Psychology 205: Research Methods in Psychology
Issues in measurement

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

April, 2021

Outline

Thought problems

Numbers

- Assigning Numbers to Observations
- Difficulties and artifacts of scaling

Preliminaries

Validity $\neq$ Reliability

Two broad classes of validity

- Internal Validity
- External Validty

Reliability

- Classical test theory
- Coefficients based upon the internal structure of a test

Thought problems

Friendly Dolphins help fishermen?

1. A recent case in the news reports how a fisherman fell off his boat but was rescued when a dolphin pushed him to shore. Several other fishermen confirmed that this happened to them as well.
2. From these stories, should we conclude that dolphin are friendly to humans and help them when they are in distress?
3. What piece of evidence is missing from these stories?

Mortality statistics

1. In 1835 the Swiss physician H. C. Lombard published the results of a study on the longevity of various professions. His data were very extensive, consisting of 8,496 death certificates gathered over more than a half century in Geneva. Each certificate contained the name of the deceased, his profession, and age at death.
2. Lombard used these data to calculate the mean longevity associated with each profession. Lombard's methodology was not original with him but instead was merely an extension of a study carried out by R. R. Madden, Esq., published 2 years earlier. Lombard found that the average age of death for the various professions ranged principally from the 40s to the mid 70s. Those were somewhat younger than those found by Madden, but this result was expected, because Lombard was dealing with ordinary people rather than the "geniuses" in Madden's study (the positive correlation between fame and longevity was well known even then).

Being a student is the riskiest profession!

Wainer (1999) reviews data from the Swiss physician H.C. Lombard who examined 8,496 death certificates gathered over a half century in Geneva. Each certificate contained the name of the deceased, his profession, and age at death. Lombard used these data to calculate the mean longevity associated with each profession. Consider the following (abbreviated) table.

Table: Age at death by occupation: data from Lombard (from Wainer)

Profession	Total Number of Deaths	Average Age at death
Students	39	20.2
Merchant assistants	58	38.9
Coachmen	12	45.0
Soldiers	338	48.4
Bakers	82	49.8
Butchers	77	53.0
Surgeons	41	54.0
Farmers	267	54.7
Wine merchant	120	56.3
Businessmen	7	57.5
Harness Makers	10	60.4
Lawyers	12	64.3
Apothecaries	19	64.3
School masters	18	64.4
Professors	10	66.6

Explain how this is a problem of sampling.

Where to place armor in airplanes

1. In World War II, many planes would return to base with bullet holes in them.
2. The question became where to add extra armor to protect them.
3. George Wald (1980) recorded where returning planes had been shot.
4. The question became where to put the extra armor.

Table: Probability of plane being hit at a location

Part	% Area	% hits
Entire plane	100	100
Engines	.27	.19
Fueselage	.37	.39
Fuel System	.15	.15
Other parts	.23	.27

Where to place armor in airplanes

1. In World War II, many planes would return to base with bullet holes in them.
2. The question became where to add extra armor to protect them.
3. George Wald (1980) recorded where returning planes had been shot.
4. The question became where to put the extra armor.

Table: Probability of plane surviving a single hit

Part	% Area	% hits	p(surviving)	p(downed)
Entire plane	100	100	.85	.15
Engines	.27	.19	.61	.39
Fueselage	.37	.39	.95	.05
Fuel System	.15	.15	.85	.15
Other parts	.23	.27	.98	.02

Where to place extra armor: The case of George Wald and the problem of missing data

An outline of a plane

A depiction of a plane with shading indicating where returning planes had been shot.

(Wainer, 1990, 1999) (Wald, 1980)

Assigning numbers to observations

Although seemingly easy, the assigning of numbers is more complicated than it appears.

Table: Numbers without context are meaningless. What do these number represent? Which of these numbers represent the same thing?

2.718281828459045 0908	3.14159265358979 3116
24	86,400
37	98.7
365.25	365.25636305
31,557,600	31,558,150
3,412.1416	.4046856422
299,792,458	$6.022141 * 10^{23}$
42	X

What is the "average" class size?

Table: Average class size depends upon point of view. For the faculty members, the median of 10 is very appealing. From the Dean's perspective, that the faculty members teach an average of 50 students per class is great. But what about the students? What do they experience?

Faculty Member	Freshman/ Sophmore	Junior	Senior	Graduate	Mean	Median
A	20	10	10	10	12.5	10
B	20	10	10	10	12.5	10
C	20	10	10	10	12.5	10
D	20	100	10	10	35.0	15
E	200	100	400	10	177.5	150
Total	280	230	440	50	1000	
Mean	56	46	108	10	50.0	39
Median	20	10	10	10	12.5	10

Class size from the students' point of view.

Table: Class size from the students' point of view. Most students are in large classes; the median class size is 200 with a mean of 223.

Class size	Number of classes	number of students
10	12	120
20	4	80
100	2	200
200	1	200
400	1	400

Time in therapy

A psychotherapist is asked what is the average length of time that a patient is in therapy. This seems to be an easy question, for of the 20 patients, 19 have been in therapy for between 6 and 18 months (with a median of 12) and one has just started. Thus, the median client is in therapy for 52 weeks with an average (in weeks) $(1 * 1 + 19 * 52)/20$ or 49.4.

However, a more careful analysis examines the case load over a year and discovers that indeed, 19 patients have a median time in treatment of 52 weeks, but that each week the therapist is also seeing a new client for just one session. That is, over the year, the therapist sees 52 patients for 1 week and 19 for a median of 52 weeks. Thus, the median client is in therapy for 1 week and the average client is in therapy of $(52 * 1 + 19 * 52)/(52+19) =$ 14.6 weeks.

Tournaments to order people (or teams)

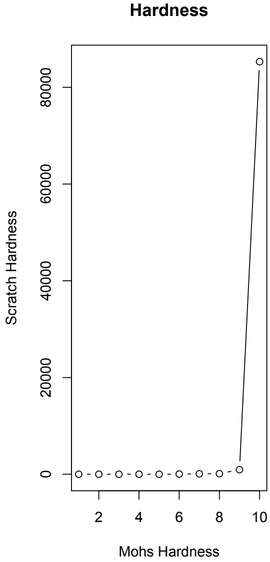
1. Goal is to order the players by outcome to predict future outcomes
2. Complete Round Robin comparisons
    - Everyone plays everyone
    - Requires $N * (N-1)/2$ matches
    - How do you scale the results?
3. Partial Tournaments – Seeding and group play
    - World Cup
    - NCAA basketball
    - Is the winner really the best?
    - Can you predict other matches

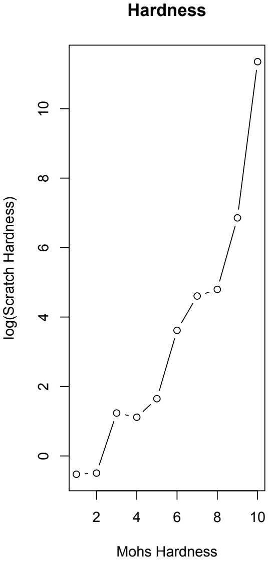
Moh's hardness scale provides rank orders of hardness

Table: Mohs' scale of mineral hardness. An object is said to be harder than X if it scratches X. Also included are measures of relative hardness using a sclerometer (for the hardest of the planes if there is a ansiotropy or variation between the planes) which shows the non-linearity of the Mohs scale (Burchard, 2004).

Mohs Hardness	Mineral	Scratch hardness
1	Talc	.59
2	Gypsum	.61
3	Calcite	3.44
4	Fluorite	3.05
5	Apaptite	5.2
6	Orthoclase Feldspar	37.2
7	Quartz	100
8	Topaz	121
9	Corundum	949
10	Diamond	85,300

Measuring Hardness – Scratch versus Mohs



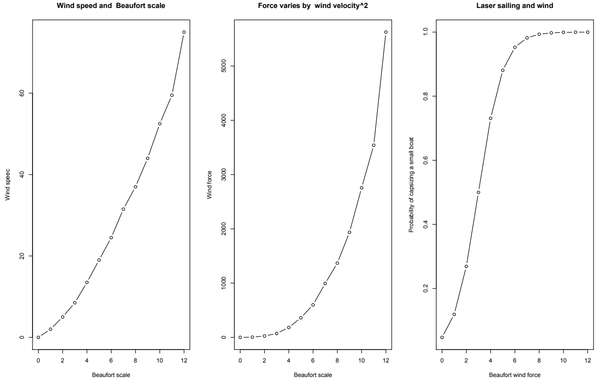


Ordering based upon external measures

Table: The Beaufort scale of wind intensity is an early example of a scale with roughly equal units that is observationally based. Although the units are roughly in equal steps of wind speed in nautical miles/hour (knots), the force of the wind is not linear with this scale, but rather varies as the square of the velocity.

Force	Wind (Knots)	WMO Classification	Appearance of Wind Effects
0	Less than 1	Calm	Sea surface smooth and mirror-like
1	1-3	Light Air	Scaly ripples, no foam crests
2	4-6	Light Breeze	Small wavelets, crests glassy, no breaking
3	7-10	Gentle Breeze	Large wavelets, crests begin to break, scattered whitecaps
4	11-16	Moderate Breeze	Small waves 1-4 ft. becoming longer, numerous whitecaps
5	17-21	Fresh Breeze	Moderate waves 4-8 ft taking longer form, many whitecaps, some spray
6	22-27	Strong Breeze	Larger waves 8-13 ft, whitecaps common more spray
7	28-33	Near Gale	Sea heaps up, waves 13-20 ft, white foam streaks off breakers
8	34-40	Gale Moderately	high (13-20 ft) waves of greater length, edges of crests begin to break into spindrift, foam blown in streaks
9	41-47	Strong Gale	High waves (20 ft), sea begins to roll, dense streaks of foam, spray may reduce visibility
10	48-55	Storm	Very high waves (20-30 ft) with overhanging crests, sea white with densely blown foam, heavy rolling, lowered visibility
11	56-63	Violent Storm	Exceptionally high (30-45 ft) waves, foam patches cover sea, visibility more reduced
12	64+	Hurricane	Air filled with foam, waves over 45 ft, sea completely white with driving spray, visibility greatly reduced

The Beaufort scale is non-linear with force or probability of capsizing



Scaling of Objects: O x O comparisons

1. Typical object scaling is concerned with order or location of objects
2. Subjects are assumed to be random replicates of each other, differing only as a source of noise
3. Absolute scaling techniques
   - Grant Proposals: 1 to 5
   - "On a scale from 1 to 10" this [object] is a X?
   - If A is 1 and B is 10, then what is C?
   - College rankings based upon selectivity
   - College rankings based upon "yield"
   - Zagat ratings of restaurants
   - A - F grading of papers

Absolute scaling: difficulties

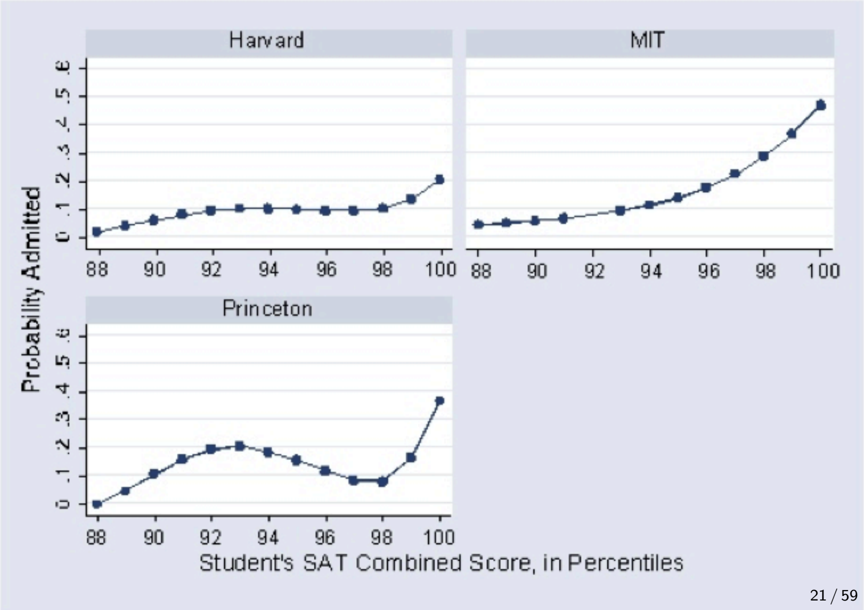
1. "On a scale from 1 to 10" this [object] is a X?
   - sensitive to context effects
   - what if a new object appears?
   - Need unbounded scale
2. If A is 1 and B is 10, then what is C?
   - results will depend upon A, B

Absolute scaling: artifacts

1. College rankings based upon selectivity
   - accept/applied
   - encourage less able to apply
2. College rankings based upon "yield"
   - matriculate/accepted
   - early admissions guarantee matriculation
   - don't accept students who will not attend
3. Proposed solution: college choice as a tournament
   - Consider all schools that accept a student
   - Which school does he/she choose?

Avery, Glickman, Hoxby & Metrick (2013)

A revealed preference ordering Avery et al. (2013)



A revealed preference ordering Avery et al. (2013)

A REVEALED PREFERENCE RANKING OF COLLEGES BASED ON MATRICULATION DECISIONS

Rank Based on Matriculation (with Covariates)	College Name	Theta	Implied Prob. of "Winning" vs. College Listed... 1 Row Below	Implied Prob. of "Winning" vs. College Listed... 10 Rows Below	Rank Based on Matriculation (no Covariates)
1	Harvard University	9.13	0.59	0.93	1
2	Caltech	8.77	0.56	0.92	3
3	Yale University	8.52	0.59	0.92	2
4	MIT	8.16	0.51	0.89	5
5	Stanford University	8.11	0.52	0.90	4
6	Princeton University	8.02	0.73	0.90	6
7	Brown University	7.01	0.56	0.78	7
8	Columbia University	6.77	0.54	0.73	8
9	Amherst College	6.61	0.51	0.71	9
10	Dartmouth	6.57	0.52	0.72	10
11	Wellesley College	6.51	0.53	0.71	12
12	University of Pennsylvania	6.39	0.56	0.71	11

The effect of schools upon writing performance

1. A leading research team in motivational and educational psychology was interested in the effect that different teaching techniques at various colleges and universities have upon their students. They were particularly interested in the effect upon writing performance of attending a very selective university, a less selective university, or a two year junior college. A writing test was given to the entering students at three institutions in the Boston area. After one year, a similar writing test was given again. Although there was some attrition from each sample, the researchers report data only for those who finished one year. The pre and post test scores as well as the change scores were as shown below:

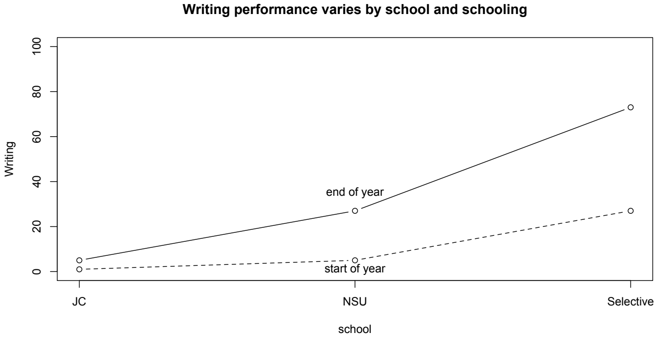
Writing Performance

Table: Writing performance by type of school

School	Pretest	Posttest	Change
Junior College	1	5	4
Non selective University	5	27	22
Selective University	27	73	45

1. From these data, the researchers concluded that the quality of teaching at the very selective university was much better and the student there learned a great deal more.
2. They proposed to study the techniques used there in order to apply them to other institutions
3. Do these results follow? What are alternative explanations?

Writing performance and teaching



The effect of school on math performance

1. Another research team in motivational and educational psychology was interested in the effect that different teaching techniques at various colleges and universities have upon their students. They were particularly interested in the effect upon math performance of attending a very selective university, a less selective university, or a two year junior college. A math test was given to the entering students at three institutions in the Boston area. After one year, a similar math test was given again. Although there was some attrition from each sample, the researchers report data only for those who finished one year. The pre and post test scores as well as the change scores were as shown below:

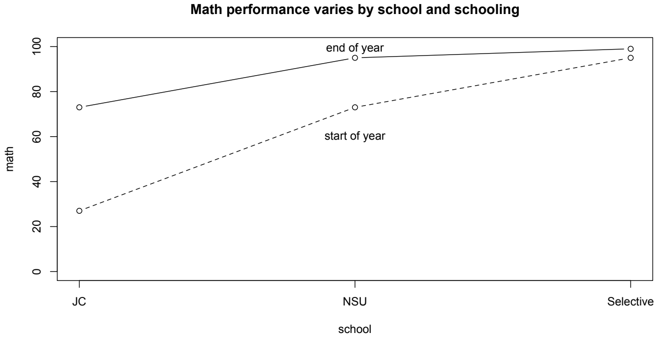
Math Performance

Table: Math performance by type of school

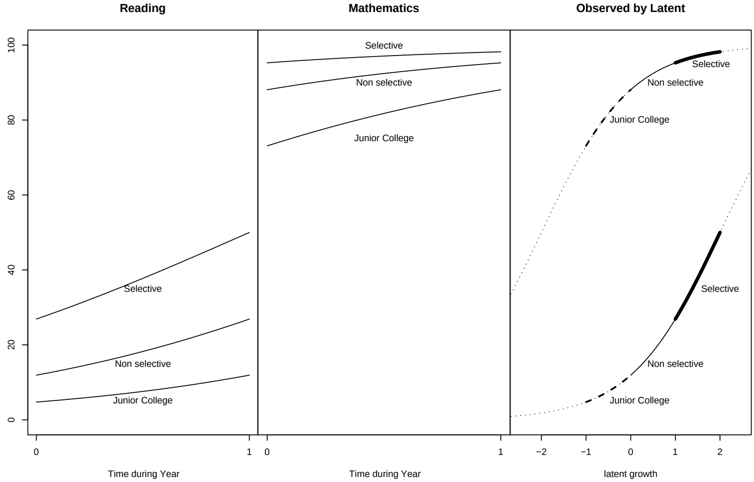
School	Pretest	Posttest	Change
Junior College	27	73	45
Non selective University	73	95	22
Selective University	95	99	4

1. From these data, the researchers concluded that the quality of teaching at the junior college was much better and the student there learned a great deal more.
2. They proposed to study the techniques used there in order to apply them to other institutions
3. Do these results follow? What are alternative explanations?

Math performance and teaching



The effect of scaling upon the latent variable - observed variable relationship



Observed Variables

X	Y
$\boxed{X_1}$	$\boxed{Y_1}$
$\boxed{X_2}$	$\boxed{Y_2}$
$\boxed{X_3}$	$\boxed{Y_3}$
$\boxed{X_4}$	$\boxed{Y_4}$
$\boxed{X_5}$	$\boxed{Y_5}$
$\boxed{X_6}$	$\boxed{Y_6}$

Latent Variables

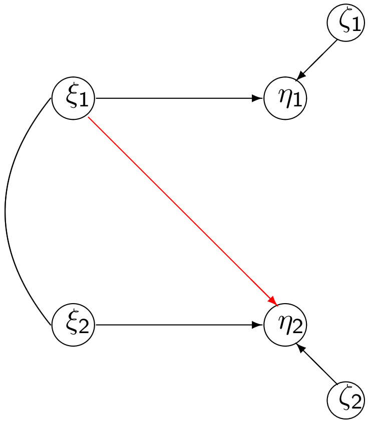
ξ η

ξ_1 η_1

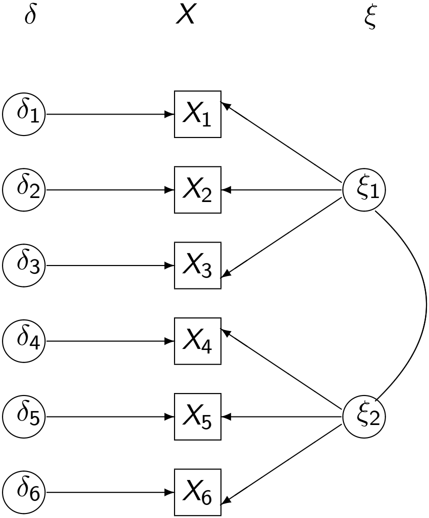
ξ_2 η_2

Theory: A regression model of latent variables

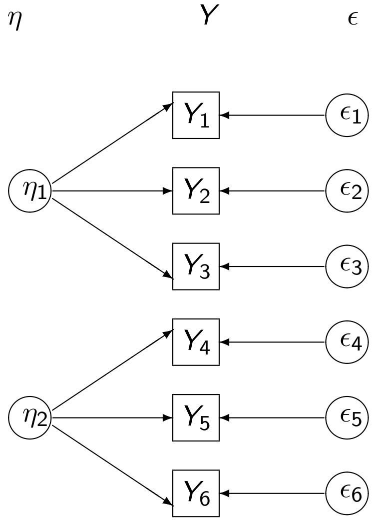
ξ η



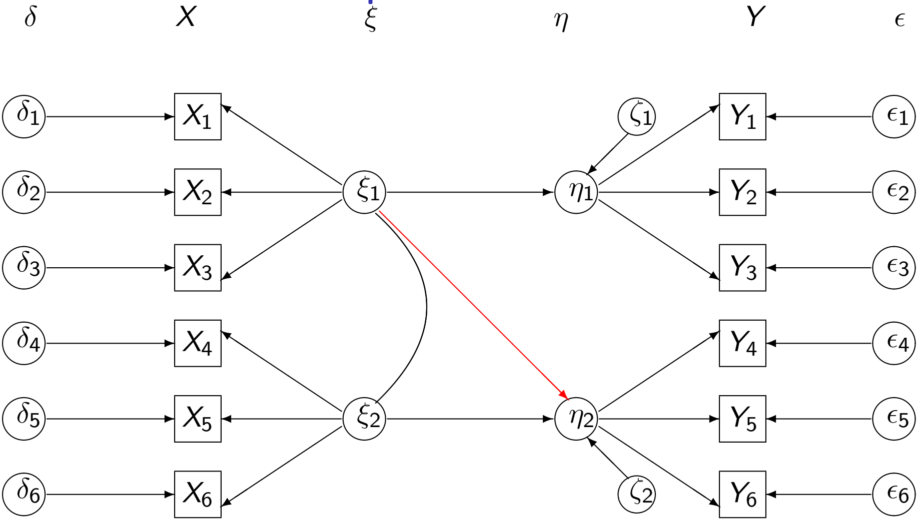
A measurement model for X – Correlated factors



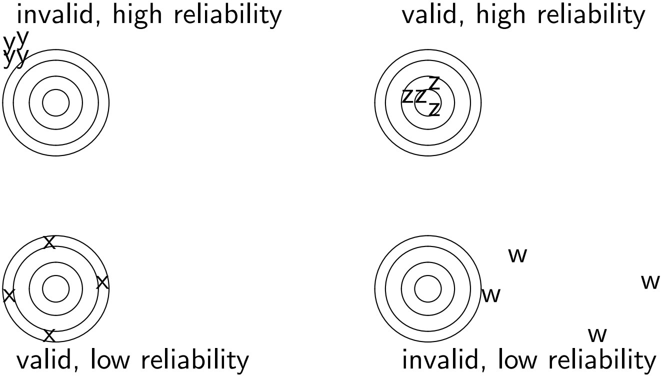
A measurement model for Y - uncorrelated factors



A complete structural model



Reliability and Validity



Reliability without validity

A personal example from when I was on an oceanographic expedition from San Diego to Bangkok while in high school.

1. To assess the oxygen content of a deep water sample, one does a chemical titration
   - The test is used to determine the concentration of dissolved oxygen in water samples.
   - This is important to understand the effects of climate on ocean circulation.
2. Add a reagent to the sea water sample until the solution turns clear.
3. After training, I was very reliable and spent the summer doing oxygen titrations.
4. But being color blind, my values were wrong!
5. I discovered my problem with doing titrations when I took a chemistry course in college.
6. I learned years later that they assumed the reagents were bad.

Two classes of validity

1. Internal Validity: Is systematic error (bias) minimized
   - Have we controlled for confounds?
   - This is the primary purpose of design.
2. External Validity: Does the study actually study what is reported?
   - Will the results generalize?
   - This is the purpose of understanding your sample and the reality of manipulations.

Internal Validity

1. Are the results of the experiment/study due to the variables considered
   - What are the constructs of interest?
   - Do the measured variables measures those constructs?
2. Are confounding variables controlled for?
   - What alternative explanations for the effect of the variables can you come up with?
   - How do you control for them?
3. What are other plausible explanations for your effect?

Major Threats to internal validity

1. Within subject experiments
   - Fatigue
   - Practice
   - Boredom
   - Order effect
2. Between subject experiments
   - Subject differences
   - Many ways subjects can differ.

Controlling for threats to internal validity within subjects

1. Every subject is their own control. Each subject is in all conditions.
2. What are the obvious sources of error, and how to control them?
3. Order effects may be controlled by counterbalancing
   - But some order effects need long delays between trials (e.g., drug studies)

External Validity: Does the study actually study what is reported?

1. Do the effects generalize across other subjects?
    - Are the effects true only for the type of subjects studied?
2. Do the effects generalize across other conditions
    - Are the effects true only for the specific situation studied?

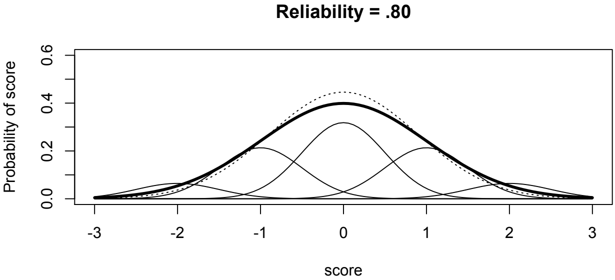
Reliability as an everyday problem

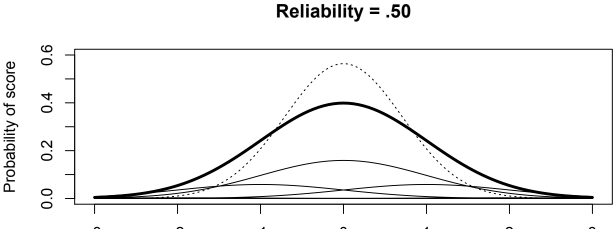
1. Baseball players and "MoneyBall" (Lewis, 2004)
   - Average batting average is .260 with a standard deviation of .027
   - Year to year correlation is .38
   - Someone who bats .343 one year is expected to bat .291 the next year!
2. Athletes and the "Sports Illustrated Curse"
   - Best performer of the year will not do as well next year
3. Pilot Trainers and the belief in punishment
   - Bad performance improves following punishment (or even without punishment)
   - Good performance decreases following reward (or even without reward)
4. School performance
5. Stock market investment advisors are rated by performance
   - But great performance does not persist

All data are befuddled with error

Now, suppose that we wish to ascertain the correspondence between a series of values, p, and another series, q. By practical observation we evidently do not obtain the true objective values, p and q, but only approximations which we will call p' and q'. Obviously, p' is less closely connected with q', than is p with q, for the first pair only correspond at all by the intermediation of the second pair; the real correspondence between p and q, shortly r_{pq} *has been "attenuated" into* $r_{p'q'}$ *(Spearman, 1904, p 90).*

All data are befuddled by error: Observed Score = True score + Error score





Regression effects due to unreliability of measurement

Consider the case of air force instructors evaluating the effects of reward and punishment upon subsequent pilot performance. Instructors observe 100 pilot candidates for their flying skill. At the end of the day they reward the best 50 pilots and punish the worst 50 pilots.

- Day 1
  - Mean of best 50 pilots 1 is 75
  - Mean of worst 50 pilots is 25
- Day 2
  - Mean of best 50 has gone down to 65 (a loss of 10 points)
  - Mean of worst 50 has gone up to 35 (a gain of 10 points)
- It seems as if reward hurts performance and punishment helps performance.
- If there is no effect of reward and punishment, what is the expected correlation from day 1 to day 2?

(Kahneman & Tversky, 1973; Kahneman, 2011)

Classical True score theory

Let each individual score, x, reflect a true value, t, and an error value, e, and the expected score over multiple observations of x is t, and the expected score of e for any value of p is 0. Then, because the expected error score is the same for all true scores, the covariance of true score with error score (σ_{te}) is zero, and the variance of x, σ_x^2, is just

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 + 2\sigma_{te} = \sigma_t^2 + \sigma_e^2.$$

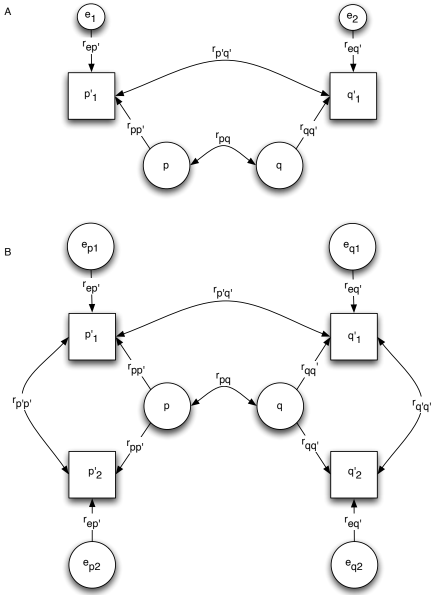
Similarly, the covariance of observed score with true score is just the variance of true score

$$\sigma_{xt} = \sigma_t^2 + \sigma_{te} = \sigma_t^2$$

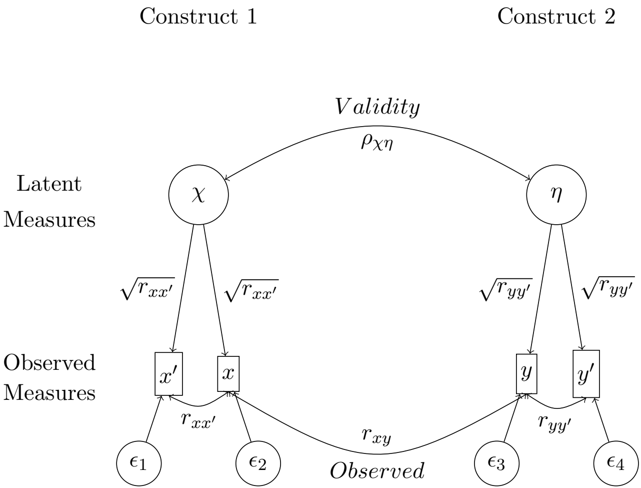
and the correlation of observed score with true score is

$$\rho_{xt} = \frac{\sigma_{xt}}{\sqrt{(\sigma_t^2 + \sigma_e^2)(\sigma_t^2)}} = \frac{\sigma_t^2}{\sqrt{\sigma_x^2 \sigma_t^2}} = \frac{\sigma_t}{\sigma_x}. \quad (1)$$

Spearman's parallell test theory



Reliability and Validity



(Revelle & Condon, 2019)

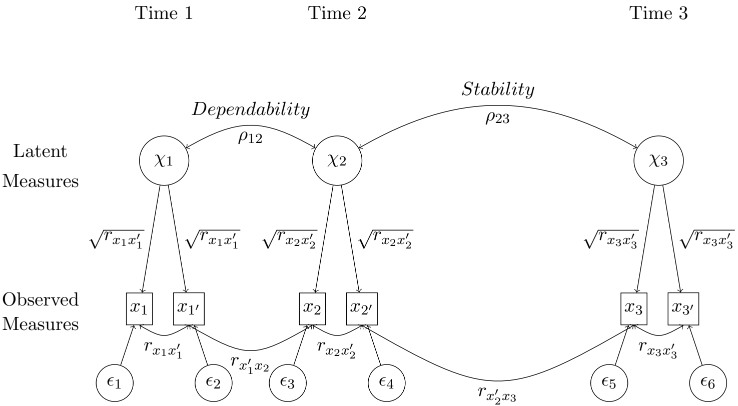
Correcting for attenuation

> *To ascertain the amount of this attenuation, and thereby discover the true correlation, it appears necessary to make two or more independent series of observations of both p and q. (Spearman, 1904, p 90)*

Spearman's solution to the problem of estimating the true relationship between two variables, p and q, given observed scores p' and q' was to introduce two or more additional variables that came to be called *parallel tests*. These were tests that had the same true score for each individual and also had equal error variances. To Spearman (1904) this required finding "the average correlation between one and another of these independently obtained series of values" to estimate the reliability of each set of measures ($r_{p'p'}, r_{q'q'}$), and then to find

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} r_{q'q'}}}. \tag{2}$$

Reliability, Dependability and stability



(Revelle & Condon, 2019)

Types of reliability

Reliability coefficient

- Internal consistency
  - α
  - $\omega_{hierarchical}$
  - ω_{total}
  - β
- Intraclass
- Agreement
- Test-retest, alternate form
- Generalizability

Reliability measurement

- Internal consistency
  - `alpha`, `score.items`
  - `omega`
  - `iclust`
- `icc`
- `wkappa`, `cohen.kappa`
- `cor`
- `aov`

Coefficient α

Find the correlation of a test with a test just like it based upon the internal structure of the first test. Basically, we are just estimating the error variance of the individual items. Known as α (Cronbach, 1951) or λ_4 (Guttman, 1945) this is just

$$\alpha = r_{xx} = \frac{\sigma_t^2}{\sigma_x^2} = \frac{k^2 \frac{\sigma_x^2 - \sum \sigma_i^2}{k(k-1)}}{\sigma_x^2} = \frac{k}{k-1} \frac{\sigma_x^2 - \sum \sigma_i^2}{\sigma_x^2} \tag{3}$$

That is, as the number of items increases, the reliability goes up. And as the items correlate more highly, the reliability goes up. This is the principle of most tests, we give items that are not necessarily very good, but by giving enough of them, we have a good test.
This can be thought of as the "Rapunzel effect". Many strands of weak hair make for a strong rope.

Signal to Noise Ratio

The ratio of reliable variance to unreliable variance is known as the Signal/Noise ratio and is just

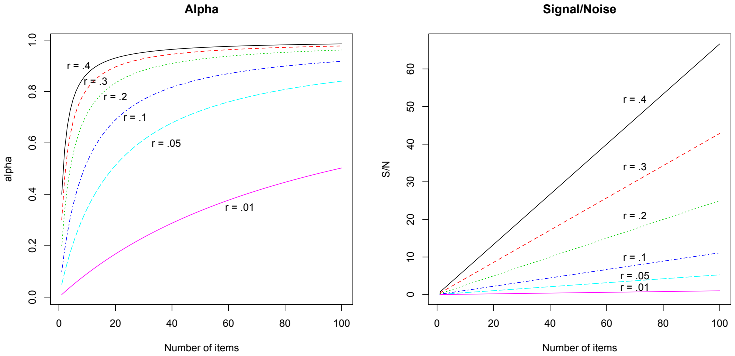
$$\frac{S}{N} = \frac{\rho^2}{1-\rho^2},$$

which for the same assumptions as for α, will be

$$\frac{S}{N} = \frac{n\bar{r}}{1-\bar{r}}. \tag{4}$$

That is, the S/N ratio increases linearly with the number of items as well as with the average intercorrelation.

Alpha vs signal/noise: and r and n



Does noise hurt learning?

1. For their class project, Alice and Bob were interested in the effect of ambient noise on learning. They selected 40 students to participate in one of two conditions: Quiet (50 dba) or Noisy (75dba). They each ran 20 subjects. Alice ran the quiet condition in the third floor of the library while Bob ran the noisy condition at a table in Norris. Students were given a text passage to study for 10 minutes and then given a 20 item multiple choice test. The scores were
2. Quiet mean = 15, sd = 3, N = 20
3. Noisy mean = 10, sd = 2, N = 20
4. These means differed significantly (t=6.2, p << .001). Alice and Bob concluded that noise hindered study efficiency when compared to quiet.
5. There at least two major problems with this design. What are they?

Does noise hurt learning, part II: random assignment

1. For their class project, Alice and Bob were interested in the effect of ambient noise on learning. They randomly assigned 40 students to participate in one of two conditions: Quiet (50 dba) or Noisy (75dba). Alice ran the quiet condition in the third floor of the library while Bob ran the noisy condition at a table in Norris. Students were given a text passage to study for 10 minutes and then given a 20 item multiple choice test. The scores were
2. Quiet mean = 15, sd = 3, N = 20
3. Noisy mean = 10, sd = 2, N = 20
4. These means differed significantly ($t=6.2$, $p << .001$). Alice and Bob concluded that noise hindered study efficiency when compared to quiet. There is a major problem with this design. What is it?

N = 1 design

1. Cynthia, an auto mechanic, wants to know which of two different brands of motor oil will make a car easier to start in a cold winter morning. She designs an experiment to find out, in which the number of seconds until the engine starts is the dependent measure. Cynthia has 1 (one) car. On each of 10 different mornings, Cynthia fills her car with brand A motor oil. Then she tests to see how long it takes to start. After waiting for the engine to cool completely, she empties out the brand A motor oil and fills her car with brand B motor oil, then tests it again. (This question does not require any knowledge of cars – this is a question about design).
2. What is wrong with this procedure?
3. If she has only 1 car, is it possible for her to determine which is the better motor oil? How could this be done?

Design problems in developmental psychology

1. Three developmental psychologists believed that happiness increases with age among married couples (Levenson, Carstensen & Gottman, 1993). They collected data from two randomly selected sets of married couples in the San Francisco area: couples who were 40-50 years old and had been married for at least 15 years and couples who were 50-65 years old and had been married for at least 25 years. All couples has been married only once.
2. They found that the older couples reported more positive affect and less negative affect than did the younger couples. They concluded from this that age does indeed lead to happiness.
3. There is a serious artifact in this study that makes the conclusions questionable. What is it?
4. Can you think of a way to get around this problem?

Avery, C. N., Glickman, M. E., Hoxby, C. M., & Metrick, A. (2013). A revealed preference ranking of U.S. colleges and universities. *The Quarterly Journal of Economics*, *128*(1), 425–467.

Burchard, U. (2004). The sclerometer and the determination of the hardness of minerals. *Mineralogical Record*, *35*, 109–120.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237–251.

Levenson, R. W., Carstensen, L. L., & Gottman, J. M. (1993). Long-term marriage: Age, gender, and satisfaction. *Psychology and Aging*, *8*(2), 301–313.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.

Revelle, W. & Condon, D. M. (2019). Reliability: from alpha to omega. *Psychological Assessment*, *31*(12), 1395–1411.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Wainer, H. (1990). Graphical visions from william playfair to john tukey. *Statistical Science*, *5*(3), 340–346.

Wainer, H. (1999). The most dangerous profession: A note on nonsampling error. *Psychological Methods*, *4*(3), 250–256.

Wald, A. (Ed.). (1980). *A Method of Estimating Plane Vulnerability Based on Damage of Survivors*, number ADA091073, ALEXANDRIA VA OPERATIONS EVALUATION GROUP. CENTER FOR NAVAL ANALYSES.