

Theories, Variables, Definitions, Hypotheses, and Measurement

Martin Kozloff

Summary of main points.

Here is what you want to see.

1. A researcher lays out a theory of the thing being studied---reading achievement, for example. What is reading? What is reading achievement? What are the variables that affect reading achievement?
2. The theory is consistent with sound scientific research. It is not a fantasy.
3. The theory identifies independent (input), dependent (outcome), and intervening variables.
4. The researcher provides conceptual (general) definitions and operational (precise) definitions for important independent (input), dependent (outcome), and intervening variables.
6. These definitions use clear terms.
7. Measures should be consistent with the definitions of the variables.
8. Measurement should be direct.
9. The researcher should measure at the proper *level* or *scale* of measurement.
10. When possible, the researcher should have several measures of the same variables.
11. Researchers should assess and report the reliability of measurement.
12. Researchers should summarize numerical data with range, mean, median, and mode.
13. Researches should report raw numbers as well as percentages, otherwise, a big percentage difference might conceal small actual numbers. “33% more projects were done by cooperative learning groups.” In fact, one

group did 3 projects The other group did 4. One more than 3 is 33%.
33% more means ONE.

14. Researchers should use proper statistical tests to report the **significance of differences** between groups or between pre-tests and post-tests.
15. Researchers should determine the **degree of association or correlation** between variables in causal research; e.g., comprehension scores as a function of children's fluency scores.

The honest and competent researcher begins with a **theory** or big picture of how he or she thinks things work. Perhaps the researcher wants to test the theory. Or perhaps the researcher is trying to fill gaps in a theory; e.g., identify additional instructional methods (variables) that will increase students' learning. In the end, the researcher will have **findings**. For example, "When teachers added ten minutes of review AFTER lessons, and ten minutes of review to the start of NEXT lessons, the percentage of students who learned new material with just ONE example rose from 50 to 75%." *Where did the findings come from?* They came from **data** that were collected. *Where did the data come from?* They came from **measurements** that were made; e.g., the researcher counted how many students got the right answer the first time. *How did the researcher know what to measure, and how to measure it?* The researcher had clear **definitions of variables**. *How did the researcher know which variables to define?* The researcher laid out a **theory** of the thing her or she was studying. The sequence is like this.

Theory → Variables → Definitions → Measures → Findings

If you read an article that does NOT cover all of these steps, then you have NO idea how the researcher ended up with the findings. And it's possible that the researcher doesn't know, either.

What is a Theory?

A theory is a set of **general statements (propositions)** about how things

(variables, or concepts) are connected. *Concepts* or *variables* are what the theory is about. The *propositions* connect the concepts, or variables. Propositions state *relationships* among the variables. A theory explains something (e.g., how students learn general ideas) by stating connections among the variables (concepts, factors) that **PRODUCE** the thing to be explained; e.g., achievement. Here is an example of a theory of learning.

Theory of the Learning Process

From Specific Events

Students Learn

General Ideas: Four Forms of Cognitive Knowledge

Input (Independent) Variables

Intervening Variables

Outcome (Dependent) Variables

Teacher Presents Examples and Treats Them the *Same Way* (e.g., names, solves, analyzes them).

+

Teacher presents Nonexamples and Treats Them a *Different Way* (e.g., names, solves, analyzes them).

+

Teacher provides *Assistance* such as Gaining attention, Review, Framing the Task, Modeling the Information, Leading Students Through the Information, Testing/Checking to Ensure Learning, Correcting

-----> The Learning Mechanism ---->

Performs a set of Logical Operations. It:

- a. *Examines* examples; *observes* their features
- b. *Compares and contrasts* examples; *identifies* features that are the same
- c. *Contrasts* examples (that share some of the same features and are treated the same way) with nonexamples (that don't have those features and are treated differently).
- d. *Identifies the differences* (in the features) between examples and nonexamples, and how they are treated.
- e. Makes a *generalization*:

Makes generalizations

- a. Verbal Association
 - (1) simple fact
X goes with Y
(Name <-> event)
 - (2) verbal chain
X goes with Y1-Y5
(New England <-> list of states)
- b. Concept
 - (1) sensory, or basic
(All defining features can be seen, heard, felt)
E.g., red, on, faster
 - (2) higher-order
(Defining features are spread out and must be synthesized)
E.g., Sandstone, justice
- c. Rule-relationship, or proposition; that is, statements that tell

Errors, Outcome
Assessment.

how concepts/variables
are related. E.g.,
“Frequent practice (one
variable) strengthens
Retention of knowledge
(another variable).”
d. Cognitive routine
(Sequences of steps
for accomplishing
task). E.g., sounding
out words, solving
math problems,
writing essays.

[Adapted from Engelmann, S., and Carnine, D. (1991). *Theory of Instruction*.
Eugene, OR: ADI Press.

**A Theory Should Identify the Important Independent/input variables,
Intervening variables, and Dependent/outcome variables.**

Notice that the theory of learning, above, lays out independent/input
variables, intervening variables, and dependent/outcome variables. On the
left are **input** variables---also called *independent* variables. These are seen as
causes of something else. What? Students’ knowledge---on the right.

Students’ knowledge is seen as the **outcome** of the effects of the independent
variables. That’s why students’ knowledge is called “dependent.”

Notice the input (independent variables). They include examples and
nonexamples that teachers use to communicate a general idea (e.g., concept).
The independent (input) variables also consist of what teachers do to gain
attention, model/present information, correct errors, etc. Please identify the
rest of the input (independent) variables....

Now look in between the input (independent) and outcome (dependent)
variables. There is a set of variables called “**intervening**” variables. In this
theory, the intervening variables are the logical operations (what students DO)
to SEE the general ideas (knowledge) revealed by the examples. In other
words, the theory states that the examples teachers use (independent
variables) **are not enough** by themselves to produce knowledge (the dependent
variables). Knowledge also requires the intervening variables of students

DOING something with the examples. This says that teachers may have to TEACH students HOW to do this. If students don't know how, then the teacher can present examples in a skillful way, but students will not be able to FIGURE out what the examples say. So, it is important that this theory lays out the intervening variables. It means that a researcher can test WAYS to teach students how to perform the operations, and then see if THAT (PLUS proper examples) increases learning.

So, the theory really says,

If teachers present a proper set of examples, and assist students to make sense of the examples,...	→ [And if students perform a set of logical operations with the examples,]...	→ Then students will learn general ideas.
[Independent variables]	[Intervening variables]	[Dependent variable]

Here is another example of a theory---how you catch a cold.

Viral dose	→ [If Weak Immune System]	→ Likelihood of Cold
[Independent variable]	[Intervening variable]	[Dependent variable]

This theory says, If you receive a sufficient dose of virus, AND your immune system is weak, then you are likely to catch a cold. In other words, the virus is NOT enough. You also need a weak immune system.

Here's another.

Tested, effective math materials	→ [If teachers are proficient]	→ Student achievement
----------------------------------	--------------------------------	-----------------------

[Independent variable]

[Intervening variable]

[Dependent variable]

This theory says, If you use tested and effective materials, AND if teachers use the materials in a proficient way, THEN students will achieve. In other words, materials alone are not enough for students to achieve. Nor is proficient teaching WITHOUT effective materials. That would be like a surgeon who is proficient but has not tools.

As a consumer, you want researchers to spell out all of the variables in their theory of how things work. What exactly are the dependent variables, the independent variables, and the intervening variables? If researchers leave out the intervening variables, it may mean that they are lazy. It also means that **they are suggesting that the input variables by themselves have the effect. This is almost NEVER the way things are.** Even catching a cold involves intervening variables. So, if the researcher concludes an article by saying, “X produced the following effects on student learning,” you KNOW that this is not the whole story. And you are warned that if you DO X (as the researcher suggests), your students are NOT likely to learn, because *there are intervening variables that you don’t know about.*

Here are some examples. Try to fill in the blanks. What do YOU think important intervening variables might be?

Teacher creates cooperate learning groups → [If teacher.....] → Accomplishment of group tasks
[Independent variables] [Intervening variables] [Dependent variable]

Teacher establishes classroom rules → [If] → Students cooperate with rules

[Independent variables] [Intervening variables] [Dependent variable]

Imagine if an author merely tells readers about how she established cooperative learning groups, and about how well the groups accomplished their tasks, but she does NOT tell you about the **intervening variables** that **made** the cooperative learning groups work. Readers might establish cooperative learning groups in their own classrooms and expect it to work; but it flops. Because the researcher did not tell readers what else had to be done.

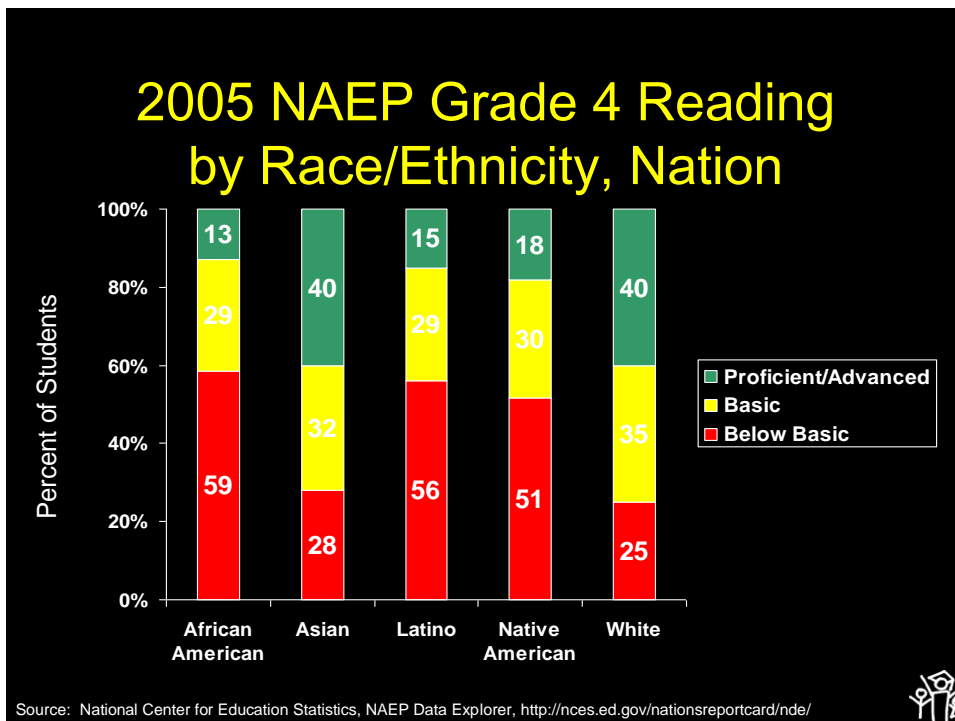
A Theory Should be Consistent with Sound Research

It is important that a theory lay out the dependent variables, the independent variables, and the intervening variables. It is also important **that the theory is derived from or is consistent with sound research**. Otherwise, it's not a theory. It's a fantasy. The above theory of learning is consistent with a large amount of research. [See references.] However, below is an example of a theory of reading that has been widely used. But this theory is **not** based on sound research.

...I offer this: Reading is a selective process. It involves partial use of available minimal language cues selected from perceptual input on the basis of the reader's expectation. As this partial information is processed, tentative decisions are made to be confirmed, rejected, or refined as reading progresses....More simply stated, reading is a psycholinguistic guessing game. It involves an interaction between thought and language. Efficient reading does not result from precise perception and identification of all the elements, but from skill in selecting the fewest, most productive cues necessary to produce guesses which are right the first time. The ability to anticipate that which will be seen, of course, is vital in reading, just as the ability to anticipate what has not yet been heard is vital in listening (Goodman, K. (1967). Reading: A psycholinguistic guess game. *Journal of the Reading Specialist*, May, 126-135. pp. 127-8).

This is the theory behind whole language. It asserts that “Efficient reading does not result from precise perception and identification of all the elements” (that is, readers do not precisely see and identify the LETTERS). Instead, they use other CUES to guess what words say. Goodman seems to admit that he is *making this theory up*. “I offer this...” And he does not cite any research to back up his claims. Yet, for 30 years, tens of thousands of teachers taught millions of children to guess at words, based on this “theory.”

Here’s the result of poor reading instruction.



What if consumers (teachers) had known that researchers must back up their “theories” with solid research? What if consumers had rejected this theory because it sounded like a fantasy---that it not only was NOT backed up by research, but was not even backed up by common sense?

Definitions of Variables

A definition is a statement that tells what a word (a name for a variable, or concept) means, or signifies, or points to. If a definition clearly tells what a variable means, then you can more easily think of how to measure the variable--measure the events that it points to. For example, if fluency (a variable) means performance that is both accurate and rapid, then to measure fluency you must measure how accurately and rapidly a person does something.

Words don't tell you what they mean. Human beings invent definitions. There are two kinds of definitions.

Conceptual definitions. Conceptual definitions are broad. They are like search lights that shine on a general area. A conceptual definition of fluency might be:

Fluency is a feature of performance: accuracy and speed.

Here is a conceptual (general) definition of decoding.

Decoding is a routine that involves translating written words into speech, using knowledge of the alphabetic principle (letters have sounds).

Notice that the conceptual definition of fluency directs your attention to two aspects of performance (accuracy and speed) and NOT to other aspects of performance, such as how independently persons performs a task, or how easily persons generalize knowledge or the performance to new situations. Likewise, the definition of decoding directs your attention to what students do when they read words, and away from things that are not part of decoding, such as guessing.

Operational definitions. Conceptual definitions are not precise enough. To create actual ways of measuring a variable, you need definitions that say EXACTLY what you would see or hear. For instance, an operational definition of fluent reading in grade 1 might be:

By the end of grade 1, the student reads grade level connected text at the rate of 60 correct words per minute.

Notice that this operational definition DOES include accuracy and speed. But it is **more precise** than the conceptual definition. It is so precise that you can think of exactly how to measure fluency:

Measuring grade 1 level connected text.

1. Present sample text.
2. The child reads the text.
3. The observer marks each error.
4. The child reads for one minute. The observer counts the number of errors and subtracts this from the total number of words read.

Likewise, here is a possible operational definition of decoding.

Decoding is a routine that consists of saying the sounds in a word, from left to right, producing a recognizable word.

Let's line up the pairs of definitions.

Conceptual

Fluency is a feature of performance:
The combination of accuracy and speed.

Decoding is a routine that involves translating written words into speech, using knowledge of the alphabetic principle (letters have sounds).

Operational

By the end of grade 1,
the student reads grade level
connected text at the rate of 60
correct words per minute.

Decoding is a routine that consists
of saying the sounds in a word,
from left to right,
producing a recognizable word.

Do you see that the operational definitions say the same thing as the conceptual definitions, but are more precise? For instance, the conceptual definition says "translating written words into speech," but the operational definition says "saying the sounds in a word, from left to right..." (a more precise way of saying translating). This is precise enough that you can measure it.

Here are examples of conceptual definitions. *Think of operational definitions for each one.* Remember, the operational definition has to say the same thing as the conceptual definition, but it is more precise; it gives examples. Also, operational definitions **depend on the situation**. For example, part of an operational definition of aggression might be hitting, but NOT if you are talking about the sport of boxing!

Conceptual definition

Aggression is behavior that is
Intended to cause injury

Operational definition

[Aggression on an elementary
school playground]

Reading fluency

Fluency is a feature of
performance:
The combination of
accuracy and speed.

Second grade reading fluency

http://reading.uoregon.edu/flu/flu_benchmarks.php

- When you evaluate research, ask:
1. Were conceptual definitions derived from or consistent with scientific research? For example, reading might be TOO NARROWLY defined as
The process of constructing meaning from text.
Is that ALL that reading is? **Comprehension alone?** Scientific research shows that reading ALSO includes knowledge of the sounds that are associated with letters (phonics); using knowledge of letter-sounds to sound out words (decoding); hearing the separate sounds in words (phonemic awareness), and vocabulary (knowing the definitions of words). <http://reading.uoregon.edu/> So, the above conceptual definition is narrow. It does not include enough of what is meant by

reading in the scientific community. Any curriculum materials, instructional methods, and assessments/measures of reading based on this NARROW definition will be INVALID.

2. Did the writer provide conceptual definitions? For example, if a writer says that “teachers were trained,” what does that mean? Trained to do what? What skills?
3. Did the writer provide operational definitions? For example, did the writer state how teachers were trained, how their learning was measured, how successful and unsuccessful performance was defined and measured? If not, then maybe different teachers were trained differently, and with different results. In other words, without operational definitions, **the word “trained” means nothing.**
4. Definitions should consist of words with clear meaning.

You saw the theory of learning above. The words are clear. Examples, compare, contrast, gain attention, etc. Here is another theory of learning. What do you think? Are the words clear?

"From this perspective, learning is a constructive building process of meaning-making that results in reflective abstractions, producing symbols within a medium." (Fosnot, C.T. (Ed.) (1996). *Constructivism : theory, perspectives, and practice*. New York : Teachers College Press. Fosnot, 1996, p. 27).

"Reflective abstraction is the driving force of learning." (Fosnot, C.T. (Ed.) (1996). *Constructivism : theory, perspectives, and practice*. New York : Teachers College Press. Fosnot, 1996, p. 29).

Do you know what Fosnot is talking about? Do you know what a “constructive building process of meaning-making that results in reflective abstractions, producing symbols within a medium” looks like? If you don’t, how could you determine whether Fosnot’s data have anything to do with her theory? Why would a person NOT write more clearly?

Measurement

Once a researcher has defined variables conceptually and operationally, the researcher can begin to select or to develop methods of measurement. There are several guidelines that should be followed.

1. Measures should be consistent with the definitions of the variables.

For example, if fluency with math problems is one outcome variable, the researcher needs to measure accuracy and speed with which students solve math problems. A measure might be the rate of correct and incorrect problems solved per minute. Likewise, if one input variable is the **faithfulness** with which teachers follow a written instructional protocol, then the researcher cannot just measure (describe) HOW teachers teach, but must measure how teachers teach **in relation to** the written protocol. The researcher would have to describe the teaching methods in the protocol AND how the teacher USES those methods.

2. Measurement should be direct.

When persons have a lung infection, they often have a fever with it. What would you want your physician to measure, to see if you are getting well: the amount of infection in your lungs, or your temperature? Temperature is an INDIRECT measure of lung infection. And it may NOT be valid. Your fever may be gone but you still have an infection. Likewise, if reading proficiency is an outcome variable, then reading proficiency (e.g., accuracy and speed of decoding, comprehension of text) is what you should measure. How much students enjoy reading, or how much they read outside of school are INDIRECT measures of reading proficiency. Students who read well are likely to enjoy reading and to read more. But these measures may not be valid.

3. The researcher should measure at the proper *level* or *scale* of measurement.

Consider the variable, color. There are four “scales” or “levels” for measuring it.

- a. You could simply take each color sample and **name** it---say the category it is in. This is called “nominal” level measurement. Think of “name.”
- b. You could **rank** each color sample from lighter to darker.

Darkest red

Dark red

Medium red

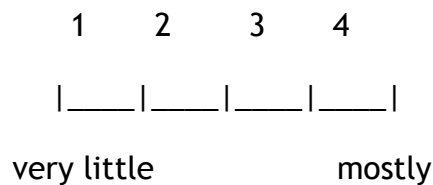
Light red

Lightest red

This is called “ordinal” level measurement. Think of order.

- c. You could use a scale of **equal intervals**.

“How much red would you say is in this fabric?”



This is called “interval level” measurement.

- d. You could use an instrument that **measures exactly how much white** is in each color sample. The instrument gives you a **number**. This number is a measure of brightness. This is called “ratio-level” measurement. One sample may have 25 white units. Another may have 50 white units. The first one has half the amount of white as the second. The ratio is 1 to 2. Ratio level.

Let’s look at each level or scale in more detail. Here are some useful websites.

<http://web.uccs.edu/lbecker/SPSS/scalemeas.htm#3>

<http://www.math.sfu.ca/~cschwarz/Stat-301/Handouts/node5.html>

<http://allpsych.com/researchmethods/measurement scales.html>

http://en.wikipedia.org/wiki/Level_of_measurement

<http://www.kimberlyswygert.com/archives/002750.html>

Again, there are four levels of measurement: nominal, ordinal, interval, and ratio. **Each next level provides more precise information than the others.**

Nominal level. The lowest level of measurement. Nominal level or nominal scale measurement implies **qualitative** (type) not quantitative (amount) differences. It refers to kinds or types of things. *Nominal measurement consists of naming or putting the things measured into categories.* For example, you could categorize students into two groups: students who receive free and reduced lunch and students who do not receive free and reduced lunch. Other examples of nominal measurement include marital status (married, divorced, separated, single), occupation, and ethnic identity.

If you are measuring some variable (e.g., error correction) on a nominal scale, you would *simply put each instance of error correction in one of several types that you had already identified.* For instance, one type might be modeling the correct answer. Another type might be explaining why the student made an error. The third type might be calling on another student to demonstrate the correct answer. After you have collected the data (put all instances of error correction in the proper categories), you would *summarize the data simply by counting the number of instances in each category.*

Data on how teacher corrected math errors during one lesson

Modeled correct answer and then tested.....12
Explained why student made error.....20
Called on another student to come to the8
board and show the correct way.

With NOMINAL data, you can

(1) Figure out how many instances are in each category.

(1) Figure out the percentage of the total that is in each category.

Model and test = $12/40 = 30\%$

Explain = $20/40 = 50\%$

Call on another = $8/40 = 20\%$

(3) Figure out the most frequent category. Explaining = 20. The most frequent category is the **mode**, or the **modal** category

Please restate the three ways that you can summarize NOMINAL data?

Ordinal level. An ordinal-scale or ordinal-level measure implies a rank order of degrees or amounts of something, but *not equal intervals* between the degrees or ranks. Probably most opinions---attitudes, perceptions and feelings---are in reality ordinal-level. *Ordinal measurement consists of placing the things measured into ranks.* For example, teachers might observe students reading and then place each student in one of three ranks:

Proficient/advanced; Basic; and Below basic. This ranking indicates differences in proficiency but, as with nominal measurement, *it does not give precise information* (such as how many correct words students read per minute). *Also the differences between the ranks are not necessarily equal.* That is, the difference in proficiency between Below basic and Basic, and between Basic and Proficient/advanced may not be equal. The difference in proficiency between Basic and Proficient/advanced may be far greater than the difference in proficiency between Below basic and Basic. ***This is why you cannot give a number to each rank, and then add up the rank scores (2, 3, 3, 2, 2, 2, 1, 1, 3, 3, 2, 2) and then divide by the number of scores (12) and find the average rank! Because the distances between the ranks are not equal. The NUMBER of a rank is NOT a numerical VALUE.*** It is nothing more than the NAME of a rank. So, if you measure things by giving their rank order (e.g., you assign each student the rank Proficient/advanced; Basic; or Below basic), you summarize the data by simply

(1) Figuring out how many students are in each rank and then perhaps figuring out the **percentage** of the total number that is each rank. For example, there are 12 students.

Proficient/advanced = 4 = 33%

Basic = 6 = 50%

Below basic = 2 = 17%

If you then use a better reading program, you hope that the DISTRIBUTION of ranks changes.

Proficient/advanced = 4 = 33%

Basic = 8 = 67%

Below basic = 0

- (2) Figuring out the most frequent rank, or the **mode**. Which, above, is “Basic.”
- (3) Figuring out the rank that is in the middle---about 50% of scores are above and below it. Here are the data from above. 1 = Below basic; 2 = Basic; 3 = Proficient/advanced.

1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3

The middle of the distribution is 2. **This is called the median.**

Here is another distribution. Income for nine persons.

\$100,000

\$60,000

\$60,000

\$20,000

\$20,000

\$20,000

\$10,000

\$8,000

\$8,000

What is the middle score—about half are above and half are below it? **\$20,000**

Interval level. Interval level measurement is the kind of information provided by thermometers. There are a series of intervals (e.g., degrees) that **are equal**, and there is no true zero (there is no such thing as zero

temperature). Interval level measurement is often provided by **rating scales** that ask persons to answer questions such as:

“Place an X in the spot that best represents how teacher-friendly (that is, well-organized, lots of instructions, easy to use) your new math materials are.”

1	2	3	4
____	____	____	____
Less friendly			More friendly

Or, “How much do you agree with the following statement? “Our school provides timely and adequate supervision and assistance?”

1. Strongly agree.
2. Agree
3. Disagree
4. Strongly disagree.

When it is assumed that the intervals are equal, it is then okay to summarize scores by calculated the **mean, or average**. *You add the scores and divide by the number of scores.*

For instance, here are the scores of 10 persons on the above question.

3 persons gave a rating of 3, or $3 \times 3 = 9$.

4 persons gave a rating of 2, or $2 \times 4 = 8$

3 persons gave a score of 1, or $1 \times 3 = 3$

Total score = $9 + 8 + 3 = 20$. 20 total divided by 10 scores = 2. The average or mean score is 2.

Ratio level. Ratio-level or ratio-scale measurement is real numbers. There can be true zero (e.g., zero episodes of aggression occurred; zero

income). In addition, there are **equal intervals between quantities**; e.g., the difference between 0 and 1, 1 and 2, etc., is 1.

Ratio level measurement is the most precise. It provides information on the **number** of times (e.g., number of questions answered correctly), or the **rate** (e.g., number of words read correctly per minute), or **percentage** of times (e.g., the percentage of errors teachers correct) that something happens. Ratio level information is usually provided through direct observation or through tests that enable the observer to count instances of identified variables (e.g., correct answers).

With ratio-level measures you can do many operations to summarize data. Here are data on reading fluency.

Billy = reading 100 correct words per minute

Sam = reading 90 correct words per minute

Slim = reading 90 correct words per minute

Darren = reading 110 correct words per minute

Nancy = 80 correct words per minute

Terri = 90 correct words per minute

Tim = 95 correct words per minute

(1) Figure out the mode, or most frequent score. 90.

(2) Figure out the median, or the middle score.

80, 90, 90, 90, 95, 100, 110 = 90 (3 scores are above and 3 are below 90)

(3) Figure out the mean, or average.

$80 + 90 + 90 + 90 + 95 + 100 + 110 = 655$, divided by 7 scores = 93 mean or average score

4. Figure out percentages. For example, if the mean fluency when the teacher used “Phud Phonics” was 93 correct words per minute, and the mean fluency **after** the teacher used a new reading program (“Fluent Phonics”) for

three months rose to 100 correct words per minute, what is the percentage increase?

From 93 to 100 = increase of 7

What percentage of 93 is 7?

$7/93 =$ approximately 8%

Going from a mean of 93 to a mean of 100 is an increase of about 8%.

A few cautionary comments

- 1. You *can* use a lower-level scale for measuring a variable that could be measured on a higher level, but you lose information.** For example, you can measure fluency on a nominal scale by categorizing each student as either “Rapid,” “Moderately fast,” or “Slow.” But this means that several students that are in the same category could actually have different EXACT fluency rates. You might treat these students the same (e.g., put them in the same reading groups based on their nominal category), when they are actually different. It also means that you don’t know EXACTLY how many words students read correctly per minute. Therefore, **it is best to use the highest (more precise) level of measurement that you can.**
- 2. However, you CANNOT (!!!) use a higher-level scale to measure a variable that is *really* on a lower scale.** For example, the three different methods of error correction (above) are just categories. The categories do NOT imply differences in the amount or quantity of anything. Therefore, you cannot give each category of error correction a number.....

Model correct answer is 1

Explain error is 2

Another student demonstrates is 3.

And then add up the number of 1’s, 2’s, and 3’s.....

Model and test = 12 $12 \times 1 = 12$

Explain = 20 $20 \times 2 = 40$

Call on another = 8 $8 \times 3 = 24$

And then figure out the mean....

$12 + 40 + 24 = 76$ 76 divided by 40 scores = 1.9 = average or mean error correction.

This makes no sense at all. The different kinds of error correction are not WORTH any points. Explaining (a 3) is not worth 3 times modeling (a 1).

These numbers are no more than names.

4. When possible, the researcher should have several measures of the same variables.

This is called “triangulation.” The idea is, if different measures say much the same thing, you can have greater confidence in the validity of the finding. For instance, a researcher might give students mastery tests every 10 lessons on a math program. The tests are based on curriculum materials that were covered. At the end of the semester, the researcher also gives students a standardized test on math. If the curriculum based measures and the standardized test (that has different kinds of items on it) both say that students have learned the material, then you can have more confidence in the findings than if you had only one measure.

Here are resources on standardized tests.

<http://www.ncrel.org/sdrs/areas/issues/students/earlycld/ea5lk3.htm>

<http://nces.ed.gov/nationsreportcard/about/>

http://en.wikipedia.org/wiki/Standardized_testing

http://www.sizes.com/society/test_scores.htm

5. Researchers should assess and report the reliability of measurement.

Observers and testers should be trained ahead of time to follow a testing or observing protocol---steps on exactly what to do. They should be observed testing or observing, and coached to use the protocol faithfully. Scores of the SAME observer or tester scoring the same thing several times should be

compared to see how much the two sets of scores agree. This is called **intra-observer** (within the same observer) **reliability**. Also, different observers or testers scoring the same thing should be compared---again to see how closely they agree. This is called **inter-observer** (between observer) **reliability**. If reliability (agreement) is below 90%, then either observers and testers need more training, or the definitions of variables need to be clearer (maybe they disagree because the definitions are vague), and the protocols need to be made easier or clearer. Researchers should describe how they trained observers and testers, and how they assessed reliability. **If not, the consumer has no way to tell if the scores are valid.**

Analyzing Statistical Data

Please examine the entries in “Vocabulary” for mean, median, mode, plot on a graph, relationship, and trend

Let’s say authors are reporting **survey research** of schools that used one of two kinds of math programs. Program A (there were several versions) taught all of the elementary math concepts and operations before it had students apply these skills to word problems. Program B (there were several versions) focused on word problems, and taught students the relevant math concepts and operations at the same time. The authors believe that Program A will yield higher achievement. So, they divide the schools in the district into schools that use Program A vs. Program B, and they also collect information using district official statistics on the percentage of students who pass end of grade tests (as an outcome measure of achievement.) The authors report, “In general, students who received Program A achieved significantly more than students who received Program B.”

Will you use Program A? Will you avoid Program B?

The authors don’t tell you what it means that “students who received Program A achieved significantly more than students who received Program B.” They are leaving out essential statistical information.

Summary statistics

For each class in each school that used Program A and Program B, you want to know:

1. **The average score---the mean.** The sum of all of the scores in a class divided by the number of scores.

$$\frac{65+ 69 + 70 + 75 + 78 + 80 + 80 + 87+ 93 + 93+ 96}{11} = \frac{886}{11} = 80.5$$

2. Notice that the mean is 80.5, but scores range from 65 to 96. **Range is another statistic to present.** Shouldn't consumers know that a program can produce a WIDE range of scores? Wouldn't you want to know if a medication produced wide range of effects?

3. **The most frequent score.** This is the **mode**, or **modal** score. What is the modal score?... 80. Again think of medicine. Can you imagine asking your physician, "What is the **most likely** outcome?" Of course.

4. **The middle score.** This is the **median**. This is an important statistic. It tells you which score is about half way in the **distribution** (spread) of scores. What is the median score from the above distribution? **80**. Why is it important? Well, imagine that five students scored in the 90s. These scores make the mean or average pretty high. If the mean were the **only** statistic you had, you might think that the class as a **whole** did well. But what if the middle score was 80? Half of the class got lower scores than 80. So, **the median tells you not to be fooled by a high mean that is really the result of a few very high scores, or vice versa.**

5. **Percentages or ratios vs. whole numbers.** Do not be satisfied if a researcher reports percentages or ratios but not the whole numbers—raw numbers. One researcher reported that students who received a certain pre-school program (vs. a different pre-school program) were **twice as likely**---two decades later--to have been arrested for felonies. **Many readers were completely fooled by this statistic.**

"Boy, I'm never going to use THAT program. It makes kids twice as likely to become criminals!"

Sure, that's how it looks if you only report percentages and ratios (twice as likely). But what if you found out that after 20 years there were **only three persons left in the samples** for each pre-school program? And what if "twice as likely" means that in one pre-school sample, ONE adult had been arrested for felony, and in the other pre-school sample, TWO adults had been arrested. In other words, percentage-wise, the difference is 200 percent. But in terms of whole numbers or raw numbers, we are talking about ONE person. Do you think THAT is significant? Could it just as easily be a difference of ONE arrest in the OTHER sample? Of course. So, if authors do NOT report the raw numbers, you have NO idea if the percentages and ratios are meaningful. **200 percent more of WHAT? One!?**

Statistical Significance

In the survey, above, the researchers collected data on student achievement when students used one of two kinds of math programs. They report that "students who received Program A achieved significantly more than students who received Program B." We wondered what that meant. The researchers told us PART of what that meant by giving us summary statistics for each class: the range of scores, the mean score, the modal score, and the median score. The researchers' claim, remember, is that the mean, median, and modal scores for students in Program A are significantly higher than the mean, median, and modal scores of students in Program B. But what does "significantly higher" mean? Significance means two things: practical and statistical.

Practical significance. You join a program to lose some weight.

"I can stand to lose a few pounds. I have to walk sideways through the doorway."

So, you join Whale Watchers. You pay 100 dollars a month for advice, feedback, encouragement, and menus. At the end of one year, you have lost 10 pounds!

Wow!

1200 dollars.

10 pounds.

Would you say that the result is of **practical significance**?

Can you walk straight through a doorway?

Can you fit into your swim suit?

Can you see your feet?

NO?

So, 10 pounds is not of practical significance.

Statistical significance. But what if almost everyone in Whale Watchers (thousands of persons) lost from between 5 and 15 pounds? What are the odds of that, if Whale Watchers did not work? What are the odds that so many persons losing weight--even if it is only a little weight---is a fluke, random, chance? *That is what statistical significance is about.* If you have large samples, even small but consistent differences between the samples on some outcome measure are probably statistically significant---NOT likely to be the result of chance.

[However, small differences may not be practically significant. Would you change an entire reading program just because program C produces on average 2 points higher achievement?]

At the same time, **with small samples, it takes larger differences for the differences to be statistically significant.** Imagine two weight loss programs. Whale Watchers and Pie Anonymous. Imagine that there are five persons in each group. At the end of the year, the mean weight loss in the five Whale Watcher clients was 6 pounds, and the mean weight loss in the five Pie Anonymous clients was 7 pounds, or 8 pounds, or even 9 pounds? Do you think those differences COULD NOT EASILY be the result of CHANCE? Of course they could be chance. Imagine you did the study again. Do you think you would get the same outcomes? No, sorry, **a small difference between small samples is NOT statistically significant.** With only samples of five persons, you could easily get small differences by chance.

There are many kinds of tests of statistical significance. It depends on the kind of data you have---nominal, ordinal, interval, or ratio. **Basically, the test**

tells you the chances that results could be chance. For example, a test might say, $p = < .05$. This means that the chances of getting the scores you got (e.g., the differences in the achievement scores for one group vs. another group) are less than 5 in a hundred. The question is, CAN you live with that? Is it okay to be wrong 5 out of 100 times? Would 95% confidence that the effect of a drug was real and not chance be high enough for you? How about the effect of a reading program? It would probably be satisfactory to have statistical significance at the .05 level. After all, you are only going to put the new program in once. The odds are 95 out of 100 in your favor. But if you used the program in 100 school districts, the results might be chance---not the result of the program---5 times.

Here are some resources on statistical tests.

http://www.graphpad.com/articles/interpret/principles/stat_principles.htm

Citation: H.J. Motulsky, *Analyzing Data with GraphPad Prism*, 1999, GraphPad Software Inc., San Diego CA, www.graphpad.com.

<http://www.itl.nist.gov/div898/handbook/prc/section1/prc13.htm>

<http://www.surveysystem.com/signif.htm>

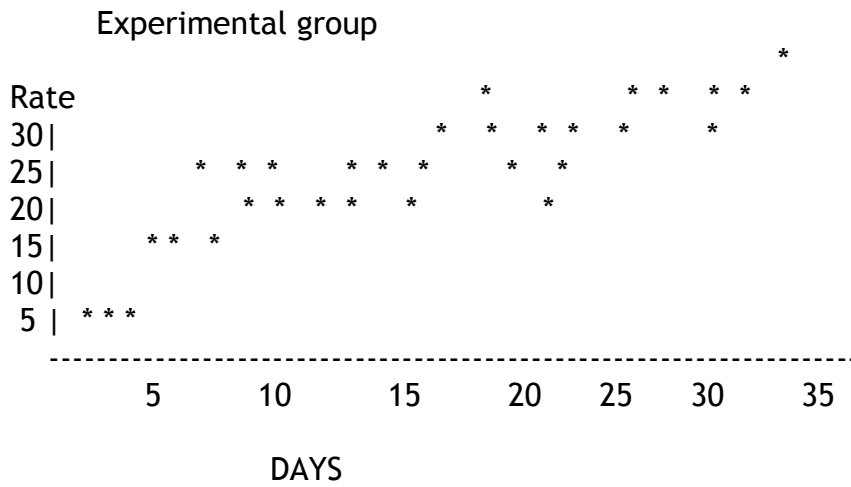
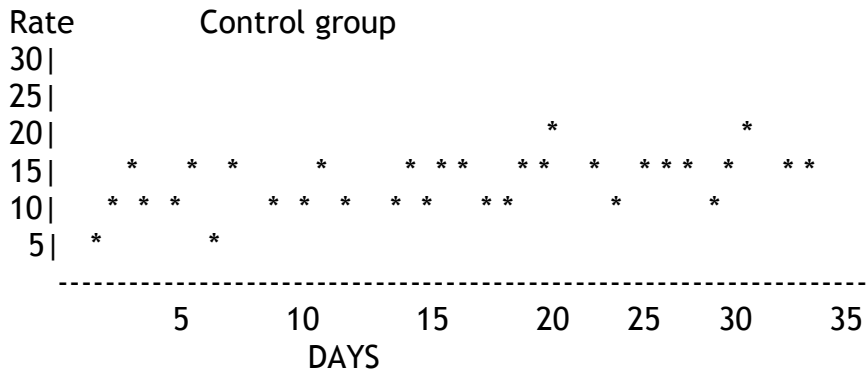
http://en.wikipedia.org/wiki/Statistical_significance

<http://www.statpac.com/surveys/statistical-significance.htm>

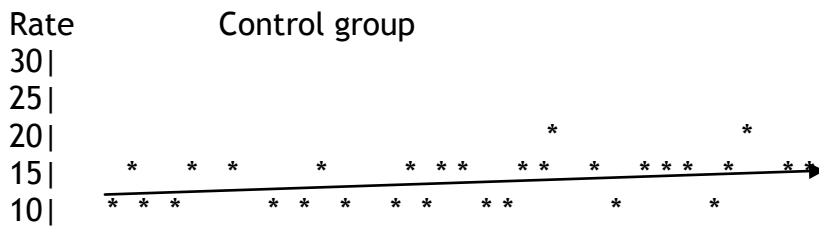
Correlation

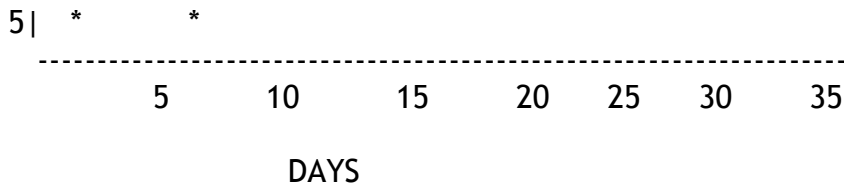
Let's say you are doing research that is looking to see IF there is a causal relationship between, say, how often teachers provide immediate and specific praise (input, independent variable, intervention), and the rate at which students give correct responses (outcome, dependent variable, effect). You have a pool of 50 children in fourth grade. The 50 children are assigned at random to two classes: Experimental group (teacher gives immediate and specific praise---"I love the way you answered with a full sentence!"---after almost every correct response); Control group (teacher gives delayed, general praise after one out of four correct responses. "Good job.").

Here are the data.

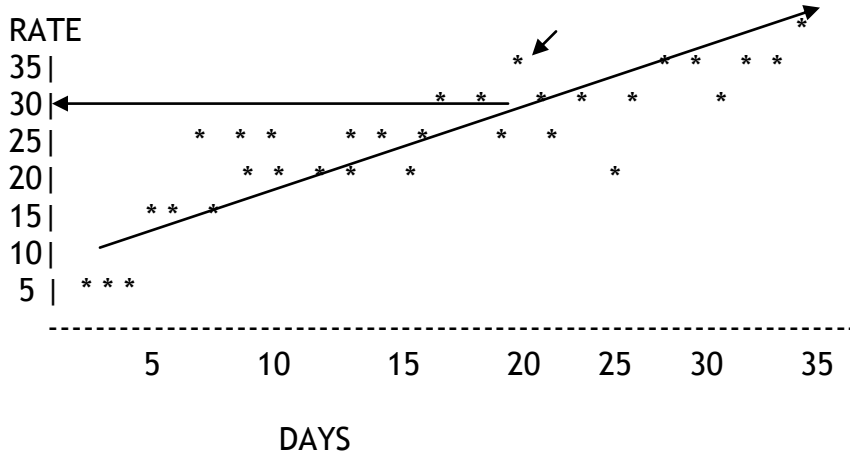


Let's draw a best fit line through the data points.





Experimental group



Notice that there is almost no change in the rate of correct responses in the control group. The rate begins at around 12 correct responses per lesson, and ends at around 15 correct responses 35 days later. However, **there IS an increase in the rate of correct responses in the Experimental group.** The group began with 6 correct responses per lesson and rose to about 35 correct responses per lesson 35 days later. **But how STRONG is the relationship between timely, specific praise and correct responses?** How accurately does the number of days students receive timely and specific praise predict the number of correct responses on that day? Well, look at the plotted data for the experimental group. If the correlation between praise and correct responses (if the prediction of correct responses from knowledge of days of praise) was perfect (100% accurate), then all of the data points would be right on the best fit line. But they aren't. This means that if on Day 20, you

predicted 30 correct responses (as the line says), you would be off by 5 responses. The actual number of correct responses on Day 20 was 35. Check some of the other data points. What does the line predict for a day, and what is the actual number for that day?

So, does knowing the day enable you to predict better than if you pulled a number out of a hat? Yes. Why? Because there IS an association (correlation = CO-relation) between days of praise and correct responses.

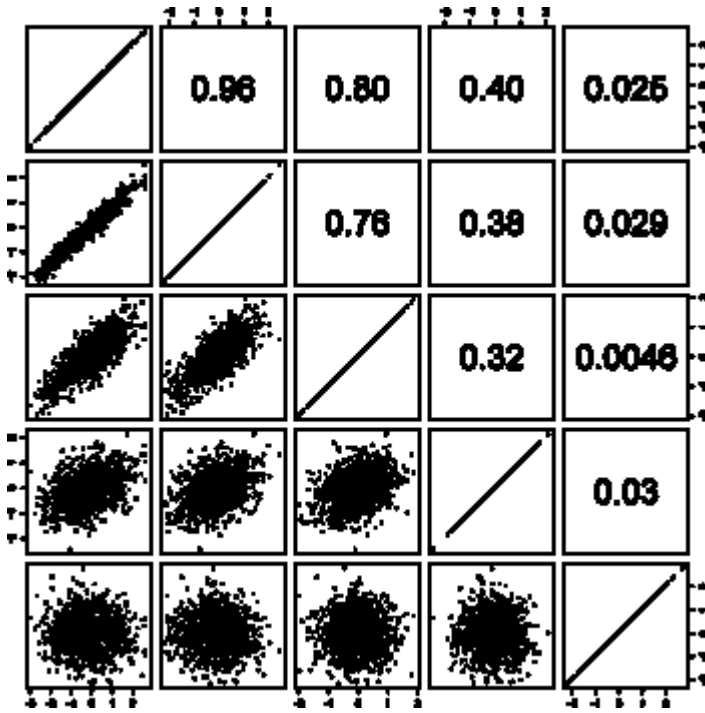
There are statistical techniques that tell you just how strong the relationship is. The number they give you is the “correlation coefficient.” The table, below, is from Wikipedia. It shows the shape of a line, and it shows data points around the line. The numbers to the right are the correlation coefficient. For instance, the top left plot shows the data points almost right on the line. This means that the correlation between one variable and the other is very high = .96.

In the second line down, the correlation coefficient is .76. Notice that there is more variation. The same spot on the across axis is associated with several values on the up axis.

The correlation in the third row down is even weaker. Notice that any value along the across axis is associated with MANY values along the up axis. The coefficient is .32.

In the fourth row down, there is hardly any association at all. And the coefficient is .03.

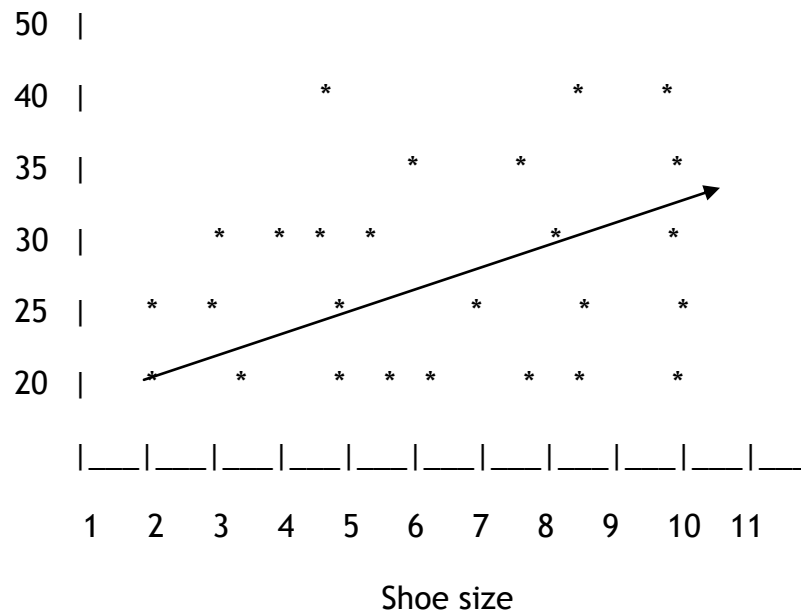
The fifth line shows zero relationship. Knowing the value on the across axis does not give you any information about what the values on the up axis might be.



<http://en.wikipedia.org/wiki/Correlation>

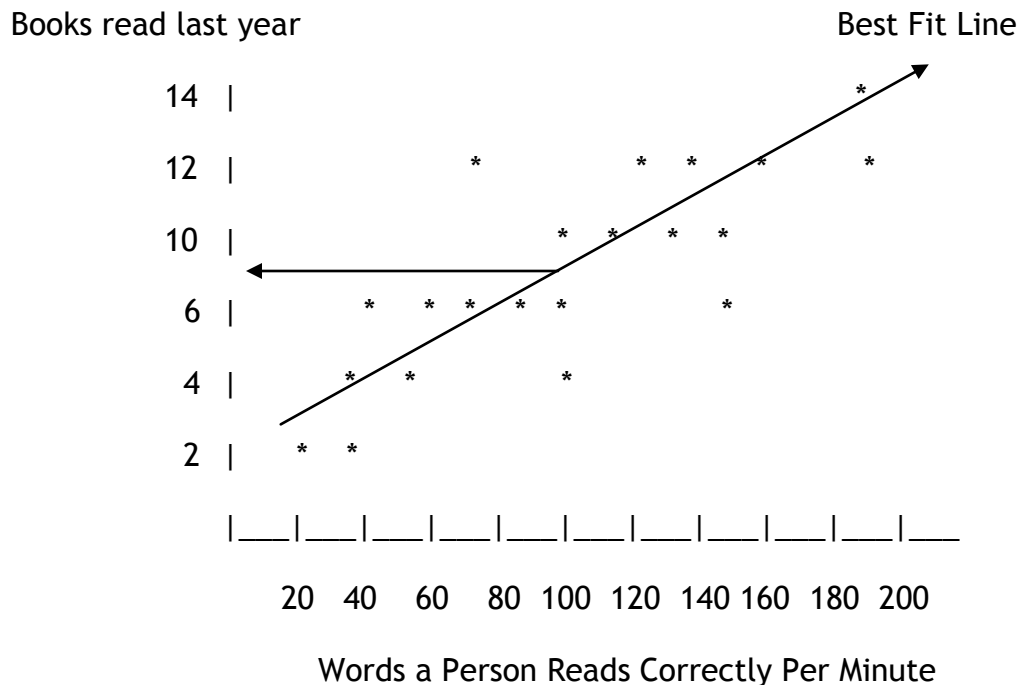
Remember this graph, below?

Books read last year



Is there a trend here? Yes, people with tiny feet (infants) don't read much. And when people get older—and their feet get bigger---they read more. But some people with big feet hardly read any books. So, how well does shoe size predict the number of books a person reads? **How strong is the association?** Look at the table above from Wikipedia. The plot above looks like the fourth row on the Wikipedia table. A correlation coefficient of .03. Almost nothing.

Here's another graph.



It shows data for 21 teenagers. We know two things about each person: how many books they read last year and how many words they read correctly per minute (reading fluency). So, if you look at the bottom left corner, it PLOTS the data for one person. He reads 20 correct words per minute (very slow) and he read 2 books in a year.

Now look at the right side of the graph. Two persons read at a rate of 200 correct words per minute; one read 12 books and the other read 14 books.

Do you see a trend? For example, **does the number of books per year change as the fluency increases?** Yes. You can see that the higher the fluency, the more books persons read. Fluency IS correlated with, and it predicts, the number of books read.

Notice that the best fit line does NOT connect the plotted data points. **It cuts through them so that there are about as many above it as below it.** Pick a value along the across (input, predictor) axis. Say, 100 words per minute. The best fit line predicts that persons reading at 100 words per minute will read how many books?... (See arrow)... 8 books. Now how many

books did our teenagers reading at 100 words per minute actually read?... 4, 6, and 10. We predict 8, but we get a range from 4 to 10. This is PRETTY strong. Check the Wikipedia table. Which plot does our book plot look like?... I'd day the third row down. The correlation coefficient is .32.

Here are more resources on correlation.

<http://www.neatideas.com/cc.htm>

<http://www.surveysystem.com/correlation.htm>

<http://ssed.gsfc.nasa.gov/lepedu/IA-CorrCoeff.html>

http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

Please review the main points made at the beginning of this document.

References

- Anderson, J.R., Reder, L.M., & Simon, H.A. (1998). Applications and Misapplications of Cognitive Psychology to Mathematics Education. Department of Psychology. Carnegie Mellon University. Pittsburgh, PA 15213.
- Binder, C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst*, 19, 163-197.
- Brophy, J.E., & Good, T.L. (1986). Teacher behavior and student achievement. In M.C. Witrock (Ed.), *Third handbook of research on teaching* (pp. 328-375). New York: McMillan.
- Carnine, D. W. (1976). *Correction effects on academic performance during small group instruction*. Unpublished manuscript. Eugene, OR: University of Oregon Follow Through Project.
- Dixon, R. (1989). Sequences of Instruction. University of Oregon.
- Dixon, R.C. , & Carnine, D. (1993). Using scaffolding to teach writing. *Educational Leadership*, 51 (3), 100-101.
- Dougherty, K.M., & Johnston, J.M. (1996). Overlearning, fluency, and automaticity. *The Behavior Analyst*, 19, 289-292.
- Ehri, L.C. (1998).

- Grapheme-phoneme knowledge is essential for learning to read words in English. In J. Metsala & L. Ehri (Eds.), *Word recognition in beginning reading* (pp. 3-40). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Ellis, E.S., & Worthington, L.A. (1994). *Research synthesis on effective teaching principles and the design of quality tools for educators*. University of Oregon: National Center to Improve the Tools of Educators.
- Engelmann, S. (1999). Student program alignment and teaching to mastery. National Direct Instruction Conference. Eugene, Oregon.
- Englert, C.S., Raphael, T.E., Anderson, L.M., Anthony, H.M., & Stevens, D.D. (1991). Making strategies and self-talk visible: Writing instruction in regular and special education classrooms. *American Educational Research Journal*, 2, 337-372.
- Greenwood, C.R., Delquadri, J., & Hall, R.V. (1984). Opportunity to respond and student academic performance. In W.L. Heward, T.E. Heron, J. Trap-Porter, & D.S. Hill (Eds.), *Focus on behavior analysis in education*. Columbus, OH: Merrill.
- Grossen, B.J., Carnine, D.W., Romance, N.R., & Vitale, M.R. (1998). Effective strategies for teaching science, in E.J. Kameenui & D.W. Carnine (Eds.), *Effective teaching strategies that accomodate diverse learners*, pp. 113-137. Columbus, OH. Merrill.
- Gunter, P.L., Hummel, J.H., & Conroy, M.A. (1998). Increasing correct academic responding: An effective intervention strategy to decrease behavior problems. *Effective School Practices*, 17, 2, 55-62.
- Rosenshein, B., & Meister, C. (1992). The use of scaffolds for teaching higher-order cognitive strategies. *Educational Leadership*, 49 (7), 26-33.
- Rosenshine, B., & Stevens, R. (1986). Teaching functions. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (Third edition) (pp. 376-391). New York: McMillan..
- Rosenshine, B. (1986). Synthesis of research on explicit teaching. *Educational Leadership*, 43, 60-69.

Walberg, H.J. (1990). Productive teaching and instruction: Assessing the knowledge base. *Phi Delta Kappan*, February, 470-478.