – 100" scale.

Calculation of the Standard Error of Measurement

As it is introduced in above, we need to calculate the standard deviation and reliability of scores in order to calculate standard error of measurement. There are some ways to estimate reliability, which are given in all measurement books. So, reliability estimation methods are not covered in this article. In classical test theory and generalization theory, mostly used reliability procedure in achievement tests is KR-20, alpha, interrater and test-retest reliability methods.

For example, when the estimated reliability of scores is 0.84 and standard deviation of this scores is 2.00 points for a 20 item multiple choice test, scored by "0-1" scoring method, standard error of measurement is calculated as follows:

$s_E = 2\sqrt{(1 - 0.84)}$ = $s_E = 2\sqrt{0.16}$ = s_E = 2(0,40) = 0,80 points.

This error score is calculated for "0-20" point scale. When we want to alter it to "0-100" point scale; we get; (0,80*100)/20 = 4,00 points of standard error of measurement.

The distribution of standard error of scores could be showed in unit normal distribution, for this example;

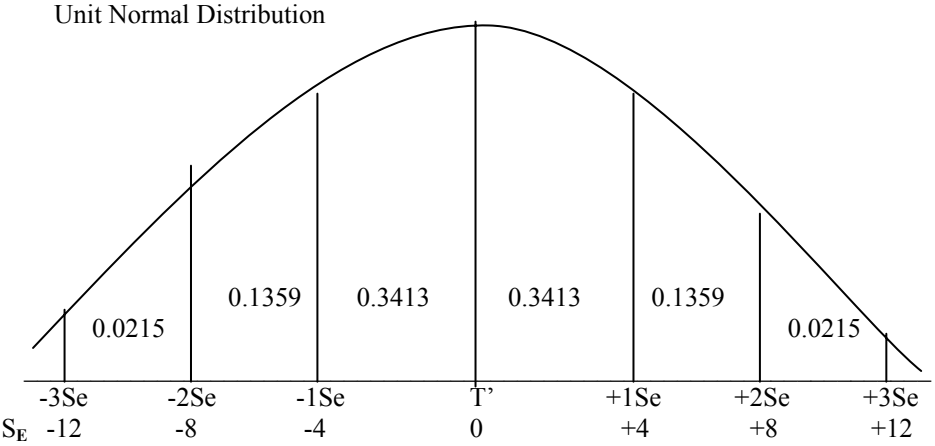

FİGURE 1. *Distribution of error scores when the standard error of measurement Score is 4 points*

The Standard Error of Measurement in Item Response Theory, SE($\hat{\theta}$)

In classical test theory, the standard error of measurement is assumed to be the same for all examinees; however, it is not assumed to be the same for all examines in item response theory. In item response theory, every estimated score has its own standard error of measurement. Standard error of measurement in item response theory is the standard deviation of $\hat{\theta}$, estimated ability level, and it is denoted as SE($\hat{\theta}$).

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

The term $I(\theta)$ is called the information function. As it is given in Hambleton & Swaminathan (1985) item response models are based on strong assumptions, which limit their applicability to many test data sets. An item response theory model requires very large subjects to meet the assumption of unidimensionality and local independence. This large sample may not be found in assessing students' achievement for a class. Item response theory (IRT) may be used, if data fits the model, for tests that are taken by large samples.

Educational Evaluation Procedures and Type I and Type II Errors

In educational assessment process we need to have two things to be able to make an assessment; one of them is having some measurement results and the second one is having some criteria. When we assess a student' achievement about a class we compare the measurement results with the criteria to give a decision. When we give a decision about achievement of a student, we can have two kinds of errors in our decisions.

Decision given

		Successful	Failure
True Condition	Successful	Correct Decision	Type I Error α
	Failure	Type II Error β	Correct Decision

FIGURE 2. Type I and Type II Errors in Educational decisions

As it is clear in Figure 2, when we decide "failure" for a successful student we have type I error. When we decide "successful" for a failure student we have type II error. These errors occur around the criterion score to be successful. A teacher can have both kinds of errors in his/her decisions in assessing academic achievement of students.

How We Can Reduce the Type I and Type II Errors

To have more accurate decisions when criteria were given us, we must have very low the standard error of measurement. One of the important things to have standard error of measurement as minimum is increasing the reliability and validity of the measurement results. Because all of the educational and psychological books have very deep explanation of increasing the reliability and validity of measurement results, it will not covered in this article. There are some other ways to reduce Type I error at the same

time to increase the Type II errors. In this article, decreasing the type I error considered more important than decreasing the type II error.

Transforming Scores to a Predetermined Scale: The distribution of raw scores could be altered to a distribution that has predetermined arithmetic mean and standard deviation. In order to do this, first each individual score need to be transformed to a z standard score by using equation given below:

$$Zi = \frac{Xi - \overline{X}}{Sx}$$

Then new scores with predetermined mean and standard deviation are calculated by using below equation (Tan, & at all. 2003):

$$Ai = \overline{X}_A + S_A (z_i)$$

Ai : i.th individual's new score in predetermined scale.

$\overline{X}_A$: predetermined arithmetic means of new scale.

S_A : predetermined standard deviation of new scale.

These kinds of transformation of scores are giving very much flexibility to teachers. It is very possible to misuse of this technique. In predetermination of the mean and standard deviation some appropriate norms should be used as much as teachers opinions.

Weighting the Scores by Their Reliability: Another way of reducing the measurement error is weighting the measurement results when multiple measurement results are available. Calculating weighted mean of scores by using reliability coefficients as weights lets us to have more reliable final scores to assess students' success. Therefore, we can have less type I (failing students who are actually successful) and type II (to pass students who are not truly successful) errors in our decisions.

A Scoring Method to Reduce Type I Error: Standard error of measurement could be added to scores of students. This is very important, especially when a score of a student is close to the criterion score, to not make wrong decision about the some of the successful students. Addition of the standard error of measurement to test scores could be done whether scores transformed to a predetermined scale or not. When we have just raw scores; we can simply add the standard error of measurement to the scores.

$$X_{SE} = X_i + S_e$$

This equation implies that new scores could be calculated simply by adding the standard error of measurement to all scores. This addition of the standard error of measurement to scores will increase the students test scores but will not improve the internal consistency of scores. However, using the standard error of measurement in scoring procedure as given above will decrease the type I error. In other word, this procedure decrease the probability of making wrong judgment about students whose score is very close to the criterion score to be judged as successful for a class.

The method that is given above has the same logic with setting up a confidence interval for the scores. With some amount of possibilities (68%, 95% or 99%) a confidence interval for students' score could be calculated then if confidence interval covers the criterion score, the student can be considered as successful.

Discussion

The standard error of measurement for an exam could be estimated very easily by hand calculation or using some statistical software such as SPSS (Statistical Packages for the Social Science). We do have standard error of measurement in measurement theory; however, we do not concern it in assessing students' achievement. It is clear that failure to consider measurement error resulted in ill-considered educational decisions.

Using the weighted means when we have multiple measurement results, adding-up the standard error of measurement (from classical test theory or item response theory) to the scores or setting-up the confidence interval for estimating students true scores may be helpful to assess students' achievement in more appropriate manner.

Using confidence interval or adding up the standard error of measurement to scores

Could be criticized in some points. Adding-up the standard error of measurement increases the type II error. Also, adding up the standard error of measurement to the scores results same thing by decreasing the criterion score itself. It is very important to know that when a student's confidence interval covers the some higher and lower scores from the criterion score, these students deserve special consideration.

Reducing the type I error (failing students who are actually successful) considered more important than type II error in applications. Not all the error in students' scores is random. There are some types of measurement errors come from teachers. Teachers as test makers can have mistakes when they writing questions, applying and scoring the tests. It is seen that there are some source of errors that are not random, and not depending on the examinees. These kinds of errors caused to increase type I error. That's way, decreasing the type I error considered more important in this article.

REFERENCES

AIRASIAN, P. (1994). **Classroom Assessment**, Second Edition, NY: McGraw-Hill.

BROWN, F. (1983). **Principles of Educational and Psychological Testing**, Third edition, NY: Holt Rinehart, Winston.

HAMBLETON, R.K. & SWAMINATHAN, H. (1985). **Item Response Theory Principles and Applications**, Boston: Kluwer.

HAMBLETON, R.K., SWAMINATHAN, H. & ROGERS H.J. (1991). **Fundamentals of Item Response Theory**, Newbury Park: Sage Pub.

HAYS, W. L. (1994). **Statistics**. Holt, Rinchart & Winston, Inc.

LINN, R.L. & GRONLUND, N.E. (1995). **Measurement and Assessment in Teaching** (7^{th} ed.). New Jersey: Prentice-Hall.

LORD, F.M. & NOVICK, M.R. (1968). **Statistical Theories of Mental Test Scores**. Reading MA: Addison-Wesley.

TAN, Ş. (2005). **Öğretimi Planlama ve Değerlendirme**. (7^{Th} Ed.). Ankara:PegemA Yayıncılık.