# Score Reliability and Language Testing

Westrick, Paul
Faculty of Languages and Cultures, Kyushu University

https://doi.org/10.15017/1312243

出版情報：英語英文学論叢. 54, pp.21-42, 2004. The English Language and Literature Society
バージョン：
権利関係：

# Score Reliability and Language Testing

Paul Westrick

## I. Introduction

In language testing literature one will find the vitally important term "reliability" mentioned when discussing norm-referenced tests (NRTs). NRTs are for comparing people, and they are often then used to make decisions about individuals, and as these decisions can have serious consequences for the individuals involved, they need to be based upon reliable information. Well-known examples of NRTs are the TOEFL and TOEIC examinations of English, tests in which the performances of individuals are compared with those of other individuals. NRTs differ from criterion-referenced tests (CRTs), also commonly called classroom tests, which deal with measuring the degree of mastery of the course material. Knowing about the reliability of test scores is paramount if one is to understand the descriptive statistics of published test reports or explain NRT scores to students. This knowledge is important because people often misunderstand what they read about reliability (and validity), and consequently they may reach wrong conclusions and make judgment errors. In this paper, reliability is defined, and variance and important points to remember about reliability are discussed. Next a few common methods of determining reliability estimates are presented, followed uses of reliability estimates and factors that influence reliability. Reliability standards and ways to improve score reliability are then covered.

## II. Explaining reliability

A common explanation of how reliability works and our personal concern with it in everyday life involves scales used to measure weight. When a person steps on a scale to measure his or her weight, steps off the scale, and then steps back on the scale, he or she expects the two measurements to be very close if not identical. If the person repeatedly weighs himself or herself on the scale, he or she will get a better idea of how consistent the measurements are and can determine how reliable the scale measurements are. If the measurements are widely different, the person will determine that the measurements are unreliable, meaningless, and unusable.

Whereas some instruments measure weight, some measure reaction time, and some measure language proficiency. In all these instances, there is a concern with the accuracy of the measurements, but in any measurement there is always an amount of error. There are two definitions of reliability, and the first one is concerned with error. This first definition of reliability is: "freedom from random error, i.e., how repeatable observations are (1) when different persons make the measurements, (2) with alternative instruments intended to measure the same thing, and (3) when incidental variation exists in the conditions of measurement" (Nunnally & Bernstein, 1994, p.213). For language testers, this means that a given group of students' scores should be consistent regardless of who administers the test and which form of the test is used, and when there is little difference between the conditions under which the test is administered. The second definition of reliability, "stability over time" (Nunnally & Bernstein, 1994, p. 243), is not often used in language testing because language learning dynamic, not static. Over time, students actively engaged in study tend to improve. With regard to language testing, the general idea is to determine how repeatable the observations are if the examinees take the same test or parallel forms of the test continually. With tests designed to measure physical reaction time, one would expect

to see stability over a period of years, but with language learners there will typically be great changes over a period of years.

With reliability defined, variance needs to be covered. In any human activity we want to measure, be it physical or mental, performance in a given activity will vary. A golfer will have good days and bad days on the same course within a single tournament, and a student will have good days and bad days when taking tests. This variance in performance pertains to the individual, and, generally speaking, at the individual level a high degree of variance is undesirable. There is also variance in performances of individuals within a group, but for testers this variance between individuals is desirable because it helps ensure high reliability estimates. In both cases, variance can be placed into two categories. There is meaningful variance, "variance related to the purpose of the test," which in language testing may cover different competences in the language, and error variance, more often referred to as measurement error, which is "variance due to other extraneous sources" (Brown, 1996, p. 186-187).

This error variance, or measurement error, can be subdivided further. Brown (1996, p.189) provides a checklist for potential sources of error variance, summarized here. There is variance due to environment, such as location, space, ventilation, noise, lighting, and weather ; variance due to administration procedures, such as directions, equipment, timing, and mechanics of testing ; variance attributable to examinees, such as health, fatigue, physical characteristics, motivation, emotion, memory, concentration, forgetfulness, impulsiveness, carelessness, testwiseness, comprehension of directions, guessing, task performance speed, and chance knowledge of item content ; variance due to scoring procedures, such as errors in scoring, subjectivity, evaluator bias, and evaluator idiosyncrasies ; and variance attributable to the test and test items, such as test booklet clarity, answer sheet format, particular sample of items, item types, number of items, item quality, and test security. Error variance is undesirable because it not related to the purpose of the test, and although it is unavoidable, all efforts should be made to minimize the influence of

these unwanted variables.

In addition to variance, there are a few other points about reliability to keep in mind. The first is that reliability is a characteristic of scores, not the test. As Thompson (2003a) emphasizes, the shorthand expression "the reliability of the test" is understood by knowledgeable, experienced testing experts as meaning the reliability of the test scores for the particular group in the study, but less knowledgeable people may believe reliability is a characteristic of the instrument. It is a problem because "authors of education and psychology journal articles" use the expressions "the reliability of the test" and "the test is reliable "(Thompson, 2003a, p.3). He believes that such shorthand expressions are "sloppy" and "cause confusion" because "we come unconsciously to ascribe literal truth to our shorthand" (Thompson, 2003b, p.94). Thompson repeatedly stresses that we should be clear by talking about the *reliability of scores,* not the reliability of tests. He notes others (Rowley, 1976 ; Crocker and Algina, 1986 ; and Gronlund and Linn, 1990 ; cited in Thompson, 2003b) who over the years have made similar statements that we must keep in mind that the reliability of scores is determined by the examinees. Thompson reminds testers and researchers that,

> The subjects themselves impact the reliability of scores, and thus it becomes an oxymoron to speak of "the reliability of the test" without considering to whom the test was administered, or other facets of the measurement protocol. Reliability is driven by variance—typically, greater score variance leads to greater score reliability, and so more *heterogeneous* samples often lead to more *variable* scores, and thus to higher reliability. Therefore, the same measure, when administered to more heterogeneous or to more homogeneous sets of subjects, will yield scores with differing reliability. (Thompson, 2003b, p.93 emphasis in original)

It should not be taken that Thompson's message cannot be found in language testing literature, as Brown (1996, p.206) expresses similar thoughts :

> Remember that reliability estimates are derived from the performances of a particular group of people. Hence, the estimate is linked to that group. In other words, the tester can only make claims about the reliability of a test with reference to a particular group of students ; or perhaps very cautiously, claims can be made about the probable level of reliability when the test is administered to a very similar group of students with about the same range of abilities.

Although the message is found in language testing literature, it is often simply forgotten by language teachers and people in other fields who give tests and conduct research. For this reason, Thompson has recently delivered his stern reminder.

In both their ways, Thompson and Brown lead us to a second point that language teachers need to remember when looking at a test-the reliability of scores from a single test administered to two different groups will not be alike, and they may be very different. Brown (1989) describes how a cloze test produced very different reliability coefficients with two different groups of students, but most language teachers do not have this thought in the forefronts of their minds when making judgments about NRTs. Although at one level language teachers instinctively know that different groups of students will perform differently on tests, they tend to forget or simply do not know how this effects reliability. If they are thinking of reliability when looking at a commercially produced test, they may look only at the reliability coefficient reported in the test manual and stop thinking about it from that point onward because the "experts" have already proven that "the test is reliable."i Unfortunately, the examinee populations for these general proficiency language tests are drawn from ESL populations or from different countries around the world, and they are far more heterogeneous than those found in typical EFL contexts. Consequently, the data collected by the makers of commercial tests are almost guaranteed not to look exactly like the data collected by individual institutions abroad using those tests. In Thompson's (2003a, p.5 emphasis in original) words, "It is important to evaluate score reliability in all studies, because it is the reliability of the data in hand in

a given study that will drive the results, and not the reliability of the scores described in the test manual."

The final point for language teachers to remember is that the reliability of the scores from a given test is important because of its relationship with score validity. In defining validity, Thompson concisely states that, "Score validity deals with the degree to which scores from a measurement measure the intended construct" (2003a, p.5). There can be no validity without reliability, for it is extremely difficult to measure a construct when the measurements (i.e., scores) are unreliable or of very low reliability. In an extreme case of having a reliability estimate of zero, Thompson (2003a, p.6) explains that,

> ...Perfectly unreliable scores measure nothing. If the scores purport to measure something/anything (e.g., intelligence, self-concept), and the scores measure nothing, the scores (and inferences from them) cannot be valid. Scores can't both measure nothing and measure something. The only time that perfectly unreliable scores could conceivably be valid is if someone was designing a test intended consistently to measure nothing.

Although perfectly unreliable scores are rare, low score reliability estimates are not rare, and when reliability estimates are low the soundness and trustworthiness of the scores are cast into doubt. For this reason, high score reliability is essential for score validity, but having high score reliability does not ensure score validity. A test with highly reliable scores may be reliably measuring something other than the intended construct. An English listening test, for example, may have a high reliability estimate, but it is not valid measure of English communication, as communication requires both receptive and productive skills.

## III. Determining reliability

Numerous methods of estimating reliability exist, but here only some of the more common methods will be covered. They are the test-retest reliability coefficient, the parallel form reliability coefficient, the split-

half coefficient, the Guttman split-half coefficient, the alpha coefficient (or Cronbach alpha, named after its creator), the Kuder-Richardson formula 20 (K-R20) reliability coefficient, and the Kuder-Richardson formula 21 (K-R21) reliability coefficient (Bachman, 1990 ; Brown, 1996 ; Carmines & Zeller, 1979 ; Hopkins, Stanley, & Hopkins, 1990 ; Nunnally & Bernstein, 1994). All of these can range from 0.00, perfectly unreliable, to 1.0, perfectly reliable.

The test-retest method of estimating reliability involves giving a test once, administering the same test later, and then determining the correlation between the test scores, something easily done in a computer spreadsheet. If the correlation is perfectly positive, 1.0, the reliability of the scores is a perfect 1.0. If the correlation is 0.50, the reliability of the scores is 0.50. As it involves using the exact same test, test-retest coefficients are sometimes referred to as stability coefficients (Hopkins, et al, 1990). The problem with the test-retest method is that students remember and students learn. If the second administration of the test comes too soon after the first administration, students will remember the items ; if it is administered much later, students may have learned more, and as individual students learn at different rates, the students' scores may be quite different from the first administration.

A way to avoid the problems associated with giving the same test twice is to give students parallel forms (also called equivalent forms or alternative forms) of the test. For two forms of a test to be considered equivalent, certain criteria must be met. As Bachman (1990, p.168) states, "...Parallel tests are two tests of the same ability that have the same means and variances and are equally correlated with other tests of that ability." Brown (1990, p.194) offers a similar definition, "From a strict statistical point of view, equivalent (or parallel) forms produce scores that have equal means, equal standard deviations, and equal correlations with some third measure of the same knowledge or skill." If these criteria are met, then one simply needs to determine the degree of correlation between the students' scores from the two forms of the test.

Though the test-retest method and the parallel forms method are

easy to understand, they are quite difficult from an administrative point of view, and it is a burden for students as well. For these reasons, statisticians developed internal consistency formulas to estimate score reliability from only one test administration. "Internal consistency is concerned with how consistent test takers' performances on the different parts of the test are consistent with each other (Bachman, 1990, p.172)." The split-half method, Cronbach alpha, K-R20 and K-R21 formulas are all methods of estimating the internal consistency reliability of test scores.

The split-half method is similar to the parallel form method, but it requires a little extra work. The single test is equally divided into two parts, typically odd numbered items and even numbered items, and the scores on the two halves are correlated. This correlation is for only half the test, so in order to get the full-test reliability estimate, the Spearman-Brown Prophesy formula is needed. The full-test reliability estimate is equal to the number of times the test length is to be increased (two) multiplied by the correlation coefficient for the two halves of the test, divided by the number of times the test length is to be increased (two) minus one, multiplied by the correlation coefficient plus one. The formula looks like this :

$$rxx' = n \times r / (n-1) r + 1$$

in which $rxx'$ is the full-test reliability ; $n$ is the number of times the test length is to be increased ; and $r$ is the correlation between the two halves of the test. For example, if the correlation coefficient for the two halves was 0.80, the full-test reliability estimate would 0.89. The formula would look like this :

$$rxx' = 2x0.80/(2-1)0.80+1 = 1.60/1.80 = 0.89$$

As with the parallel form method, it is assumed that the means and variance of both halves are equal. Additionally, the two halves must be independent of each other-performance on one half should not influence performance on the other half.

A way of getting around these problems is to use the Guttman split-half estimate as explained by Bachman (1990) :

$$rxx' = [k/k-1] \ [VAR_{h1} + VAR_{h2}/VAR_t]$$

in which *rxx'* is the reliability coefficient ; $k$ is the number of items ; $VAR_{h1}$ is the variance for one half of the test ; $VAR_{h2}$ is the variance for the other half of the test ; and VARt is the variance for the total test (variance is often represented as the standard deviation squared). In Brown (1996), this formula is presented as one form of the Cronbach alpha coefficient, with the exception that the variance for each half is explicitly referred to as the odd or even half. This formula is widely used because of its great flexibility.

The Cronbach alpha can also take the form of the K-R20 formula, when all the items are dichotomously-scored, i.e., correct or incorrect (Thompson, 2003a). The K-R20 is the more accurate of the two Kuder-Richardson formulas, and it is one of the most commonly reported reliability coefficients. The formula is as follows :

$$K-R20=[k/k-1]\ [1-(\Sigma IV/VAR]$$

in which $k$ is the number of items ; $\Sigma$ is the sum ; IV is the item variance ; and VAR is the variance for the total test.

To know the item variance of a test, the item facility (IF) of each item on the test needs to be known. The item facility is determined by dividing the number of examinees who answered the item correctly by the total number of examinees. If 12 examinees out of 20 answered the item correctly, the IF for that item is 0.60 (12 / 20). To determine the IV, this formula is used :

$$IV=IF(1-IF)$$

in which IV is the item variance, and IF is the item facility. The IV for each item on the test is calculated, and then they are added together to get the sum of item variance ((ΣIV). The K-R20 is considered quite accurate, but it is limited in that it can only be used with dichotomously scored answers, and all items must be equally weighted, making it less flexible than other forms of the Cronbach alpha.

Whereas the K-R20 is popular for its accuracy, the K-R21 formula is popular because of its ease of use. The number of items, mean (average score), and standard deviation are the only information needed to calculate the K-R21, and an assumption of the K-R21 is that all items are of

equally difficult. The K-R21 formula is as follows :

$$K-R21=[k/(k-1)]\ [1-(M(k-M)/(k-VAR)]$$

in which $k$ is the number of items ; $M$ is the mean (average) score ; and VAR is the variance for the entire test. The downside of using the K-R21 is that it is the most conservative reliability coefficient described here ; that is, it is more likely to underestimate the reliability of the test scores more than the other methods, but it is very useful when one wants to quickly get a preliminary reliability estimate in a research study. If the reliability coefficient produced from the K-R21 is quite high, the researcher can be confidant that the test scores are reliable and that the estimation of reliability will be higher using one of the other methods. It should be noted though, that as the level of the K-R21 reliability estimate rises, the smaller the gain observed when using another formula such as the K-R20.

## Ⅳ. Looking at individuals

Determining the reliability estimate for scores on a test is important, but it is not the final step in the analysis of the scores. The reliability estimate of test scores is very useful when looking at the score results of the group that generated the scores (Thorndike & Dinnel, 2001), but when looking at individuals more information is needed. With a reliability estimate one can determine two other useful pieces of information that are important when looking at individual examinees, the true score and the standard error of measurement.

In classical testing theory, the score an examinee receives on a test is the observed score, and the observed score consists of two parts that are independent of each other. These components are the error of measurement and the true score (or universe score). As explained by Hopkins et al(1990,p.118), "True scores represent the average score a person would obtain on an infinite number of parallel forms of a test, assuming that the person is not affected by taking the tests (i.e., assuming no practice effect or fatigue effect)." To estimate an examinee's true

score, one uses the following formula :

Estimated true score $= M + [rxx'(X - M)]$

in which $M$ is the mean score ; $rxx'$ is the reliability estimate ; and $X$ is the individual examinee's score. For example, if a student has an observed score of 60, and the mean is 50, and the reliability estimate is 0.90, the student's estimated true score would be 59. For a student with an observed score of 40, the estimated true score would be 41. In these cases, the amount of regression to the mean is very small because of the high reliability coefficient. With a lower reliability estimate, the regression increases. Changing the reliability estimate to 0.50, the first student's estimated true score would be 55, and the second student's estimated true score would be 45.

The estimated true score will always be closer to the mean than the observed score is unless the reliability estimate is a perfect 1.0, but as there will always be some degree of measurement error, this regression to the mean should be expected. Examinees who score above the mean on a test will tend to score above the mean on a retest, but most of their retest scores will be closer to the mean than observed score on the first test, and most examinees who scored below the mean will score below it again on a retest, but most of their retest scores will also be closer to the mean. Some examinees will move from one side of the mean to the other, particularly those who were close to the mean on the first test. The main point to be kept in mind is that lower the estimate of score reliability, the more teachers and administrators should expect to see regression to the mean due to the large amount of error in the observed scores.

Another useful figure that can be determined with the reliability estimate is the standard error of measurement (SEM). To get the SEM, get the square root of one minus the reliability estimate, and then multiply this number by the standard deviation of the test scores. For example, with a reliability estimate of 0.90 and a standard deviation of 9, after subtracting 0.90 from one, the next step would be getting the square root of 0.10, which is 0.3162277. Multiplying 0.3162277 by 9 results in an SEM of, after rounding, 2.85.

SEMs are to individuals what standard deviations are to groups. With a normal distribution, we can expect 68% of the people to fall within one standard deviation (plus or minus) of the mean. We can expect 95% of the people to fall within two standard deviations, and 99.7% to fall within three standard deviations. With the SEM, we can estimate that an examinee will score within one SEM of the observed score 68% of the time if continually retested, within two SEMs 95% of the time, and within three SEMs 99.7% of the time. As an example, with an SEM of five, one can estimate that if an examinee with an observed score of 50 is continually retested, 68% of the time that person would score between 45 and 55, 95% of the time between 40 and 60, and 99.7% of the time between 35 and 65.

These estimations of an examinee's true score with the SEM are referred to as confidence intervals. For the examinee above, the tester could be 68% confident that the examinee's true score falls between 45 and 55. With the SEM, one can establish band scores that cover the ranges of scores above and below the observed score (Bachman, 1990 ; Hughes, 2003), and these band scores are more informative than an examinee's observed score.

Too often decision makers put too much faith in the observed score on norm-referenced tests in which the performances of students on the test are compared. What is the difference between one person with an observed score of 81 and another with an 85? If the SEM on the test equals five, their 68% band scores are 75 to 87 and 79 to 91 respectively. Going to two confidence intervals, their 95% band scores are 69 to 93 and 73 to 97 respectively. Seeking 99.7% confidence, their band scores are 63 to 99 and 67 to 103 respectively. There are four points of difference at each of the extremes, but there are 33 points of overlap, four times as large as the combined differences found at the extremes. Making ranking decisions based on observed scores alone in this case would be rather unfair. Now, if the observed scores were 80 and 90, and the SEM equals 1.60, one could estimate with more than 99.7% confidence that there is no overlap. Decisions based on the test results here are much easier to explain and

defend. In borderline situations in which the picture is not so clear, as in the first case, and the fate of students or potential students hang in the balance, the responsible course of action would be to gather more information about the students' ability that was originally tested (Brown, 1996).

## V. Factors influencing reliability

As can be seen, the score reliability estimate is a very important tool for looking at how well a accurately a test separates and orders a given group of examinees, and it allows testers to estimate an individual examinee's true score and band scores, but one should be aware that a number of factors influence reliability. These factors are the item similarity, the number of items, the quality of each item, the number of examinees, the homogeneity or heterogeneity of the examinees, and the dispersion of scores (Bachman, 1990 ; Brown, 1996 ; Hughes, 2003, McNamara, 2000 ; Thompson, 2003)

The first factor to keep in mind is that, with everything else being equal, items that cover similar material tend to lead to higher score reliability than items that cover a wide variety of material. For example, for a group of students taking both a 40-item reading test and a 40-item reading/listening/grammar test, the score reliability estimate for their reading test will typically be higher than the score reliability estimate for the multi-skill test. For this reason, when different materials are covered, they are generally put into different sections or subtests, and each section or subtest has enough items needed for an adequate or high reliability estimate.

This idea of having a minimum number of items in a section or subtest leads to the second factor influencing score reliability-test length. Again, with everything else being equal, the reliability estimates for examinees' scores on longer tests are generally higher than those on short tests. With more items, there can be a greater range of scores, which typically increases score variance and the score reliability estimate.

The third item factor that influences score reliability is the quality of individual items on a test. As the purpose of an NRT is the dispersion of examinees' scores along the scale in a normal distribution so that the examinees can be accurately compared to one another, it is desirable to have items that discriminate between high scoring and low scoring examinees. When the percentage of the highest scoring examinees getting the item correct is higher than the percentage of the lowest scoring students getting the item correct, the item is discriminating between the high scorers and the low scorers in a positive way. Usually the top scoring 33% of the examinees are called the uppers, and the lowest 33% are called the lowers. The IFs for each item for each group is figured, then the IF lower is subtracted from the IF upper, giving the test developer (or researcher) the item discrimination (ID) of each item. According to Ebal' s (1979, cited in Brown, 1996) guidelines of ID quality, items with IDs above 0.40 are good items those with IDs between 0.20 and 0.39 are typically in need of improvement to varying degrees, and those with IDs below 0.19 should be rejected or revised. Ideally, a tester would want all items to have IDs above 0.40. Furthermore, the testers would want each item's IF at 0.50 (Brown, 1996), but in practice language testers should be reasonably content with IFs between 0.33 and 0.66. (McNamara, 2000).

Along with the three item factors discussed above, there are two examinee factors that influence the reliability of scores, the first of which concerns the number of examinees. Generally speaking, when everything else is equal, increasing the number of examinees will increase the score reliability. With any NRT, a normal (bell shaped) distribution is desired, and the minimum number of examinees required for the distribution of scores to approach that of a normal distribution is roughly 30, but with only 30 examinees one should not be surprised to see peaks and valleys in the actual distribution. If the number of examinees rises to 300 or 3,000, the distribution curve tends to become less jagged. With millions of examinees, as ETS has for their TOEFL and TOEIC tests, the picture becomes even better.

This global characteristic of the TOEFL and TOEIC examinees leads

to the second examinee factor that influences score reliability, the homogeneity or heterogeneity of the examinees. Generally speaking, when everything else is equal, the more heterogeneous the examinees are, the greater the variation in observed test scores. When examinees come from many countries and have a multitude of educational backgrounds, one should expect large variations in performance on the test, and large amounts of variance in test scores helps produce high reliability estimates. With homogenous groups, the results are often quite the opposite. Within a single country, and especially within a single university or college, the amount of variance can be quite small, and this has a negative influence on score reliability. With a lack of variance, there will a low score reliability estimate, which in turn effects the size of the SEM.

All the item and examinee factors covered above have an impact on the final factor that influences score reliability, the dispersion of scores. In the words of Nunnally and Bernstein (1994, p.261), "The size of the reliability coefficient is directly related to the standard deviation of obtained scores for any sample of subjects since the reliability coefficient is a correlation coefficient. ... It should be clear that the reliability coefficient will be larger in more variable samples." Ideally, testers want scores to be spread out in a normal distribution. If scores are bunched together in a narrow range, there will not be enough variance for a high reliability estimate. Additionally, the scores should be well centered on the scale. If the test is too easy or too difficult for the examinees, the scores will be negatively or positively skewed respectively. In sum, having well centered scores spread out over a wide range helps prevent low score reliability.

## Ⅵ. Standards of reliability

Standards of reliability differ depending on whether groups or individuals are being compared. As Nunnally and Bernstein (1994, p.265) explain,

> Group research is often concerned with the size of correlations

> and the mean differences among experimental treatments, for which a reliability of .80 is adequate. However, a great deal hinges on the exact test scores when decisions are made about individual. ...When selection standards are quite rigorous, decisions depend on very small score differences, and so it is difficult to accept any measurement error. We have noted that the standard error of measurement is almost one-third as large as the overall standard deviation of test scores when the reliability is .90. If important decisions are made with respect to specific test scores, a reliability of .90 is the bare minimum, and a reliability of .95 should be considered the desirable standard.

Hopkins et al. (1990, p.136) similarly state, "Standardized tests that such as those used for college admissions or special education placement should have reliability coefficients of at least .90 ..." One can also find references to standards of reliability in works specifically dedicated to language testing. McNamara (2000, p.62) writes, "We normally look for reliabilities on comprehension tests, or on tests of grammar or vocabulary, of 0.9 or better." McNamara does not discuss one would see with other types of tests, but according to Lado (1961, cited in Hughes, 2003), while good tests of vocabulary, structure and reading typically have reliability estimates between 0.90 and 0.99, reliability estimates for listening tests usually fall between 0.80 and 0.89, and those for oral tests between 0.70 and 0.79.

If these different types of tests are subtests within a larger, broader language test, the score reliability of the entire test can be quite high. Of course, the degree of reliability desired depends much on the decisions being made based on the test scores. As Hughes (2003, p.39) states, "The more important the decisions, the greater reliability we must demand : if we are to refuse someone the opportunity to study overseas because of their score on a language test, then we have to be pretty sure that their score would not have been much different if they had taken the test a day or two earlier or later." High stakes tests such as admissions tests and placement tests play a part in determining the future of each examinee, and it is the responsibility of the decision makers to ensure that the score

reliability estimates from those tests are high.

When the score reliability of an NRT is lower than desired, testers can do two things to improve the reliability estimates. The first is to conduct item analysis. As noted earlier, items with IDs under 0.40 generally need to be looked at more closely to see if they can be improved, and those under 0.19 need to be seriously revised or discarded completely. When replacing discarded items, a good idea is to pilot twice as many items as needed with students that are similar to those who will take the revised test. Some of the new items will not discriminate very well, and therefore will not be good replacements, but by piloting twice as many as needed one should be able to choose from an adequate number of new items that do discriminate well.

Another way to increase the score reliability is to lengthen the test. For example, a 30-item grammar test had a score reliability coefficient of 0.80, but the test developer wants a reliability estimate of at least 0.90. Here the Spearman-Brown Prophesy formula is useful.

$$rxx' = n \times r / (n-1)(r) + 1$$

in which $rxx'$ is the reliability estimate for the revised-test, $n$ is the number of times the test length is to be increased ; and $r$ is the reliability coefficient for the original test. By increasing the test by half, the test developer gets the following estimate :

$$rxx' = 1.5 \times 0.80 / (1.5-1)(0.80) + 1 = .86$$

By doubling the test length, the test developer gets this estimate :

$$rxx' = 2 \times 0.80 / (2\text{-}1)(0.80) + 1 = .89$$

As can be seen, creating twice as many items, the developer still does not reach his or her objective, 0.90. For this reason, it is often easier for the test developer (and future examinees!) if item analysis is conducted and individual items are improved rather than increasing the size of the test.

## Ⅶ. Examining test results

Having an understanding of score reliability and the standard error of measurement allows one to look at the descriptive statistics from a test

and decide whether the scores are useful or not, and the descriptive statistics of a test administration are presented here as an example. In a small study, The Quick Placement Test-Paper and Pen Test (2001), a test of English reading, vocabulary and grammar skills that is produced by the University of Cambridge Local Examinations Syndicate and marketed as a placement test, was given to 155 students in a Japanese university. The students were in two groups, and each group took two versions of this test back to back on the same day.

In Table 1, the descriptive statistics from this study are shown. A look at the means and medians on both forms of the test, along with the high and low scores for each test administration, suggest that the distribution is fairly well centered, and skewness is not much of a problem, but the narrow ranges of scores for each group on both tests are very narrow in relation to the entire scale. This lack of distribution has a negative impact on the reliability estimates (Cronbach alpha), which range from 0. 37 to 0.69, nowhere near 0.90. The correlations between Test 1 and Test 2 for both groups provide another reliability estimate, the parallel forms estimate, and these are also quite low at 0.41 and 0.52. Combining the test results for both forms of the test to get the correlation for all 155 students generated a correlation coefficient of only 0.43.

Using the Cronbach alpha reliability estimates for each test with each group, the SEMs were calculated, and by using the SEMs to convert observed scores to band scores, one can see that there is a large amount of overlap among the students. Only the people at the extremes in the first test administration can be safely identified as being among the top or bottom half of the group. In the second administration, one can see regression to the mean in both groups. Though the ranges of scores changed very little (Group 1) or not at all (Group 2), more scores were closer to the mean, and as score variance declined so did score reliability. Looking at Group 1, Test 2, going out three confidence intervals from the mean allows one to estimate with 99.7% confidence that a student with an observed score of 38 has a true score somewhere between 29 and 47, which actually exceeds the observed scores. This student may be among the

strongest or weakest students in the group in regard to reading, vocabulary and grammar.

Table 1.

| | Group 1 Test 1 | Group 1 Test 2 | Group 1 T 1 & 2 | Group 2 Test 2 | Group 2 Test 1 | Group 2 T 1 & 2 | G 1 & 2 Test 1 | G 1 & 2 Test 2 |
|---|---|---|---|---|---|---|---|---|
| *N* | 81.00 | 81.00 | 81.00 | 74.00 | 74.00 | 74.00 | 155.00 | 155.00 |
| *k* | 60.00 | 60.00 | 120.00 | 60.00 | 60.00 | 120.00 | 60.00 | 60.00 |
| Mean | 35.36 | 37.80 | 73.16 | 36.08 | 36.18 | 72.26 | 35.75 | 36.98 |
| Median | 35.00 | 37.00 | 72.00 | 36.00 | 37.00 | 71.50 | 36.00 | 37.00 |
| High | 44.00 | 47.00 | 90.00 | 47.00 | 46.00 | 86.00 | 46.00 | 47.00 |
| Low | 25.00 | 30.00 | 59.00 | 28.00 | 27.00 | 55.00 | 25.00 | 28.00 |
| *S* | 4.64 | 3.91 | 7.18 | 4.42 | 4.15 | 7.47 | 4.42 | 4.24 |
| Cronbach a | 0.69 | 0.37 | 0.70 | 0.59 | 0.48 | 0.65 | 0.60 | 0.51 |
| SEM | 2.57 | 3.11 | 3.93 | 2.85 | 2.99 | 4.44 | 2.78 | 2.98 |
| Correlation | | 0.41 | | | 0.52 | | | 0.43 |

This uncertainty about students at the middle of the distributions is supported by the observed high and low scores. By combining Tests 1 and 2 into one larger test, the combined scores allow one to infer useful information about the performances of the highest and lowest scorers on the individual tests. Looking at the high scores for Group 2 on Test 1 (47) and Test 2 (46), and then looking at the high combined score (86) one can tell that some students had scores bouncing around within broad ranges. The student who scored 47 on Test 2 could have scored only 39 at best on Test 1, and the student who scored 46 on Test 2 could have score only 40 at best on Test 2. Are these students among the elite in their group or are they just slightly above average? Looking at the low scores in Group 1 tells us that the student who scored 25 on Test 1 had to have scored at least 34 on Test 2, an improvement of nine points. This is more than the roughly eight points, plus or minus, estimated by going out three confidence intervals using the SEM from Test 1. Is this student the weakest in the group or just slightly below average? Moving away from the high and low scorers, it would be unsurprising to find other students who, if they were ranked by percentiles, moved from, say, the 20th percentile on one

test to the 80th percentile on the other test.

In sum, making placement decisions among these students based on these observed scores alone would be unwise. More information about the students needs to be gathered if one wants to be accurate. This does not necessarily mean that the test is bad—it just is not a very useful tool for separating these particular students. If these students were added to a larger group of students from around the world, the descriptive statistics would almost assuredly be much better, and dividing this larger body of students into groups based on those scores would be more defensible.

## Ⅷ. Conclusion

In this paper, the importance of understanding score reliability for norm-referenced tests has been briefly discussed. Score reliability matters because the decisions made affect the futures of the examinees. Language teachers need to be aware of how score reliability estimates are calculated, and how the standard error of measure can be used to interpret test scores. Teachers also need to be aware of various factors influencing score reliability, standards of reliability, and how score reliability can be improved. The descriptive statistics provided as an example above demonstrated that when the distribution of scores is poor, the score reliability estimates for a group will be low, and the standard error of measurement will be so large that it is impossible to accurately compare the performances of the examinees. Norm-referenced tests designed for the global market are blunt tools that work fairly well when separating masses of people from around the world with various educational backgrounds, but they tend to be inappropriate for use with homogenous groups. In designing their own norm-referenced tests, teachers need to ensure that their test items discriminate well between the higher performing and lower performing examinees so that scores are well distributed along the scale. Developing tests that match well with the students to be tested is hard work, and it involves piloting and revising, but when the examinees' futures are to be determined by their scores on high-stakes

tests, test developers must ensure that the reliability of the test scores is as high as possible.

## References

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford : Oxford University Press.

Brown, J.D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly,* (1), 65-83.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ : Prentice Hall Regents.

Carmines, E. & Zeller, R. (1979). *Reliability and validity assessment.* Newbury Park, CA : Sage Publications.

Hopkins, K., Stanley, J. & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ : Prentice Hall.

Hughes, H. (2003). *Testing for language teachers* (2nd ed.). Cambridge : Cambridge University Press.

McNamara, T. (2000). *Language testing.* Oxford : Oxford University Press.

Nunnally, J. & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York : McGraw-Hill, Inc.

Thompson, B. (2003a). "Understanding reliability and coefficient alpha, really." In B. Thompson (Ed.), *Score reliability : Contemporary thinking on reliability issues* (pp. 3-23). Thousand Oaks, CA : Sage Publications.

Thompson, B. (2003b). "Guidelines for authors reporting score reliability estimates." In B. Thompson (Ed.), *Score reliability : Contemporary thinking on reliability issues* (pp. 91-101). Thousand Oaks, CA : Sage Publications.

Thorndike, R. & Dinnel, D. (2001). *Basic statistics for the behavioral sciences.* Upper Saddle River, NJ : Merrill Prentice Hall.

University of Cambridge : Local Examinations Syndicate, (2001). *Quick*

University of Cambridge : Local Examinations Syndicate, (2001). *Quick placement test : Paper and pen test.* Oxford : Oxford University Press.

---

[i] For example, the *Test of English as a Foreign Language* has a reputation for high reliability, and as Brown (1996 : 231) points out, the *TOEFL Test and Score Manual : 1992-1993* (Educational Testing Service, 1992) reports that the reliability coefficient for the entire test was *.95* with an SEM of 14.1.