HOW Journal
Volume 27, Number 2, pages 135-155
https://doi.org/10.19183/how.27.2.541

Statistics for Classroom Language Assessment: Using Numbers Meaningfully

Estadística para la evaluación en el aula de idiomas: el uso significativo de los números

Frank Giraldo[1]

Abstract

Large-scale language testing uses statistical information to account for the quality of an assessment system. In this reflection article, I explain how basic statistics can be used meaningfully in the context of classroom language assessment. The paper explores a series of statistical calculations that can be used to examine test scores and assessment decisions in the language classroom. Therefore, interpretations for criterion-referenced assessment underlie the paper. Finally, I discuss limitations and include recommendations for teachers to use statistics.

Keywords: criterion-referenced testing, language testing, language assessment, score interpretation, statistics

Resumen

La evaluación de lenguas estandarizada utiliza datos estadísticos para evaluar la calidad de un sistema de evaluación. En este artículo de reflexión, explico cómo se puede usar la estadística de manera significativa en la evaluación en el aula de idiomas. El artículo explora una serie de cálculos estadísticos que pueden usarse para estudiar las notas y decisiones provenientes de instrumentos de evaluación en la clase de idiomas. Por ello, la evaluación criterial es la que utilizo para las in-

[1] Frank Giraldo holds an MA in English Didactics from Universidad de Caldas and an MA in TESL from the University of Illinois at Urbana-Champaign in the USA. He works for the foreign languages department of Universidad de Caldas. His main research interests are language assessment literacy and teachers' professional development.
frank.giraldo@ucaldas.edu.co
ORCID: https://orcid.org/0000-0001-5221-8245

Received: February 12th, 2020. Accepted: July 27th, 2020.

terpretaciones en el artículo. Finalmente, discuto unas limitaciones, y hago recomendaciones para que los docentes de idiomas puedan usar la estadística.

Palabras claves: estadística, evaluación criterial, evaluación de lenguas, interpretación de puntajes

Introduction

Numbers have power in language testing. Language tests such as TOEFL iBT and IELTS Academic yield scores that are used to make life-impacting decisions about people. These decisions should be sound given correct interpretations of test scores, and certainly a fundamental consideration for interpreting scores is that they represent the state of language ability as the main construct about which tests provide information (Chapelle, 2012; Fulcher & Davidson, 2007). Thus, scores from a language assessment can be considered a bridge between language ability as a construct and decisions in educational and other contexts.

To produce numbers, language testing professionals utilize basic and advanced statistical calculations through which the quality of assessments is scrutinized. For example, correlation coefficients are calculated to find out whether and to what extent scores from two different tests seem to be representative of similar constructs: Generally speaking, a correlation coefficient of 0.89 between two sets of scores can mean good news for test developers. However, in a statistical calculation known as Item Difficulty (ID), a set of items with ID levels between 0.80 and 0.95 may be bad news as these items are overly easy, even for students with a low level of proficiency. Thus, numbers in language testing are full of meaning.

Given that advanced calculations are fundamental to interpret scores or to evaluate large-scale language testing –it is in fact a core condition–, there is a belief that language teachers tend to fear or dread statistics and mathematical calculations in general (Brown, 2013; Fulcher, 2012). In fact, in Fulcher's (2012) study, language teachers from various cultural contexts reported that they need statistics explained conceptually rather than through mere calculations: They seem to want the meaning around the numbers.

When it comes to educating teachers in language testing and assessment, scholars have diligently answered this call. Publications on language testing have evolved and are now more teacher-friendly and practical in nature (Malone, 2017). However, statistics are, in my opinion, still presented without context and in rather abstract terms. To paraphrase the teachers in Fulcher (2012): Numbers are presented as numbers but not much is explained regarding their possible meanings. When interpretations are presented (for example, see Brown, 2011), they are limited to what the data present in a table, with limited allusion to classroom purposes for assessment. This may seem sensible, because these textbooks are written for a wide audience, so perhaps standard procedures suffice to explain statistics.

Interestingly, language teachers inevitably deal with numbers that should account for language learning. Scores in educational contexts are means by which teachers and other stakeholders can be informed about whether and to what extent students have learned course contents and/or achieved learning objectives (Carr, 2011). Authors such as Inbar-Lourie (2012) and Popham (2009) have suggested that teachers have at least a basic understanding of statistics so that they can be in a better position to evaluate assessment instruments and/or the decisions that are based on scores.

Against this backdrop, the purpose of this paper is to explain foundational statistics with an emphasis on context-specific interpretations for the language classroom. To provide context for this journal's readership, I will use English language education in Colombia (high school and university) as a point of reference. However, the statistical calculations and interpretations in this paper may be relevant in other contexts where teachers are tasked with evaluating their assessments.

I start by providing a general framework for the aforementioned context in Colombia, along with some assumptions that ideally should be met before doing statistics. Then, I illustrate the use of descriptive statistics (frequencies, percentages, distributions; mode, median, mean, and standard deviation) and possible related interpretations. Further, I explain what I call evaluative statistics, used for examining test items, tasks, and decisions. In evaluative statistics I include possible contextual interpretations; the evaluative statistics in this paper include item facility, difference index, B-index; agreement coefficient and kappa coefficient. I end the paper with limitations, conclusions, and recommendations.

Context and Theoretical Framework

Context

Although the overall reflection and explanations in this paper can be geared towards language teachers in general, I refer to the assessment system that is common in the Colombian educational context, specifically English language teaching in elementary, high school, and universities. In this system, scores are commonly called grades (*notas* in Spanish), and the scale that is used to assess students goes from 0.0 to 5.0. Generally, students pass a task or a course with a grade of 3.0, which translates to students having developed or learned 60% of task/course skills, contents, or objectives. This information may be different according to specific institutional policies, but the aforementioned scores represent the trend in Colombia.

Specifically, in elementary schools and high schools, the English language curriculum is driven by state-mandated standards (Ministerio de Educación Nacional de Colombia, 2016).

Such standards establish communicative competence as the major construct for language teaching, learning, and assessment. This construct is operationalized through the skills of listening, reading, writing, and speaking (monologues and conversation).

Largely, universities in Colombia follow a general English language curriculum based on communicative skills, while others drive learning through ESP syllabi; these universities are not expected to follow the aforementioned standards. Notwithstanding these specifics, their scale and passing grade are generally as just explained.

Theoretical Framework

Scores from language assessments are interpreted differently depending on whether they are norm-referenced or criterion-referenced. In norm-referenced testing, a person's score is compared to other people's scores and their relative standing, i.e., from lowest to highest performance (Douglas, 2014). Tests such as TOEFL and IELTS are norm-referenced. On the other hand, criterion-referenced testing examines an individual's score against a criterion. For example, the criterion can be a passing grade (e.g. 3.0), or a percentage (70% of course skills), or whether they can or cannot do something in the target language. In this case, decisions are not relative but absolute, which means a person's language performance is not compared to that of others (Fulcher, 2010).

I will discuss scores in this paper mostly from a criterion-referenced perspective, as it is the one with which most teachers may be engaged. For this, educational purposes for language assessment need to be considered. To begin, one of the purposes for classroom language assessment is to diagnose students' constructs before a course starts. The idea is to find out what students can and cannot do in the language, so, appropriate instructional adaptations are devised (Hughes, 2010). Another purpose for classroom language assessment is to analyze progress. In progress assessment, teachers are interested in finding out how students are (not) learning the specified curriculum objectives; with the data from progress assessment, teachers make other instructional decisions that positively impact student learning (Fulcher, 2010). Finally, achievement assessment "summarizes" what students have learned during a course. This type of assessment generally leads to a score which tells stakeholders (teachers, students, parents, etc.) to what extent students achieved the criterion for a unit or course. Through achievement assessment, teachers report how much of the construct (e.g. communicative competence, speaking, or others) students learned (Brown, 2011). I will use these three purposes –diagnostic, progress, and achievement– to explain and discuss the interpretations and related implications for the statistical calculations in this paper. Additionally, there are some assumptions that should be met for calculations to be useful.

First and foremost, classroom language assessments are supposed to yield information about students' constructs. This means instruments should provide data about the language

curriculum objectives teachers are helping students to attain; this key consideration is what scholars call content validity (Brown & Hudson, 2002). The assumption then is that language assessments and curriculum objectives should have a clear and direct relationship. If this is not the case, the data from instruments may lead to invalid decisions about students' language ability.

Another condition for scores to be meaningful is that assessment instruments have been designed soundly. A poorly designed assessment is not likely to trigger the relevant constructs and, by default, leads to unfair decisions about students' abilities. Indeed, the design of assessment instruments is a science of its own: Authors have dedicated extensive treaties on how to create useful items and tasks for language testing (Alderson, Clapham, & Wall, 1995; Brown, 2011; Carr, 2011; Giraldo, 2019).

A third condition for meaningful interpretation of scores deals with test administration. In ideal circumstances, when an assessment instrument is given to students, there should be no glitches: Sound systems should work properly, there are enough test copies for students, seating arrangements deter cheating, and there exist no background noises that can annoy students, among others. Even though teachers seem to discourage test administration as an important dimension of language assessment (see reports by Fulcher, 2012; Giraldo & Murcia, 2018; Vogt & Tsagari, 2014), administration is a key moment in the assessment enterprise.

Finally, readers of this paper should be familiar with Excel in order to perform basic calculations. In this paper, I include tables with the results of the calculations but do not explain how to arrive at such results; in other words, this paper is not a tutorial on how to use Excel for doing statistics. Interested readers may consult Brown (2011) and Carr (2011) for step-by-step guides. As explained, the core of this paper lies in the meaning and interpretations that can be made of basic statistics for classroom language assessment.

Descriptive Statistics

The purpose of descriptive statistics is to describe scores or numbers. They should be organized in a clear, informative way to allow teachers to report test results to interested parties (Gravetter & Wallnau, 2014). Additionally, as I suggest throughout this section, descriptive statistics can be used to evaluate patterns in assessments and student language performance, although their main purpose is to describe a score distribution and report overall results.

Frequencies and Percentages

The first statistics in this paper are frequencies and percentages. Table 1 presents these data based on the scores for a fictional reading test taken by 20 students. As stated earlier,

the interpretations are discussed through the lens of diagnostic, progress, and achievement purposes, within the overall framework of criterion-referenced testing.

Table 1. *Scores, Ranges, Frequencies, and Percentages for a Reading Test*

A	B	C	D	E
Student	**Score**	**Range**	**Frequency**	**Percentage**
1	1.5	0-0.9	0	0%
2	1.7	1.0-1.9	2	10%
3	2	2.0-2.9	5	25%
4	2.1	3.0-3.9	7	35%
5	2.5	4.0-4.9	6	30%
6	2.6	5.0	0	0%
7	3.2		20	100%
8	3			
9	3.1			
10	2.9			
11	3.4			
12	3.5			
13	3.7			
14	3.9			
15	4			
16	4.2			
17	4.3			
18	4.3			
19	4.7			
20	4.8			

Column A has all the numbers assigned to each student in this test; they might as well be your students' names. Column B contains every student score for this test, while C has the range of scores; these ranges are of course arbitrary and may represent levels of performance. For instance, students who get between 0.0 and 0.9 may be said to have an elementary level in the construct(s); students between 1.0 and 1.9 a basic level, and so on and so forth. Column D details how many students got scores within the specified range. For example, six students got scores between 4.0 and 4.9. If you count every score (Column B) that has a value between 4.0 and 4.9, then you get a total of six. If you add all of the values in column D, you should obtain twenty, the total of scores. Finally, column E tells you the percentage of students who fell within a specified range. So, out of twenty students, 25% got scores between 2.0 and 2.9.

Based on the results in Table 1, we can conclude that 35% of the students (or seven students) do not seem to have the constructs that the reading aimed to activate. Sixty-five % of the students do seem to have them. However, the percentages cannot be interpreted as pass-fail; for example, if these results came from a diagnostic assessment, they would be telling you that 65% of the students (or thirteen students) already seem to have mastered the reading skills under consideration. In other words, if the diagnostic assessment was based on the reading objectives for a course (and that should in fact be the case), then thirteen students have already learned them, without being in this particular course. Thus, based on these results, teachers might need to make changes to the language curriculum and devise ways to help the seven students who did not achieve a minimum score of 3.0.

A radically different interpretation is that the seven students who did not get a minimum of 3.0 are actually ready to be in the course. In other words, the diagnostic test is suggesting that the course is right for them, and they do not have the constructs the course aims to help students learn. This begs the question as to whether or not the other students (seventeen) are in the right course and should be in one with more advanced objectives.

If this assessment was being used for progress purposes, the results might be considered good news. Thirteen students seem to be learning the reading skills for the course, and seven of them need remedial work. The teacher in this course might decide to move on to other reading constructs for the course (or reading objectives), while assigning extra work for those who are not progressing well.

Finally, if this was used as an achievement test, you can conclude that thirteen students achieved the objectives in the reading course and seven did not. In such case, a teacher using this test would need to study why seven failed, given that the idea in an educational context is that all students learn the relevant constructs (Brown, 2011; Fulcher, 2010). Additionally, student 8 and student 9 got scores of 3.0 and 3.1 respectively, which means they barely passed the test. This begs the question as to whether the students really have the reading skills they were supposed to have.

Mode, Median, and Mean

As statistics, modes, medians, and means help you understand the central position of numbers (or scores) in a set of numbers. In Table 2, the same scores from Table 1 are reproduced with some minor modifications; additionally, this time the mode, median, and mean are added as statistical values.

Table 2. *Mode, Median, and Mean Values for a Set of Scores*

A	B	C	D
Student	**Score**	**Statistic**	
1	1.5	*Mode*	*4.3*
2	1.7	*Median*	*3.3*
3	2	*Mean*	*3.27*
4	2.1		
5	2.5		
6	2.6		
7	2.9		
8	3		
9	3.1		
10	3.3		
11	3.3		
12	3.5		
13	3.7		
14	3.9		
15	4		
16	4.2		
17	4.3		
18	4.3		
19	4.7		
20	4.8		

The **mode** is the most frequent score in a group of scores. In Table 2, students 10 and 11 got 3.3, which is the mode for this set of scores. No other number happens twice or more times. The **median** is the score that divides the set of scores into two. If the scores are ordered between 1.5 and 4.8, the number in the middle would be 3.3. Finally, the **mean** (3.27) is the mathematical average of a set of scores. To get the mean, one must add all the values in column B and divide this result by the number of scores, by 20 that is.

The mean is a useful statistic because it provides information about the group of students who took the test (Gravetter & Wallnau, 2014). In fact, means are widely used in applied linguistics research (Brown, 1988) because they help you to compare test scores for different groups of language learners. In the present case, if the mean was used for diagnostic purposes and with a "passing" grade of 3.0, then it is telling us that the group have a fairly good level of the assessed constructs. The mean then confirms, in one single number, that the interpretations of percentages made in the previous section are supported. It seems that most students already have the constructs that this test targeted.

If this assessment was used as a progress test, then a mean of 3.27 is telling us that students are doing well (i.e. progressing) and therefore learning the reading objectives for the course. On the other hand, should 3.27 be the mean for an achievement test –with 3.0 as the passing score–, then it would represent a fair level of achievement: In percentages, students in this group got 65.4% of the course reading skills.

Score Distributions

There is an assumption related to norm-referenced and criterion-referenced testing that I must address. In norm-referenced situations, the distribution of scores should be normal, as Figure 1 shows (from Carr, 2008, p. 51). Technically, the mode, median, and mean scores should be around the middle of the distribution.

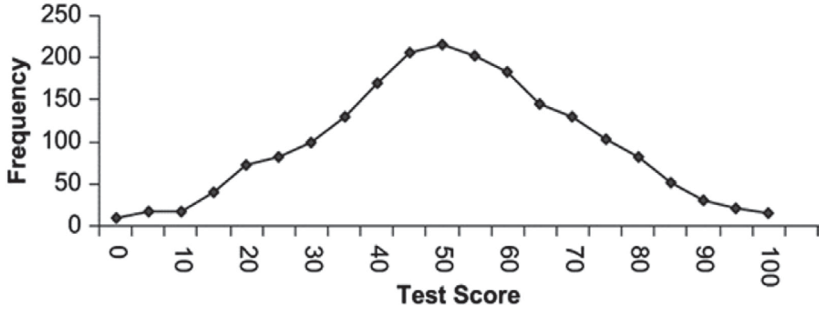

Figure 1. *Example of a Normal Distribution in Norm-Referenced Testing*

This distribution tells us that some test-takers scored low, the majority scored between 40 and 70, and a small fraction (approximately 50 test takers) scored between 80 and 100. Such distribution is normal and expected in tests such as TOEFL and IELTS; these tests are designed on a norm-referenced framework in which test takers are compared to one another based on their scores. If most people scored between 80 and 100, that would produce a bell-shaped curve at the far right, and, importantly, it would mean the test was too easy: Even students with a low level of language ability passed for reasons other than the constructs of interest, which would make interpretations and decisions relatively invalid for these norm-referenced tests.

In criterion-referenced testing, there are different expectations for the shape of a distribution of scores. For example, in Figure 2 (from Carr, 2008, p. 55), most scores are on the right end, and thus mode, median, and mean should be on this end, too. This can entail that, in this particular assessment, most students passed (if the passing score were 50). So, we would expect a shape as that in Figure 2 for achievement assessment. In diagnostic assessment, the bell-shaped curve should be on the left end, meaning most students would "fail" the test and are ready for instruction.

To summarize, modes, medians, and means give information about sets of scores. In norm-referenced testing, the values for these statistics should be roughly located in the middle of a score distribution as in Figure 1; in contrast, in criterion-referenced testing these values should be on the left end of a score distribution for diagnostic testing and right end for achievement testing.

Standard Deviation

Standard deviations are usually presented alongside modes, medians, and means. However, I find this statistic worthy of special attention and discussion, more so when

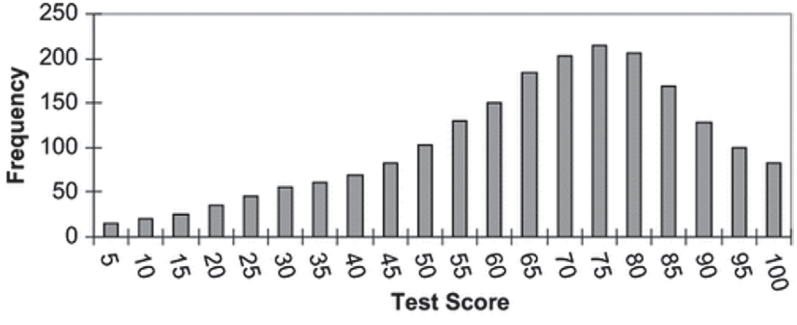

Figure 2. *Example of a Criterion-Referenced Score Distribution for Achievement Purposes*

interpretations from assessment are needed. Like means, standard deviations are common in language testing research. Brown (2011, p. 294, my emphasis) states that the standard deviation is "used to summarize the variation or distribution of scores around the mean; an averaging process considered a strong estimate of the dispersion of scores". A simple example may help illustrate conceptually the standard deviation.

Suppose Roberto got a 4.0 and Luisa a 4.4 on a listening test in which the group mean was 4.2. In this case, Roberto's score is away from the mean below by -0.2 and Luisa's above by +0.2. If you add these two scores and divide by two (0.2 + 0.2 / 2), then the result is 0.2. On average, Roberto and Luisa are 0.2 decimals away from the mean. So, 0.2 is the standard deviation: The average distance between all scores and the mean.

If we use the same values from Table 2, we get a mean of 3.27 and a standard deviation of 0.98. This standard deviation means the scores are rather spread: Some are really low and some are really high. Notice in Table 2 that Student 1 got a 1.5 and Student 20 got a 4.8. This is a big difference when constructs are to be interpreted. Student 1 does not have the constructs and Student 20 definitely does. To add to the dispersion, there are scores from ranges 2.0-2.9 to 4.0-4-9 (see Table 1). If we factor in the mean, we can say that, statistically speaking, some students got 2.29 (mean - standard deviation: 3.27 - 0.98) whereas some got 4.25 (mean + standard deviation: 3.27 + 0.98).

Additionally, the standard deviation has implications for the different language assessment purposes. If this standard deviation came from the scores in a diagnostic test, then we could argue that students have wide differences when it comes to the constructs under consideration. Thus, such a high standard deviation means some have the constructs and some do not, as I have argued so far in this article. Under ideal circumstances, scores from a diagnostic test should be low (no one should "pass") and the standard deviation should be low, too. These values would tell you that students do not have the constructs and are ready for instruction. Table 3 has an example.

Table 3. *Sample Values for a Diagnostic Test*

Mean	Standard Deviation
1.6	0.3

With the values in Table 3, we can conclude that some students got a score of 1.9 (+0.3 above the mean) or 1.3 (-0.3 below the mean). The interpretation is that the students have, similarly, a low level in the construct of interest.

A high value for a standard deviation may be fine for a progress test, because such assessment should tell you who is doing well and who is not. Students learn at different rates, and so dispersion around a mean is expected. Finally, as in diagnostic assessment, a low standard deviation is expected in achievement assessment, but with hopefully a high value for the mean, as Table 4 exemplifies.

Table 4. *Sample Values for an Achievement Test*

Mean	Standard Deviation
4.6	0.3

The values in Table 4 tell you that some students got a score of 4.9 (+0.3 above the mean) or 4.3 (-0.3 below the mean). You can now argue with certain confidence that the test used was fit for achievement purposes and that the students who took it have the constructs they studied during the course, i.e. they achieved the criterion.

In conclusion, the standard deviation is a useful statistic for criterion-referenced testing because it can tell you how similar or different students are in terms of their language constructs. Of course, this statistic needs to be analyzed against the mean and the purposes for which an assessment is used.

Evaluative Statistics

The name evaluative statistics is arbitrary. With descriptive statistics, we describe scores and their behavior; surely, we can also derive evaluations related to language constructs. With evaluative statistics, the purpose is to aim for test quality: They help us understand if something is off with assessment instruments or how they are used and help us to study possible solutions (Brown, 2003).

Item Facility, Difference Index, and B-Index

Item Facility. These three statistics are calculated specifically for tests which contain close-ended items; for example, multiple-choice questions (MCQs) and true/false are amenable to these statistics. Item facility (IF) tells you the proportion of students who got an item right or wrong, as the following example shows.

Suppose your students took an MCQ listening test, and you want to know the IF for all its items. If twenty students took this test, and nine students got Item 5 correctly, then

IF is 9/20 = 0.45. That is, 45% of the students got this item correct. Thus, MCQ 5 was a somewhat difficult item. Excel can help you to calculate IF for all items in a test by creating a spreadsheet as Table 5 illustrates.

Table 5. *Item Facility Values for a 10-Item Test*

Student	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
1	1	0	1	0	1	1	1	0	1	1
2	1	1	0	0	0	0	0	0	1	1
3	1	1	1	0	1	0	1	0	0	0
4	0	1	0	0	0	1	1	0	0	0
5	1	0	1	1	0	1	0	0	1	1
6	1	1	0	0	1	1	0	0	1	0
7	1	0	1	0	1	0	0	0	0	1
8	0	1	1	0	0	1	0	0	1	1
9	1	1	1	0	1	0	0	0	0	1
10	0	0	1	1	1	1	1	0	0	1
IF	**7**	**6**	**7**	**2**	**6**	**6**	**4**	**0**	**5**	**7**

In all cells, a 1 means the student got the item right, and a 0 wrong. The values in Table 5 tell you that the IF for Item 1 was seven: 70% of students got this item right. Item 8 was the most difficult because no one got it right. For a diagnostic test, IFs should be low, perhaps 45% or lower. If IFs on a diagnostic test are 50% or higher, then it may mean the items were easy or the students already have the constructs under consideration. In a progress assessment, IFs should have varying values, because they can be interpreted as some students having the constructs and some students not having them. Finally, on an achievement test, IFs should be high, meaning students have the construct. To further use IFs meaningfully, the following two simple calculations can be useful.

Difference Index and B-Index. These statistics are related to measurable language learning in a course. The difference index (DI) tells you to what extent a group of items shows what students learned during the course. To calculate it, two assumptions must be met. First, the same test or a similar test (assessing the same constructs) must be used as

diagnostic (pretest) and then as achievement (posttest). And second, IFs for the diagnostic and achievement must be calculated separately. Table 6 (based on Brown, 2011, p. 81) shows the DI for a fictional grammar test.

Table 6. *Values for Difference Index in a Grammar Test*

Item #	Post-Test IF	minus	Pre-Test IF	equals	DI
1	0.823	-	0.245	=	0.578
2	0.789	-	0.425	=	0.364
3	0.654	-	0.639	=	0.015
4	0.688	-	0.145	=	0.543
5	0.712	-	0.223	=	0.489
6	0.611	-	0.129	=	0.482
7	0.521	-	0.227	=	0.294
8	0.123	-	0.423	=	-0.3
9	0.742	-	0.514	=	0.228
10	0.645	-	0.396	=	0.249

In Table 6, Item 1 had an IF of 24% in the pretest (diagnostic) and an IF of 82% in the posttest (achievement). The DI tells you that there was a difference of approximately 57%, which means students started the course without the construct and now they seem to have it. Thus, the higher the DI, the more language learning seems to have occurred. One problem with DI is that an assessment has to be used twice. To solve this issue, Brown (2011) suggests the use of B-Index.

B-Index is a statistical value that tells you to what extent an item separates those students who have the construct from those who do not. In the case of an achievement test, B-Index helps you to identify how each item contributes to making a decision: Either pass or fail. Consequently, for B-Index to be useful, a cut score (or minimum passing grade) should be set. Recall that a passing grade in Colombia, generally, is 3.0 on a scale from 0.0 to 5.0. Table 7 shows B-Index values for a fictional achievement listening test.

Table 7. *B-Index Values for All Items in an Achievement Listening Test*

Student	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Score
1	1	1	1	1	1	1	1	1	1	1	5
2	1	1	1	0	0	1	0	1	1	1	3.5
3	0	0	1	0	1	1	1	1	1	1	3.5
4	1	1	1	0	1	0	1	0	0	0	2.5
5	1	1	0	0	1	1	0	0	1	0	2.5
6	0	1	1	0	0	1	0	0	1	1	2.5
7	1	0	1	0	1	0	0	0	0	1	2
8	1	1	0	0	0	0	0	0	1	1	2
9	1	0	1	0	1	0	0	0	0	1	2
10	0	0	0	0	0	1	1	0	0	0	1
IF pass	0.67	0.67	1.00	0.33	0.67	1.00	0.67	1.00	1.00	1.00	
IF fail	0.71	0.57	0.57	0.000	0.57	0.43	0.29	0.000	0.429	0.57	
B-Index	-0.04	0.09	0.42	0.33	0.09	0.57	0.38	1.00	0.57	0.43	

According to the B-Indices in Table 7, we can conclude that Item 8 undoubtedly separates those who learned the construct from those who did not. On the contrary, Item 1 is not doing a good job because it was easier for students who failed than it was for students who passed. Other good items are Item 6 and Item 9. As with the DI statistic, the higher the B-Index, the easier it is to ascertain that the assessment is working properly. In the case of the data from Table 7, Item 4 was difficult (IF = 0.33), even for those students who passed and five items (1, 2, 4, 5, and 7) are not separating "passers" from "failers" well. These items need revision. In conclusion, if this test was used, then its quality should be questioned and interpretations and decisions contested.

Agreement Coefficient and Kappa Coefficient. These last two statistics help you to analyze the decisions that you make based on data from an assessment. In other words, these coefficients will not tell you anything about the internal workings of items, as DI and B-Index do. Rather, they help you to determine whether pass-fail decisions are reliable, i.e. the degree to which a set of decisions is consistent across students and administrations. These statistics

can also be used with performance assessments of speaking and writing, in which raters use rubrics for decision-making. Since rubrics are not amenable to the calculations explained thus far, the main focus is on ascertaining to what extent raters agree with one another.

To calculate agreement coefficient, the same test needs to be administered twice so that the coefficient can tell you to what extent you were consistent in deciding whether students passed or failed the test on both occasions. This means you need two sets of scores for each student. The coefficient can easily be calculated by hand. The first step is to create a table as Table 8. Then, each cell in the table needs to be filled accordingly.

Table 8. *A Table for Calculating Agreement Coefficient*

A (pass-pass)	**B (pass-fail)**	A+B
C (fail-pass)	**D (fail-fail)**	C+D
A+C	B+D	A+B+C+D

In cell A, write the number of students who passed on both administrations. In cell D, write the number of students who failed on both occasions. Cell B should contain those who passed the first time and failed the second time. Lastly, in cell C write the number of students who failed the first time and passed the second time. Once cells A to D are filled in, proceed to calculate the rest of the information in Table 8. Table 9 includes an example for a test administered twice to a 30-student group.

Table 9. *Sample Values for Calculating Agreement Coefficient*

8	**8**	16
4	**10**	14
12	18	30

Now the values are ready for calculation. Add values A + D and divide them by the total number of students. Thus, 8 + 10 / 30 = 0.6, which translates into 60% in agreement. For this test, the teacher made the same pass or fail decision consistently 60% of the time. However, the values in the other cells (B and C) are not considered. In this case, the kappa coefficient is necessary as it uses all of the values for a more accurate number representing the consistency of pass-fail decisions. Before calculating kappa, another statistic is necessary:

pchance. Pchance uses all of the values from the table with pass-fail decisions. Here is the formula:

$$Pchance = \frac{[(A+B)(A+C)+(C+D)(B+D)]}{N^2}$$

With the values from Table 9, calculate Pchance.

$$Pchance = \frac{[(16)(12)+(14)(18)]}{30^2}$$

$$Pchance = \frac{[192+252]}{900}$$

$$Pchance = \frac{444}{900} = 0.49$$

Now that we have Pchance (**0.49**), the calculation for kappa is:

$k = \frac{(Po - Pchance)}{(1 - Pchance)}$ where Po means agreement coefficient. Thus, $k = \frac{(.60-.49)}{(1-.49)} = k\frac{.11}{.51} = 0.21$

In conclusion, because kappa used all values in Table 9 and because there were inconsistencies (pass-fail and fail-pass cases), there was in fact little consistency (21% of the cases) in the decisions the teacher made. This is a serious issue with several possible causes: Students did not perform similarly on both occasions given extraneous circumstances, there was a problem scoring test responses on one or both occasions, or the students knew the answers to the test the second time they took it. Further investigation is warranted.

As scholars agree, it is not practical to have students take the same test twice, though scores from comparable tasks could be used, as Fulcher (2010) argues. As I see it, there is another useful application of the agreement and kappa coefficients: A speaking or writing test scored by two teachers. In this situation, every student will have two scores, so consistency can be calculated as before.

For example, suppose that you and another teacher diagnosed students' writing skills before a course started. Both used a rubric and produced a score for each student. Then, upon going over your decisions, you have these data for 40 students in Table 10.

Table 10. *Decisions for a Diagnostic Writing Test Scored by Two Teachers*

18 (pass-pass)	**2 (pass-fail)**	20
2 (fail-pass)	**18 (fail-fail)**	20
20	20	40

With these numbers, the agreement coefficient is 0.90, or 90%. Kappa turned out to be 80%. In general, this means both teachers reached a substantial level of agreement. Fulcher (2010, p. 83) provides this rule of thumb for interpreting kappa:

.01–.20	slight agreement
.21–.40	some agreement
.41–.60	moderate agreement
.61–.80	substantial agreement
.81–.99	very high agreement

There are some interpretations for the resulting kappa (80%). First and foremost, it seems like both teachers knew what they were assessing, which then means the way the construct of writing was specified in the writing rubric was clear for both. Second, in a related manner, both teachers used the rubric fairly well: It seems that they were not influenced by extraneous factors not considered in the assessment. Notwithstanding this good news, the teachers did not agree 20% of the time (eight students out of 40), so they should discuss what they differed on and substantiate their decisions so they can reach a fairer score for the eight students involved. Criterion-referenced assessments should aim for substantial agreement or above.

To summarize, agreement and kappa coefficients help you to ascertain to what extent decisions were consistent (i.e. reliable) across two administrations of a test. Low levels of agreement should alert teachers so that they can evaluate what is happening with assessments and the way they are used.

Limitations

Since this paper is mostly concerned with criterion-referenced language testing, I paid little attention to large-scale testing, even though it has the potential to influence classroom assessment, as some authors have indicated (Fulcher, 2012; Inbar-Lourie, 2008; 2012). However, my aim was to show how numbers can be useful in classroom language assessment through simple yet useful calculations that teachers can do. For advanced statistical calculations and interpretations in norm-referenced situations, readers may wish to consult especially Bachman (2004), but also Brown (2011) and Carr (2011).

As I mentioned in the introduction, this paper does not deal with issues pertaining to the design of assessments. Statistical calculations can provide information about items or tasks that are not working properly, so a more design-based approach is needed to identify what happens at the level of instruments themselves. Thus, teachers can conduct expert reviews

of items and tasks to further investigate the quality of their assessments after they have done the calculations presented here. For qualitative expert review, Brown and Hudson (2002) and Brown (2011) provide suggestions.

Fulcher (2010) emphasizes the fact that, in language assessment, we live with uncertainty. It is unlikely for an assessment system to provide perfect numbers (i.e., a kappa of 100%; a standard deviation of 0.0 on an achievement test), but we need to make every possible effort to ensure that our assessments are useful for their intended purposes. To do so, statistics can help monitor quality. In the particular case of classroom language assessment, the major focus is on substantiating the consequences of our actions (Fulcher & Davidson, 2007; Moss, 2003), to which statistics can contribute. Living with uncertainty is expected and inevitable.

Conclusions and Recommendations

Large-scale language testing largely depends on using appropriate statistics to argue for the quality of tests and the consequences that can derive from interpreting their scores. Similarly, as I have attempted to show in this paper, interpretations of scores, and numbers in general, can provide information about classroom language assessment: state of students' constructs, quality of items and tasks, and appropriateness of decisions based on scores. For interpretations, I used two types of statistics, descriptive and evaluative, to describe scores and numbers and what they could mean for diagnostic, progress, and achievement purposes in the language classroom. With these available statistics, the following recommendations may prove useful for language teachers in general.

Teachers new to the area of statistics for language assessment may see this task as daunting. However, statistics can be used for tests which have high stakes in their school. For example, a final achievement test worth 50% of a course should have a high quality. Teachers in this context may use the pretest/posttest treatment and calculate descriptive statistics, IF, DI, and B-Index. The resulting numbers can certainly help in raising the quality of such assessment.

Other recommendations based on the statistics presented in this paper are the following:

- Calculate descriptive statistics and compare and contrast groups that are in the same grade and should have similar levels in the constructs of interest.
- Calculate IF, DI, and B-Index with different groups of learners; for instance, a group of students who already passed a course and a group who is about to start the course (pre and post). Brown and Hudson (2002) call this a differential groups study.
- Calculate agreement coefficient and kappa with a speaking or writing test you can assess two times. Assess performance the first time, wait a few days, and then as-

sess again. Finally, do the statistics. This is commonly called intra-rater agreement (Hughes, 2010; McNamara, 2000): To what extent do you agree with yourself when you decide on a passing or failing grade?

- Use all of these calculations, and others, with your colleagues to collectively learn what numbers can tell you about language assessment in the classroom.

In closing, a modicum of statistics nurtures your language assessment literacy, which in turn can help you become more critical towards the language assessment enterprise. In the end, if used appropriately, knowledge and skills in basic statistics will have positive consequences on language assessment, teaching, and learning.

References

Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Bachman, L. F. (2004). *Statistical analysis for language assessment*. Cambridge University Press.

Brown, J. D. (1988). *Understanding research in second language learning*. Cambridge University Press.

Brown, J. D. (2003). Questions and answers about language testing statistics: Criterion-referenced item analysis (The difference index and B-index). *SHIKEN: The JALT Testing & Evaluation SIG Newsletter*, 7 (3), 18-24.

Brown, J. D. (2011). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw Hill.

Brown, J. D. (2013). Teaching statistics in language testing courses. *Language Assessment Quarterly, 10*(3), 351-369. https://doi.org/10.1080/15434303.2013.769554

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press.

Carr, N. T. (2008). Using Microsoft Excel® to calculate descriptive statistics and create graphs. *Language Assessment Quarterly, 5*(1), 43-62. https://doi.org/10.1080/15434300701776336

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.

Chapelle, C. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21-33). Routledge.

Douglas, D. (2014). *Understanding language testing*. Routledge.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Fulcher, G. (2010). *Practical language testing*. Hodder Education.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113-132. https://doi.org/10.1080/15434303.2011.642041

Giraldo, F. (2019). Designing language assessments in context: Theoretical, technical, and institutional considerations. *HOW Journal, 26*(2), 123-143. https://doi.org/10.19183/how.26.2.512

Giraldo, F., & Murcia, D. (2018). Language assessment literacy for pre-service teachers: Course expectations from different stakeholders. *GiST: Education and Learning Research Journal, 16*, 56-77. https://doi.org/10.26817/16925777.425

Gravetter, F., & Wallnau, L. (2014). *Essentials of statistics for the behavioral sciences. 8th edition*. Cengage Learning.

Hughes, A. (2010). *Testing for language teachers*. Cambridge University Press.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385-402. https://doi.org/10.1177/0265532208090158

Inbar-Lourie, O. (2012). Language assessment literacy. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-9). John Wiley & Sons. https://doi.org/10.1002/9781405198431.wbeal0605

Malone, M. (2017). Training in language assessment. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment: Encyclopedia of language and education* (3rd Ed., pp. 225-240). Springer. https://doi.org/10.1007/978-3-319-02261-1_16

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

Ministerio de Educación Nacional de Colombia (2016). *Pedagogical principles and guidelines suggested English curriculum*. Team Toon Studio.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice, 22*(4), 13–25. https://doi.org/10.1111/j.1745-3992.2003.tb00140.x

Popham, W. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice, 48*(1), 4-11. https://doi.org/10.1080/00405840802577536

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374-402. https://doi.org/10.1080/15434303.2014.960046