A SHORT, NARROW, AND BIASED INTRODUCTION TO STATISTICS

Stephen Krashen

Contents

This short book attempts to provide the basics of statistics. It can serve as a "pre-statistics" course, to give students some basic ideas before taking an actual course, or provide a quick review. The goal is to help you understand papers that use basic statistical methods.

All examples are straightforward; my view is that if you understand the easy examples, you will have no trouble with the more complex cases.

There is no attempt to cover the same material introductory courses cover. My focus is on concepts that I consider to be important in order to understand current research publications.

The Mean

The mean is simply an average. To compute a mean, all you do is add up the scores and divide by the number of scores. So given these numbers, 2,3,4,5, if you want to compute the mean you add them (and get 14), then divide by the sum of the number of scores (4), giving you 14/4 = 3.5. If you understand this, you understand the mean and are ready to go on.

The symbol most frequently used to signify the mean is a capital X with a bar on top: X.

There are two other ways of calculating averages in statistics, the mode and the median, but we won't discuss them here. The mean is the most common kind of average used in studies.

The Standard Deviation

When researchers report a mean, they nearly always report the standard deviation with it, which is a very good idea.

I will not show you how to calculate the standard deviation, but only explain what it means.

Whenever you see "standard deviation" you know that about 2/3 of the scores fall between one standard deviation above the mean and standard deviation below the mean. Don't worry about why: For now, it is enough to understand that this is what the standard deviation means.

Here are a few examples.

A teacher gives a test to his students. The test had 100 items. The mean was 50 and the standard deviation was 20. This means that about 2/3 of the students' grades fell between 70 and 30. (50, the mean, + 20, the standard deviation, = 70; 50, the mean, - 20, the standard deviation, = 30).

If the standard deviation had been 10, 2/3 of the scores would fall between 60 and 40. This is because 50 + 10 = 60 (mean plus one standard deviation) and 50 -10 = 40 (mean minus one standard deviation).

The larger the standard deviation, the more "spread out" the scores are: when the standard deviation is 10, 2/3 of the scores fall between 60 and 40. When it is 20, 2/3 of the scores fall between 70 and 30.

Standard deviations thus measure "variability," how scattered or spread out scores are from an average or mean score.

If the standard deviation is small, all the scores are bunched up close to the mean. A large standard deviation means greater variability.

Here are some real-life (or nearly real-life) examples: Researchers Haeyoung Kim and Kyung-Sook Cho gave a test of English vocabulary to 420 college students in Korea. The mean score was 50, and the standard deviation was 2. (Actually it was 2.26, but let's keep things simple.) This means that about 280 (2/3 of 420) of the students scored between 48 and 52. There wasn't much variation in this case; most students scored near the group average.

Researcher Beniko Mason gave an English test to 90 college students in Japan. The mean score was 30 and the standard deviation was 10. Thus, 2/3 of 60, or about 60 students scored between one standard deviation below the mean (30 -10 = 20) and one standard deviation above the mean (30+10 = 40).

Warning: All this information about standard deviations is only perfectly accurate if we have what is called a "normal distribution": In a normal distribution, most of the scores are concentrated near the mean at the center, half the scores are above and half are below the mean, and there are fewer and fewer of them as they get farther from the mean.

If a study only includes a few subjects, the distribution may not be normal. The more subjects, the better the chance that the distribution is normal.

Let's move to two standard deviations. About 95% of a group falls within two standard deviations of the mean (to be precise, 95.4%). Returning to Haeyoung Kim and Kyung-Sook Cho's study, the mean was 50 and the standard deviation was 2. This means that 95% of the subjects (400 out of 420) scored between 46 and 54.

Work through the next example. If we come to the same conclusions, you know enough about standard deviations to continue to the next section.

The PIRLS organization collected data from 45 different countries. One of the questions they asked was the number of books in classroom libraries. The average (mean) response was 66, and the standard deviation was 58. This means that 2/3 of the countries reported between 110 and 22 books in their classroom libraries. (2/3 of 66 = 44. 66 + 44 = 110,66 – 44 = 22). (Data can be found in Krashen, Lee and McQuillan, 2012).

Here is another example, in case you want to be sure.

Mason and Krashen administered a test to 22 university students of English as a foreign language in Japan. The test had 100 items. The mean (average) score was 22 correct. The standard deviation was 12.

This means that about 15 of the 22 subjects (2/3 of 22) scored between 10 (one standard deviation below the mean) and 24 (one standard deviation above the mean).

Several standardized tests have a mean of 500 with a standard deviation of 100. This is true of the SAT (Scholastic Aptitude Test) taken by high school students in the US and used for admission to universities, and also the PISA test (Programme for International Student Assessment), an international test given to 15 year olds in 70 different countries in math, reading and science.

Thus, if a country gets a score of 600 on the PISA, it is one standard deviation above the mean – a score of 400 is one standard deviation below. Usually the top scores are less than 550.

The t-Test and Statistical Significance

In the TPRS (Teaching Proficiency Through Reading and Story Telling) method of teaching second and foreign languages, the focus is on providing interesting, comprehensible input, largely in the form of stories that teachers and students co-create.

A study by Barbara Watson compared TPRS students in two classes with traditionally taught students in one class. Here are the results from the final test, given at the end of the school year, which covered listening, reading, and writing:

Comparison of TPRS and Traditional Instruction

Group	N	mean (sd)
TPRS	50	63.9 (4.0)
Traditional	23	58.2 (7.9)

from: Watson (2009).

It looks like the TPRS group did better; their scores on the final exam were five and half points higher. Can we conclude that the TPRS group really did better? What if the traditional students had scored 59? Still, according to my intuition, the readers seemed to have done better. What if the traditional students had scored 60, or 61? Now it is not so clear.

A t-test tells us when the difference between means is so great that it is unlikely that the two groups are the same. In other words, a t-test tells us where to draw the line, how to tell whether differences are likely to be "real."

To compute a t-test you need the "raw scores," the actual ratings each student received in on each kind of measure. A fairly simple formula is used to compute the t-scores. Computers do this for us very well and the internet offers various free t-test calculators.

When the results of t-tests are included in scientific reports, the authors give the means, the standard deviations, and a "t-score," followed by some other mysterious symbols, numbers and terms, such as $p < .05$ or $p < .01$ or "not significant." So the whole thing typically looks like this: $t = 4.06$, $p < .001$. The goal of this section is to explain to you what these symbols and terms mean.

We first look at the "p" part.

To explain what "p" means, we need to discuss statistical significance. We will discuss statistical significance in terms of the t-test, but this idea is used with all statistical tests. We will then tell you what t-scores are and have another look at Watson's results.

Statistical Significance

Let's flip a coin. What are the chances, or the probability of getting a heads on the first flip? The odds are 50-50 or 50%. In mathematical language, $p = .50$. "p" stands for "probability."

How about two heads in a row? The chances of this happening are one in four. In other words, it will typically happen about 25% of the time. Now p, the probability, = .25.

For three heads in a row, the chances are about one in eight, or 12.5%. Now $p = .125$.

For four times in a row (please be patient!), the chances are one in 16, or .0625%, and $p = .0625$. And the chances of getting five heads in a row is one in 32, or .03%, and $p = .03$.

Most people would agree that it is not unusual to get one, two, or three heads in a row.

Four is a bit strange. When you get five heads in a row, you suspect something is wrong with the coin: The chances of getting five heads in a row is highly unlikely.

When results of t-tests (and other statistical tests, as we will see later) are given, they are followed by a statement that tells us how likely it was that such a difference between means could have happened by chance.

The .05 level and the .01 level

If the chances of getting a certain result is .05 (when p = .05), this is roughly equivalent to getting more than four heads in row in flipping a coin. p = .01 is about as likely as getting more than five heads in a row.

Statisticians have agreed that when p = .05 or less (p< .05), the difference is "significant," that is, it is probably real and did not happen by chance. When p is .01 or less (p < .01) the difference is considered to be "very significant."

Let's take another look at Watson's results. I have added the results of Watson's t-test in the following table. It tells us that the difference between the means is very significant, because p is less than .01. In fact p is much less than .01.

TPRS vs. Traditional Instruction (Watson)

Group	N	mean (sd)
TPRS	50	63.9 (4.0)
Traditional	23	58.2 (7.9)

t = 4.06, p = .0001.
from: Watson (2009).

t-values

Here are the results of another study, done by Joseph Dziedzic, again comparing TPRS with a more traditional approach to Spanish for high school students. The test was given after one academic year, and Dziedzic supplies us with results for all four components of the test.

TPRS vs. Traditional Instruction (Dziedric)

	Listening	Reading	Writing	Speaking
TPRS	11.63 (3.02)	12.89 (3.52)	8.25 (1.39)	3.5 (.66)
Traditional	11.82 (2.77)	12.12 (3.95)	6.77 (2.31)	2.8 (.78)
t	0.5889	0.023	3.08	3.82
P	0.558	0.82	0.0031	0.0003

Number of subjects:
Listening – TPRS = 30, Traditional = 28; Reading – TPRS = 28, Traditional = 26;
Writing – TPRS = 32, Traditional = 30; Speaking – TPRS = 32, Traditional = 30
From: Dziedric (2011)

As you can see in table 1, higher t-score are associated with lower p-values. For writing and speaking, the t-scores are greater than 3.0 and the p values are well below .001, which means the differences are very significant and it is very likely that the TPRS students really did do better than the traditional students. For listening and reading, the t-values are quite small and the p-values are high, not even close to the .05 level. This means that the two groups did not perform differently on the listening and reading tests.

A rule of thumb

Some researchers recommend the following "rule of thumb": t-values of greater than two are generally statistically significant. Here is an example where this rule of thumb works: Beniko Mason compared students of English as a foreign language in Japan who either spent a year in a class in which the focus was on extensive, self-selected English reading and writing summaries of what they read in English, or a traditional class. The readers did better on a test of reading comprehension, and the result was significant. Note that t is just barely more than 2, and the p level is significant at the .05 level.

Comparison of extensive reading + summary writing in English and traditional methodology

Method	n	means	sd
Extensive reading	36	69.39	7.62
Traditional	37	65.57	8.35

$t = 2.09$, $p < 0.05$;
from : Mason and Krashen (1997, study III).

Another group of readers wrote summaries in Japanese, not English, and they did even better on the reading comprehension test and the p-value tells us that it is even less likely that this happened by chance.

Comparison of reading + summary writing in Japanese and traditional methodology

Method	n	means	Sd
Extensive reading	36	70.5	7.08
Traditional	37	65.57	8.35

$t = 2.72$, $p < 0.01$
from : Mason and Krashen (1997, study III).

[Technical note: For those interested in more detail, the p-value is determined by a formula that takes two things into consideration: the t-value (as noted above, the higher the t-value, the more likely it is that the difference is significant) and the number of subjects (for a given t-value, the more subjects, the easier it is to get statistical significance).]

Effect size

Effect sizes allow us to attach a number to the size of differences. P-values don't do this. Remember that "p" stands for "probability" – p-values only tell us how likely it is that a result was the result of chance. Effect sizes tell us how much impact a treatment had, the size of the effect.

Effect sizes are simple. When used with the results of experiments, effect sizes tell us how much better (or worse) the experimental group did than the comparison group: the experimental group is the one that got the treatment (eg a different teaching method). If a treatment had zero impact, the effect size is zero. If the experimental group did better, the effect size is positive, if the experimental group did worse, the effect size is negative.

According to common practice in education, an effect size of about .2 is considered small, an effect size of .4 or .5 is considered modest, and an effect size of .8 or .9 is considered large.

Let's go back over the examples presented in the t-test section. (No you don't have to go back and re-read this section, I'll repeat the essentials here.)

Here is the table for Dreidzic's study, with an additional row added, labeled "d". "d" is the usual abbreviation for effect size (I have no idea why the letter "d" is used.)

TPRS vs. Traditional Instruction (Dreidzic) with effect sizes

	Listening	Reading	Writing	Speaking
TPRS	11.63 (3.02)	12.89 (3.52)	8.25 (1.39)	3.5 (.66)
Traditional	11.82 (2.77)	12.12 (3.95)	6.77 (2.31)	2.8 (.78)
t	0.5889	0.023	3.08	3.82
p	0.558	0.82	0.0031	0.0003
d	-0.16	0.006	0.79	1

Number of subjects:
Listening – TPRS = 30, Traditional = 28
Reading – TPRS = 28, Traditional = 26
Writing – TPRS = 32, Traditional = 30
Speaking – TPRS = 32, Traditional = 30

The first effect size, for listening, is negative and very small, meaning that the traditional students were a bit better. The reading effect size is very close to zero. These effect sizes agree very well with the impression one gets from looking at the means. The next two effect sizes are very large, both over 3, agreeing with the clear differences in means in the writing and speaking conditions.

[Technical note: You may have noticed that smaller p-values are associated with larger effect sizes, but this isn't always true. It is possible to have a large p-value and a small effect size when sample sizes are very large and differences are very small but consistent.

Just take my word for it for now: Effect sizes are the way we measure impact or size of a treatment, not the t-score or p-value.]

I am going to depart from my usual style and actually show you how effect sizes are calculated. The formula is simple and to make it even easier, I will use a made-up example.

The effect size formula is this:

The mean of the treatment group – the mean of the comparison group/ the pooled standard deviation.

The pooled standard deviation is the standard deviation of the treatment group and comparison group combined, in other words, the average of the two standard deviations.

The first example is straightforward because the standard deviation is the same in both groups. But the means are different.

Experimental group mean = 120.0 standard deviation = 10
Comparison group mean = 100.0, standard deviation = 10

The formula gives us: 120-100/10 = 20/10 and d = 2.0.

In other words, the experimental group scored two standard deviations better than the comparison group.

Here is another one:

Experimental group mean =105.0 standard deviation = 10
Comparison group mean = 100.0, standard deviation = 10

The formula gives us: 105-100/10 = 5/10. d = .5

In other words the experimental group scored ½ of a standard deviation better than the comparison group.

The next example is only a little more complicated. Each group has the same number of subjects, let's say 100 (we would say n = 100, where "n" stands for number.). If the number of subjects in each group is the same, we don't have to worry about it in the calculation of the effect size.

Experimental group mean =105.0 standard deviation = 10
Comparison group mean = 100.0, standard deviation = 20

The standard deviations are different, and we want the average, or pooled standard deviation, which would be 15.

So the effect size would be 105-100/15 = 3/15 = .33.

That’s it!

The next table contains the effect sizes for all the examples presented in the t-test section:

Study	t	P	d
Watson	4.06	0.001	1.02
Driedzic: listening	0.5889	0.558	-0.16
Driedrzic: reading	0.023	0.82	0.006
Driedrzic: writing	3.08	0.0031	0.79
Driedrzic: speaking	3.82	0.0003	1
Mason: summaries written in English	2.09	0.05	0.48
Mason: summaries written in Japanese	2.72	0.01	0.64

The Correlation Coefficient

The correlation coefficient, sometimes just called the correlation, is one of the most important statistical concepts. It shows us to what extent two sets of data are related.

Here is an example from very current research. Professor N. Pratheeba of the Kamaraj College of Engineering and Technology in India wanted to know if those who read more had larger vocabularies.

She gave a vocabulary test to a group of 20 engineering students who were very advanced in English. The test was difficult, including words such as’ zealot’, ‘liability’, and ‘overindulgence.’ She also gave them a questionnaire that asked them about their reading, including questions about different kinds of reading (e.g. newspapers, journals, political novels, historical novels, and science fiction). The questionnaire also had several questions asking about reading from the internet.

Prof. Paratheeba ran a correlation between results of the questionnaire and results of the vocabulary test. If those who read more have larger vocabularies, the correlation will be positive. If there is no relationship, the correlation, or correlation coefficient, will be small, close to zero. If those who read more have smaller vocabularies, the correlation will be negative.

Correlation coefficients are expressed in a simple way. Positive correlations are greater than zero but not larger than one or smaller than -1.0. A correlation of one (written as r = 1.0, where “r” stands for “correlation”) means that there is a perfect relationship between the two variables. As one goes up, the other goes up; as one goes down, the other goes down. The closer r is to one, the stronger the positive correlation. A correlation of .8 or .9 is considered to be quite strong. A correlation of .5 is considered to be "modest" and .2 is a "weak" correlation.

Similarly, a correlation of r = -1.0 means that as one variable gets larger, the other gets smaller.

(Note that effect sizes can be larger than 1, but correlations can't be larger than 1.)

Pratheeba reported that the correlation between the results of the questionnaire and scores on the vocabulary test was .63. As we said just above, correlations are usually represented with the letter "r", so we can write r = +.63, a positive correlation. Those who said they read more did better on the vocabulary test. This is very significant, p = .001.

Pretheeba was also interested in knowing if reading from the computer resulted in better vocabulary: she correlated scores on the vocabulary test with those items from just the questionnaire that referred to reading from the internet. The correlation was .42, or r = .42, which was significant, p = .03. Yes, reading from the computer is related to having a larger vocabulary.

Here is a fascinating table that illustrates the use of the correlation coefficient on a topic that is of interest to nearly everybody: The chances of dying from a heart attack.

Predictors of heart attacks

	men	Women
Smoking	0.28	0.44
Saturated Fat	0.64	0.62
Wine	- .70	- .61
Beer	0.23	0.31
Hard liquor	-. 26	-. 32

From: St Leger, Cochrane, and Moore (1979).

The correlations show that more smoking means a greater chance of dying of a heart attack, and the risk is slightly higher for women. The more saturated fat you eat (such as animal fat), the greater a chance you have a dying of a heart attack, and the correlation is quite substantial. Drinking alcohol has different effects, depending on what you drink. The more beer you drink, the greater your chances of dying of a heart attack. But the correlation between drinking hard liquor or wine is negative: That means that more wine or liquor drinking is associated with fewer deaths by heart attack. Notice also that the negative correlation between wine consumption and death by heart attack is quite high, -.61 for women and -.70 for men.

"Correlation is not causation"

A very important point is that "correlation is not causation." Showing that two variables are correlated does not necessarily mean that one caused the other. Pratheeba, as you recall, found that those who did better on a vocabulary test reported reading more. This could mean that

1. More reading results in more vocabulary knowledge.
2. Those who study vocabulary more and learn more vocabulary also read more.

Pratheeba's results do not tell us which one is correct. (Other research, however, strongly suggests that (1) is correct and (2) is not. See e.g. Krashen (2004), The Power of Reading.)

More examples

The following set of correlations comes from a project of ours, Krashen, Lee, and McQuillan (2011):

Predictors of Performance on the PIRLS Examination.

predictor	Reading Score
Poverty	-0.71
Independent Reading Time	0.5
School Library	0.56
Amount of Instruction	-0.26

From: Krashen, Lee, and McQuillan, 2011

These correlations are taken from a huge international report on the PIRLS test, a test given to tenth graders in over 40 countries. We calculated these correlations based on the data supplied by PIRLS. We had data on 44 countries.

"Poverty" in this study was based on "The Human Development Index" and is an average of three factors: education (adult literacy rates, school enrollment), life expectancy, and wealth (see http://hdr.undp.org/en/statistics/indices/hdi/.) The correlation between poverty and scores on a reading test was high, r = -.71, meaning higher poverty was related to lower reading scores. This result is very significant (p < .001).

PIRLS supplied us with data on the percentage of students in each country who were given time to read independently in school every day or almost every day. The correlation of independent reading and scores on the reading test was r = .5, also very significant (p < .001). More students reading in school was related to higher reading scores.

"School library" meant the percentage of schools in each country that had a school library containing at least 500 books. This correlated quite highly with scores on the reading test as well: r = .56, p < .001.

"Amount of instruction" meant the hours per week each country said were devoted to reading instruction in its schools. The correlation between amount of instruction and reading test scores was -.26. This means that more instruction was related to lower reading scores, a surprising result. This result was not quite significant (p = .09).

These results appear to show that high poverty is related to lower reading scores, that more independent reading in schools and the presence of libraries is related to higher scores, and more instruction means lower reading scores. All of these, poverty, independent reading, libraries, and amount of instruction, appear to be good predictors of reading scores.

This may or may not be true. The next table includes correlations among all the predictors.

Correlations among all predictors of the PIRLS test results

Correlations among all predictors	Reading Score	Poverty	Independent Reading Time	School Librar
Poverty	-0.71			
Independent Reading Time	0.5	-0.43		
School Library	0.56	-0.37	0.51	
Amount of Instruction	-0.26	.4	0.04	0.17

Notice that poverty is correlated with ALL the other three predictors: Countries with high levels of poverty provide fewer students with time for independent reading in school (r = -.43), have fewer school libraries (r = -.37) and give children more instructional time in reading (r = .4). To make matters even more complicated, countries that have more school libraries also provide more students with time for independent reading (r = .51).

The next section provides a way of dealing with all these inter-correlations.

Multiple Regression

Multiple regression is amazing. With multiple regression, a researcher can determine the impact of one variable, while holding the effect of other variables constant.

Multiple regression solves the problem of all of those correlations among the predictors in the previous table. Multiple regression is one of the great breakthroughs of statistics. Multiple regression allows us to examine the effect of several predictors at the same time, while controlling for their effects on each other. It allows us to pretend that the predictors are not correlated with each other. In this example, it allows us to not worry about the fact that countries with higher levels of poverty provide fewer school libraries and allow fewer students to do independent reading.

Here is a multiple regression analysis for the data just presented in the previous section. It was accomplished by the computer, using complex mathematics.

Multiple Regression Analysis: PIRLS

Predictor	Beta	P
Poverty	-.42	0.003
Independent Reading	.19	0.09
Library	.34	.005
Instruction	-.19	0.07
r2 = .63		

The beta column is the important one. Betas can be used to compare the strength of different predictors, similar to the way effect sizes do.

In this study, poverty was the strongest predictor, because it had the highest beta, beta = -.42. Note that the beta was negative, as was the case with the correlation of poverty and reading test scores (r = -.71).

The next highest beta is for school libraries (beta = .34). It is very interesting that the impact of libraries on reading scores is nearly as high as the impact of poverty, with libraries having a positive impact and poverty a negative impact. Studies have found that children of poverty have little access to books at home and in their communities. The results presented here suggest that providing access to books through a school library can make up for this lack.

Note that the p-values for poverty and libraries are very significant, that is, the betas are very unlikely to have occurred by chance.

The betas for independent reading and instruction are the same size, one positive and one negative, and neither quite reaches statistical significance, but both close. It appears that a larger percentage of students doing independent reading in school is positively associated with scores on the reading test, but modestly, while more instruction predicts slightly poorer performance.

Let’s compare betas and correlations. As seen in the next table, the pattern of both is similar: poverty and the library have a positive impact, while reading time is positive and instruction is negative.

	Correlation	Beta
Poverty	-0.71	-0.42
Independent Reading Time	0.50	0.19
School Library	0.56	0.34
Amount of Instruction	-0.26	-0.19

But the multiple regression changes things somewhat: Note that independent reading is now clearly weaker than the library as a predictor of reading test score performance. Just why this happens is beyond the scope of this presentation. The important point is that the betas are "purer" predictors of reading ability. They are not influenced by the intercorrelations among the predictors. We can, in effect, pretend that the predictors are not correlated with each other at all, that they are completely independent.

[Technical note: Multiple regression works very well as long as the correlations among the predictors are not super-high. This is called "multicollinearity." Multicollinearlity is not a problem in the examples we are considering here.]

Here are more multiple regressions:

Competence in the subjunctive in Spanish as foreign language in the US

Predictor	Beta	P
Study	0.0052	0.72
Residence	0.051	0.73
Reading	0.32	0.034
subjunctive study	0.045	0.76

From: Stokes, Krashen & Kartchner, 1998

This example deals with the success in acquiring the subjunctive in Spanish, a form that is traditionally difficult to master for students of Spanish.

In this study, speakers of Spanish as a second language who were living in the US were given a test of Spanish in which required the use of spontaneous, unrehearsed speech, and their speaking scores were rating by experts. The speaking situation was set up so that the subjects would have to use the subjunctive quite a bit.

All subjects filled out a questionnaire that asked them how many years they had studied Spanish in school ("study"), how long they had lived in a Spanish-speaking country ("residence"), how much pleasure reading they had done in Spanish and how much formal study they had done specifically of the subjunctive.

A look at the table shows that the winner is reading: Reading has by far the largest beta, and it is statistically sigificant (p = .034). The betas of the other three predictors are much lower and not statistically significant (all are clearly larger than .05). The results indicate that those who read more have acquired the subjunctive better, not those who studied it more or lived longer in a Spanish-speaking country.

Because this is mulitple regression, beta indicates the impact of each predictor independent of the influence of other predictors: In the last example, it is likely that those who have lived longer in a Spanish speaking country have also read more in Spanish. But the beta tells us the influence of each predictor assuming this is not true, assuming that the correlation between living in a Spanish-speaking country and reading Spanish is zero. In statistical terminology, the influence of other predictors is "controlled," or "held constant."

Another example:

Prof. Kyung Sook Cho from Busan National University of Education in Korea was interested in what motivates her students to keep reading in English. She asked 32 undergraduate students studying English as a foreign language to read the first chapter of the novel *Twilight* in English. For her study, she only included those who had not read *Twilight* before, in English or in Korean.

She asked the students to fill out a questionnaire in Korean, asking them:

-if they were pleasure readers in English (They were asked, "Do you read English books (fiction, non-fiction, magazine, etc.) for fun? (1) not at all (2) no (3) moderately (4) yes (5) a lot.").

-if they had seen the *Twilight* movie,

-if they found the chapter of *Twilight* difficult (They were asked "Was the chapter easy to read? (1) very difficult (2) difficult (3) moderately (4) easy (5) very easy"),

-whether they found the chapter enjoyable (They were asked "How much did you enjoy reading chapter 1? (1) not interesting at all (2) no fun (3) moderately (4) interesting (5) very interesting), and

-if they were interested in reading more of the *Twilight* book: ”If you had the time, would you like to read the entire book? (1) not at all (2) no (3) moderately (4) yes (5) a lot."

Here are the results, presented as a multiple regression analysis:

READ MORE TWILIGHT?		
Predictor	beta	p value
read English for fun?	0.15	0.26
seen the movie?	0.07	0.44
chapter easy?	0.27	0.16
enjoy chapter?	0.44	0.004

From: Cho (2010).

The best predictor of wanting to read more *Twilight* was NOT whether the student had seen the movie, NOT whether the student was already a pleasure reader in English, and NOT how easy the student found the chapter. The best predictor was whether the student enjoyed reading chapter one of *Twilight*. In fact, it was the only significant predictor (p = .004).

This result makes sense. Cho then performed another analysis, asking this time which were the best predictors of interest in reading in English in general. Her subjects were also asked this question: "Do you think reading *Twilight* motivated you to read in

English? (1) not at all (2) no (3) moderately (4) yes (5) a lot.").

Here are the results:

READ MORE IN ENGLISH?		
Predictor	beta	p value
read English for fun?	0.15	0.34
seen the movie?	0.11	0.23
chapter difficult?	0.07	0.16
enjoy chapter?	0.55	0.01

The results are similar to the previous multiple regression analysis: The more students enjoyed reading the first chapter of *Twilight*, the more it encouraged them to read more in English in general. The other predictors were not significant.

Syying Lee and I looked at predictors of grades in an English composition class for university students in Taiwan. Here is what we found:

predictor	Beta	p
free reading	0.26	0.04
free writing	-0.17	0.13
focus on grammar	-0.14	0.23
focus on content	0.14	0.24
apprehension	-0.41	0.001

From: Lee and Krashen (2002)

Two predictors were statistically significant: the amount of free reading students said they did, which was a positive predictor – more free reading meant higher grades, and writing apprehension, as measured by a questionnaire, which was a negative predictor – more writing apprehension meant lower grades. Students who said they focused on grammar while revising got lower grades, and students who said they focused on content when revising their essays got higher grades, but neither of these betas were statistically significant.

I close this section with an example from an area different from language education.

The following table presents predictors of high rankings of chess players, based on a sample of 158 chess players who ranged from "moderately skilled" to international grandmaster. The interesting result here is that "experience" in playing chess is not a significant predictor. This is reflected in "hours serious practice with others" and "club joining age," how old the player was when joining a chess club. Those who played more chess were not necessarily better (although clearly all of the players had played a lot of chess).

What counts among serious chess players is serious study: “hours serious analysis alone” and “chess books owned”. The same result has been found in other areas requiring a great deal of expertise, such as skating and music.

Predictors of Chess Expertise

	beta	p
Age	-0.32	0.002
hours serious analysis alone	0.56	< .0001
chess books owned	0.29	< .0001
hours serious practice with others	0.04	0.68
club joining age	-21	0.19

Charness, Krample and Mayr (1996).

Case Histories

Case histories are an important source of data. Discussions of case histories, however, have not emphasized the most important feature of using case histories: You have to have a lot of them. Only when you have a lot of them can you see what they have in common and test hypotheses. Too often, people base their opinions on just one case, their own, and do not focus on the relevant feature.

I present one collection of case histories in detail that I hope makes this point in a paper included separately, “Case Histories and the Comprehension Hypothesis.”

Sources

Charness, N., Krape, R. and Mayr, U. 1996. The role of practice and coaching in entrepreneurial skill domains: An international comparison of life-span chess skill acquisition. In K. A. Ericsson (Ed.) The Road to Excellence. Mahweh, NJ: Erlbaum.

Cho, K. (2010). Is Twilight a Home Run Book? Indonesian Journal of English Language Teaching, 4(1), 18-25.

Dziedzic, J. 2011. A comparison of TPRS and traditional instruction, both with SSR. International Journal of Foreign Language Teaching,7: 4-6.

Kim, H.Y. and Cho, K.S. 2005. The influence of first language reading on second language reading and second language acquisition. International Journal of Foreign Language Teaching 1 (4): 13-16.

Krashen, S. 2004. The Power of Reading. Portsmouth: Heineman and Westport: Libraries Unlimited.

Krashen, S., Lee, S.Y., and McQuillan, J. 2010. An Analysis of the PIRLS (2006) Data: Can The School Library Reduce the Effect of Poverty on Reading Achievement? CSLA Journal (California Association for School Librarians) 34: 26-28.

Mason, B. & Krashen, S. (1997). Extensive reading in English as a foreign language. System, 25, 91-102.

Lee, S.Y. and Krashen, S. 2002. Predictors of success in writing in English as a foreign language: reading, revision behavior, apprehension, and writing. The College Student Journal 36(4): 532-543.

St Leger A S, Cochrane A L, Moore F. (1979) Factors associated with cardiac mortality in developed countries with particular reference to the consumption of wine. Lancet: 1017-20.

Stokes, J., Krashen, S., and Kartchner, J. 1998. Factors in the acquisition of the present subjunctive in Spanish: The role of reading and study. ITL: Review of Applied Linguistics 121-122:19-25.

Varguez, K. 2009. Traditional and TPR Storytelling instruction in the Beginning High School Spanish Classroom. International Journal of Foreign Language Teaching 5 (1): 2-11.

Watson, B. 2009. A comparison of TPRS and traditional foreign language instruction at the high school level. International Journal of Foreign Language Teaching 5 (1): 21-24.