Nunes K. L., & Choy, A. (2018, October). *Understanding and evaluating research studies: The basics*. Pre-conference seminar given at the 37th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers (ATSA), Vancouver, British Columbia, Canada.

Understanding and Evaluating Research Studies: The Basics

Kevin L. Nunes — Carleton University — kevin.nunes@carleton.ca

Alberto Choy — Alberta Hospital Edmonton — Alberto.Choy@albertahealthservices.ca

Agenda

- Research Design
- Validity
- Statistics

2

Exercises

1. Complete each exercise on your own
2. Compare and discuss your answers in small groups
3. Review answers all together

3

Validity

Introduction

- Research is the foundation of effective assessment, intervention, and policy aimed at reducing sexual offending
- However, studies vary in how informative and conclusive they are, and there are differences of opinion about standards for interpreting evidence

5

More Rigorous Designs Permit Stronger Conclusions

- Description of a case or sample, but no comparisons or associations examined
- Cross-sectional/retrospective non-experimental
- Single wave longitudinal non-experimental
- Multi-wave (the factor was assessed at two or more time points) longitudinal non-experimental
- Randomized experiment

Associated

Predictor

Cause

6

More Rigorous Designs Permit Stronger Conclusions

Weakest inferences

- Description of a case or sample, but no comparisons or associations examined
- Cross-sectional/retrospective non-experimental
- Single wave longitudinal non-experimental
- Multi-wave (the factor was assessed at two or more time points) longitudinal non-experimental
- Randomized experiment

Strongest inferences

7

Threats to Validity

- Methodologically rigorous studies minimize the number and plausibility of threats to validity, thereby permitting stronger inferences

8

- "Threats to validity are reasons why we can be partly or completely wrong when we make an inference about covariance [statistical conclusion validity], about causation [internal validity], about constructs [construct validity], or about whether the causal relationship holds over variations in persons, settings, treatments, and outcomes [external validity]" (Shadish et al., 2002, p. 39)

9

4 Main Types of Validity (Shadish et al., 2002)

- Internal validity
  - To what extent does the evidence reflect a causal relationship between the variables as measured or manipulated?
- Construct validity
  - To what extent do the people, places, and variables as measured or manipulated represent the presumed constructs?
- External validity
  - To what extent does the causal relationship generalize to other people, places, and variables?
- Statistical conclusion validity
  - To what extent do the results accurately reflect the statistical significance or strength of association between variables?

10

Some Ways to Minimize the Plausibility of Alternative Interpretations

- More rigorous research designs
- Better measures; i.e., scores have demonstrated acceptable levels of reliability and construct validity
- Larger samples
- More representative sampling of population
- Multiple comparison groups
- Measurement and statistical control of plausible alternative causal variables

11

One Study Cannot Rule Out Every Threat

- Typically a trade-off between different types of validity in any one study

12

Complementary Evidence

- Use the complementary strengths and weaknesses of different studies to gather and produce more informative body of evidence. For example,
  - Randomized experiments (usually with non-forensic/correctional samples and analogous variables)
  - Studies of forensic/correctional samples (usually observational cross-sectional or, less often, longitudinal)

13

Steps in Testing Causal Hypotheses

- Definition
  - Clear and precise conceptualization of the construct
- Measurement
  - Accurate measurement of the construct
- Causal role
  - Test of the extent to which the construct plays a causal role

14

Speculation

- When evidence is limited, speculation is the best we can do
- Past research and theoretical models are important, valuable, and useful

15

Strength of Inferences Should Match Strength of Evidence

- But, sensitivity to the limits methodology places on inferences is important for the sake of accuracy and integrity, and to stimulate more informative research (e.g., Harris & Rice, 2015; Nunes et al., 2017)
- Stronger research can provide more conclusive evidence to guide practice and policy

16

Statistics

OUTLINE: BASIC STATISTICS IN RESEACH

- How statistics are used in research
- Various statistics common in sexual offender research
  - Descriptive statistics
  - Significance
  - Effect sizes
  - Confidence intervals
  - Meta-analyses

18

How Statistics Are Used

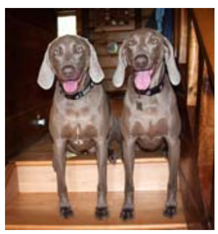
Who are we looking at?

- Who is this population?
- How old, big, small, young, tall, green, wide, smiley, silly, hungry…

20

What question are they trying to answer?

- Is there a difference between these?
- How different are these?



What question are they trying to answer?

- Are these (statistically) the same?
- What features are the same?
- How similar are these?

What question are they trying to answer?

- Do these guys have some relationship to each other?

The 10,000 foot summary

- Statistics are used in the research in our field to find relationships:
  - Describe who we are studying
  - Differences and similarities in more direct comparisons
  - Also, when considering multiple things or multiple comparisons:
    - Is there a relationship here amongst things?
    - What is the nature of that relationship?

Nunes K. L., & Choy, A. (2018, October). *Understanding and evaluating research studies: The basics*. Pre-conference seminar given at the 37th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers (ATSA), Vancouver, British Columbia, Canada.

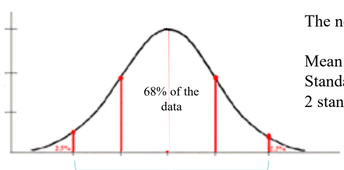
Assumptions

Descriptive statistics

- Tell us basic information about a population
  - Who is this population?
  - Mean, mode, median, standard deviation
  - Also descriptions of our population (like recidivism rate)

- May be useful to compare to the normal curve

The normal curve

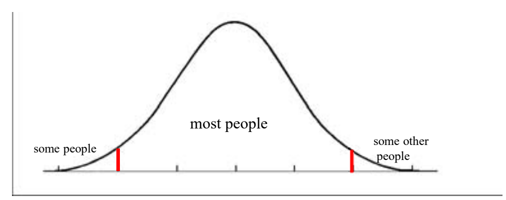


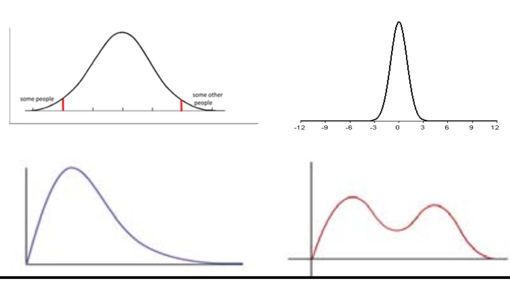
The normal curve

Mean = average
Standard deviation
2 standard deviations

- 95 % of the data

What you really really need to know





Layperson's guide to descriptive statistics

- Descriptive statistics are mostly for:
  a) understanding who the group of study is, and
  b) setting the record for generalizability and future studies to compare populations

Looking for significance

p

- Probability that there would be no difference or no association if the analysis were repeated with other samples of the same population (how likely is it that this is just a fluke?)
- If p is less than a pre-determined cutoff (usually $p < .05$) then the difference or association is statistically significant
- e.g., $r = .47$, $p = .03$
  - if the same population were sampled 100 times, the correlation should be zero or negative in only 3 of those samples
- e.g., $r = -.47$, $p = .03$
  - if the same population were sampled 100 times, the correlation should be zero or positive in only 3 of those samples

32

Cheater's guide to results

- Step 1: Look for the *, or **, or *** (or the "ns")
- Step 2: Look for the associated *p*-value
- Step 3: re-read the results to try and understand

Effect Sizes vs. Significance Tests

- Magnitude of association or difference
  - How strong is the relationship between variables?
  - How large is the difference between groups?
- Different than significance testing
- Some examples of effect sizes:
  - Correlation, Cohen's *d*, area under the curve (AUC), odds ratio, hazard ratio

34

Some Relevant and Common Analyses/Statistics

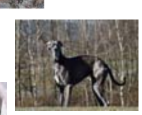
- Correlation
- Cohen's *d*
- Area under the curve (AUC) of the Receiver Operating Characteristic (ROC)
- Logistic regression
- Cox regression

35

Measuring the relationship between things: correlations

- When these things appear, how often do these other things appear?

- When I see a lot of fluffy dogs, I think I see a lot of smaller dogs, is there a correlation between fluffiness and size of dog?

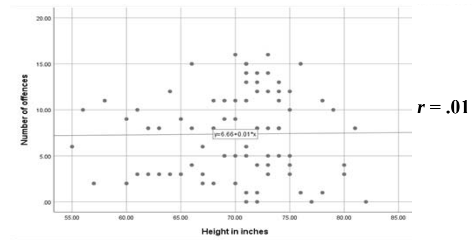


Correlation

- Pearson product-moment correlation
  - 2 continuous variables
    - e.g., what is the relationship between IQ and empathy?
  - Can range from -1.00 to +1.00
  - .00 indicates no association
  - Convention: r = .10 is small, .30 is medium, and .50 is large (Cohen, 1988, 1992)

37

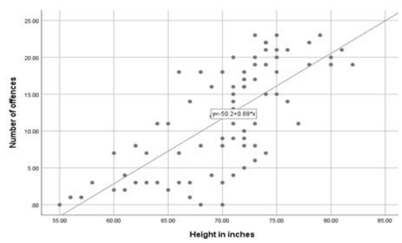
No Correlation Between Height and Crime (fake data)



r = .01

38

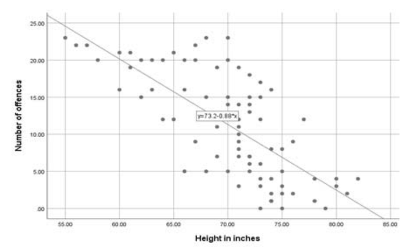
Positive Correlation Between Height and Crime (fake data)



r = .72

39

Negative Correlation Between Height and Crime (fake data)



r = -.72

40

Correlation coefficients in medicine

Pearson *r*

+/- .00 to .30 = negligible correlation

+/- .30 to .50 = low or weak correlation

+/- .50 to .70 = moderate correlation

+/- .70 to .90 = high or strong correlation

+/- .90 to .99 = very high correlation

- Point-biserial correlation
  - Correlation for 1 dichotomous variable and 1 continuous variable
    - e.g., to what extent do men and women differ in empathy?
  - Can range from -1.00 to +1.00
  - .00 indicates no association
  - Convention: r_{pb} = .10 is small, .24 is medium, and .37 is large, assuming 50% split on dichotomous variable (Cohen, 1988)

42

Nunes K. L., & Choy, A. (2018, October). *Understanding and evaluating research studies: The basics*. Pre-conference seminar given at the 37th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers (ATSA), Vancouver, British Columbia, Canada.

- Phi (ϕ)
  - Correlation for 2 dichotomous variables
    - e.g., to what extent is gender associated with recidivism (vs. no recidivism)?
  - Can range from -1.00 to +1.00
  - .00 indicates no association
  - Convention: ϕ = .10 is small, .24 is medium, and .37 is large, assuming 50% split on dichotomous variable (Cohen, 1988)

43

Cheater's guide to psychology correlations

		Small association	Medium association	Large association
"Pearson"	r	.10	.30	.50
"point-biserial correlation"	r_{pb}	.10	.24	.37
"Phi"	ϕ	.10	.24	.37
Medicine (approximately)	r	*.30*	*.50*	*.70*

Finding an r of even .20 is important in complex, multifactorial phenomena
The average r for social psychology is .21

44

Measurement between groups

- Correlations can be considered measurements of association
- What about differences between groups?

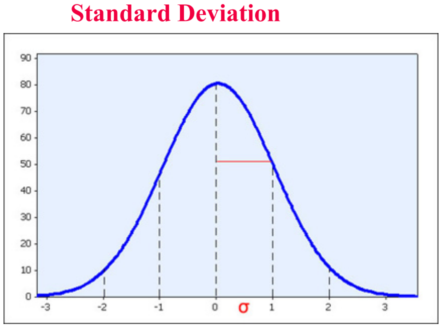

45

Cohen's *d*

- Number of standard deviations by which two groups differ
  - $(M_1 - M_2)$ / Pooled *SD*
- Can be positive or negative, with no limits
- 0.00 indicates no difference
- Convention: d = 0.20 is small, 0.50 is medium, and 0.80 is large (Cohen, 1988, 1992)

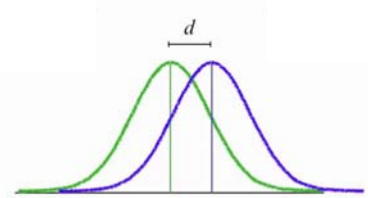
46

Standard Deviation



47

Cohen's *d*



48

Cheater's guide to Cohen's d

		Small	Moderate	Large
Cohen's *d*	*d*	0.20	0.50	0.80
Correlation coefficient	*r*	0.10	0.30	0.50

49

Prediction

50

Correct Predictions

- True positive (TP)
  - The assessor predicts an offender will **reoffend** and he does **reoffend**
- True negative (TN)
  - The assessor predicts an offender will **not reoffend** and he does **not reoffend**

51

Incorrect Predictions

- False positive (FP)
  - The assessor predicts an offender will **reoffend** but he does **not reoffend**
- False negative (FN)
  - The assessor predicts an offender will **not reoffend** but he does **reoffend**

52

Predictive Accuracy

PREDICTION	OUTCOME: Reoffence	OUTCOME: No reoffence
Reoffence	TP ✔	FP ✖
No reoffence	FN ✖	TN ✔

Sensitivity: Proportion of recidivists correctly predicted to recidivate

Specificity: Proportion of non-recidivists correctly predicted to not recidivate

53

Fake Assessment Instrument by Kevin and Bert (FAIKB) Fake data

FAIKB score	Did not recidivate	Recidivated
0	6 (86%)	1 (14%)
1	14 (78%)	4 (22%)
2	18 (64%)	10 (36%)
3	7 (44%)	9 (56%)
4	2 (22%)	7 (78%)
TOTAL	47	31

54

With a FAIKB cutoff score of 2
Predict recidivism for everyone with a FAIKB score of 2 or more

FAIKB score	Did not recidivate	Recidivated
0	6 (86%)	1 (14%)
1	14 (78%)	4 (22%)
2	18 (64%)	10 (36%)
3	7 (44%)	9 (56%)
4	2 (22%)	7 (78%)
TOTAL	47	31

5 false negatives

26 true positives

Sensitivity: 26/31 = .84

55

With a FAIKB cutoff score of 2
Predict recidivism for everyone with a FAIKB score of 2 or more

FAIKB score	Did not recidivate	Recidivated
0	6 (86%)	1 (14%)
1	14 (78%)	4 (22%)
2	18 (64%)	10 (36%)
3	7 (44%)	9 (56%)
4	2 (22%)	7 (78%)
TOTAL	47	31

20 true negatives

27 false positives

Specificity: 20/47 = .43

56

Area under the curve (AUC) of the receiver operating characteristic (ROC)

- Indicates overall accuracy considering both sensitivity and specificity
- Plots sensitivity vs. specificity (1 – specificity) for each cutoff score

Carleton University
Canada's Capital University

57

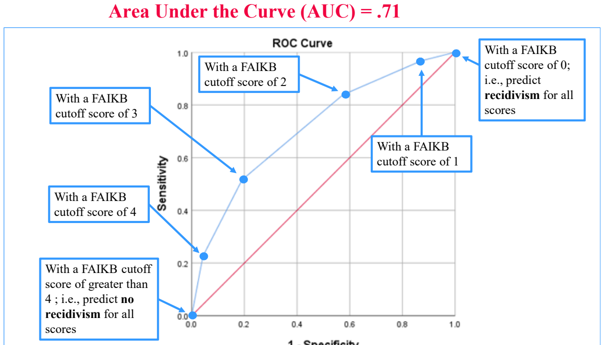
AUC / ROC

- Probability that a randomly selected person from one group has a higher score than a randomly selected person from the other group
- Can range from .00 to +1.00
- .50 indicates variable distinguishes groups at chance level of accuracy
- Convention: AUC = .56 is small, .64 is medium, and .71 is large (Rice & Harris, 2005)

Carleton University
Canada's Capital University

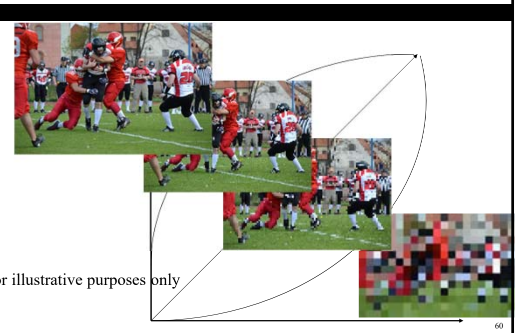
58

Area Under the Curve (AUC) = .71



59

Carleton University
Canada's Capital University



AUC for illustrative purposes only

60

For illustrative purposes only

Logistic Regression

- Test association between one or more predictor variable with one categorical outcome variable (e.g., recidivism vs. no recidivism)
- Odds ratio ($\exp^B$)
  - Indicates how much higher or lower the odds of the outcome are given a one-point higher score on the predictor
  - 1.00 = no association between the predictor and the odds of the outcome

62

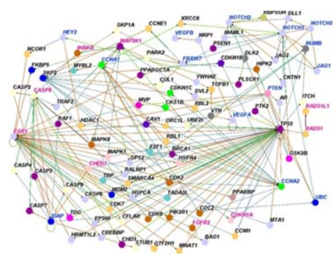
Cox Regression

- Test association between one or more predictor variable with the rate over time of one categorical outcome variable (e.g., recidivism vs. no recidivism)
- Hazard ratio ($\exp^B$)
  - Indicates how much higher or lower the rate of the outcome is given a one-point higher score on the predictor
  - 1.00 = no association between the predictor and the odds of the outcome

63

Layman's understanding of regression

- Statistically isolating one factor from all of the others to examine the importance of that item



- Credit to: Hicks C, Pannuti A, Miele L - Cancer Inform (2011)

Some tips on understanding common statistics in risk studies

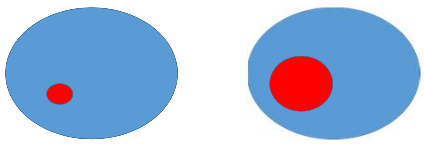
- Hanson meta-analysis for the Static 99
  - r = .10 for variables studied really meant it was not useful for prediction
- In complex phenomena, r=.20 may be important
- r^2 can be described as "the VRAG score accounted for r^2 percent of the variance"
  - VRAG total score was correlated to recidivism at r=.20
    - So a higher score was correlated to greater recidivism rates
  - r^2 can be described as "the VRAG score accounted for r^2 percent of the variance"
  - "The VRAG score accounts for 4% of the variance of recidivism"

Layperson's language and risk

- AND: this r is sensitive to base rate.
- r will be lower with lower base rates (higher false positive)
- So, it is difficult to make "apples to apples" comparison between instruments

Explaining AUC

- Probability that a randomly selected person from the offending group has a higher score than a randomly selected person from the non-offending group



Risk tool A   Risk tool B

67

Confidence Intervals

- Estimate of the range of values within which the statistic (e.g., effect size) would be expected to fall some proportion of the time (e.g., 95% for 95% CI) for the population being sampled
- e.g., r = .47, 95% CI [.05, .75]
  - if the same population were sampled 100 times, the r should fall between .05 and .75 in 95 of those samples

68

- If the confidence interval does not include the value that indicates no relationship/difference (e.g., r = .00, d = 0.00, AUC = .50, odds/hazard ratio = 1.00), then the effect size is statistically significant ($p < .05$)
- Confidence intervals become narrower with larger samples, and narrower is better
- e.g., r = .47, 95% CI [.40, .54]
  - … the r should fall between .40 and .54 in 95 of 100 samples

69

Relative vs. Incremental Predictive Validity

- Relative
  - Which variable most strongly predicts offending (e.g., recidivism)
  - Which variable is the best predictor?
- Incremental
  - Which variables independently predict offending (e.g., recidivism)
  - Which variables are non-redundant predictors?

70

Practical Examples

71

Nunes et al. (2007)

Table 1 Comparison of sexual offenders who admitted their index offenses to those who denied their index offenses

Variable	Admitter		Denier			95% CI	
	n	*M* (*SD*) or %	*n*	*M* (*SD*) or %	*d*	Lower	Upper
PCL-R	250	17.99 (7.70)	78	19.55 (8.06)	0.20	−0.06	0.45
RRASOR-M	350	1.11 (1.22)	135	1.34 (1.17)	0.19	−0.01	0.39
Recidivism							
Sexual	352	14.8%	137	15.3%	0.02	−0.18	0.21
Violent	352	23.6%	137	27.0%	0.08	−0.12	0.28

72

Nunes et al. (2007)

Table 2 Sequential logistic regression predicting sexual recidivism

Predictor	*B*	SE *B*	Wald	Odds ratio	95% CI
Block 1					
PCL-R	0.07	0.02	12.31*	1.07	1.03–1.11
RRASOR-M	0.43	0.12	13.40*	1.54	1.22–1.95
Block 2					
PCL-R	0.07	0.02	12.41*	1.07	1.03–1.11
RRASOR-M	0.44	0.12	13.51*	1.55	1.23–1.96
Denial	−0.12	0.35	0.12	0.89	0.45–1.75
Block 3					
PCL-R	0.07	0.02	10.91*	1.08	1.03–1.12
RRASOR-M	0.44	0.12	13.51*	1.55	1.23–1.96
Denial	0.38	0.99	0.15	1.46	0.21–10.04
PCL-R by denial	−0.02	0.04	0.28	0.98	0.90–1.06
Block 4					
PCL-R	0.08	0.02	10.78*	1.08	1.03–1.13
RRASOR-M	0.64	0.15	18.99*	1.90	1.42–2.53
Denial	1.44	1.06	1.83	4.20	0.53–33.61
PCL-R by denial	−0.02	0.04	0.21	0.98	0.90–1.07
RRASOR-M by denial	−0.70	0.28	6.40*	0.50	0.29–0.85

χ^2(2, *N*=326)=28.25 for Block 1 (*p*<.05). χ^2(1, *N*=326)=0.12 for Block 2 (*p*>.05). χ^2(1, *N*=326)=0.28 for Block 3 (*p*>.05). χ^2(1, *N*=326)=6.75 for Block 4 (*p*<.05). *SE* standard error. *CI* confidence interval.
* *p*<.05.

73

Nunes et al. (2007)

Table 2 Sequential logistic regression predicting sexual recidivism

Predictor	*B*	SE *B*	Wald	Odds ratio	95% CI
Block 1					
PCL-R	0.07	0.02	12.31*	1.07	1.03–1.11
RRASOR-M	0.43	0.12	13.40*		1.22–1.95
Block 2					
PCL-R	0.07	0.02	12.41*		1.03–1.11
RRASOR-M	0.44	0.12	13.51*		1.23–1.96
Denial	−0.12	0.35	0.1		0.45–1.75
Block 3					

Each additional point on the PCL-R was associated with 7% greater odds of sexual recidivism.

PCL-R	0.08	0.02	10.78*	1.08	1.03–1.13
RRASOR-M	0.64	0.15	18.99*	1.90	1.42–2.53
Denial	1.44	1.06	1.83	4.20	0.53–33.61
PCL-R by denial	−0.02	0.04	0.21	0.98	0.90–1.07
RRASOR-M by denial	−0.70	0.28	6.40*	0.50	0.29–0.85

χ^2(2, *N*=326)=28.25 for Block 1 (*p*<.05). χ^2(1, *N*=326)=0.12 for Block 2 (*p*>.05). χ^2(1, *N*=326)=0.28 for Block 3 (*p*>.05). χ^2(1, *N*=326)=6.75 for Block 4 (*p*<.05). *SE* standard error. *CI* confidence interval.
* *p*<.05.

74

Olver et al. (2014)

Table 1
Predictive Accuracy of Sexual Offender Risk Measures

Risk measure	Sexual recidivism *N*	*r*	*AUC*	95% CI
All pretreatment cases				
Static-99R	673	.17	.71	[.65, .78]
VRS-SO dynamic total	673	.15	.67	[.59, .76]
Sexual deviance	668	.06[ns]	.57[ns]	[.49, .65]
Criminality	671	.15	.68	[.60, .76]
Treatment responsivity	671	.08[b]	.59[ns]	[.50, .68]
Pre- and posttreatment cases				
VRS-SO dynamic total (pre)	572	.13[a]	.66[a]	[.56, .75]
VRS-SO dynamic total (post)		.16	.67	[.58, .77]
Sexual deviance (pre)	572	.05[ns]	.56[ns]	[.47, .65]
Sexual deviance (post)		.06[ns]	.57[ns]	[.48, .66]
Criminality (pre)	576	.13	.66	[.57, .75]
Criminality (post)		.16	.68	[.59, .77]
Treatment responsivity (pre)	574	.08[ns]	.59[ns]	[.49, .69]
Treatment responsivity (post)		.12[a]	.62[b]	[.51, .72]

Note. Unmarked point biserial correlations and AUC values are significant at *p* < .001. ns = not significant. AUC = area under the curve; CI = confidence interval; VRS-SO = Violence Risk Scale–Sexual Offender version.
[a] *p* < .01. [b] *p* < .05.

75

Olver et al. (2014)

Table 2
Cox Regression Survival Analysis: Incremental Validity of the Static-99R and VRS-SO Dynamic Items in the Prediction of Recidivism Criteria

Regression model	Sexual recidivism *B*	*SE*	Wald	*p*	e^B	95% CI *LL*	*UL*
Pretreatment							
Static-99R	.18	.07	7.49	**.006**	1.20	1.054	1.371
Dynamic total	.04	.02	3.76	.052	1.05	1.000	1.093
Posttreatment							
Static-99R	.16	.07	5.00	**.025**	1.18	1.020	1.355
Dynamic total	.05	.02	4.52	**.034**	1.05	1.004	1.100

Note. Pretreatment *N* = 670; posttreatment *N* = 572. Significant *p* values are in bold font. VRS-SO = Violence Risk Scale–Sexual Offender version; *SE* = standard error; CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

76

Olver et al. (2014)

Table 2
Cox Regression Survival Analysis: Incremental Validity of the Static-99R and VRS-SO Dynamic Items in the Prediction of Recidivism Criteria

Regression model	Sexual recidivism *B*	*SE*	Wald	*p*	e^B	95% CI *LL*	*UL*
Pretreatment							
Static-99R	.18	.07	7.49	**.006**	1.20	1.054	1.371
Dynamic total	.04	.02	3.76	.052		1.000	1.093

Each additional point on the Static-99R was associated with a 20% greater/faster rate of sexual recidivism.

Note. Pretreatment *N* = 670; posttreatment *N* = 572. Significant *p* values are in bold font. VRS-SO = Violence Risk Scale–Sexual Offender version; *SE* = standard error; CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

77

Meta-Analysis

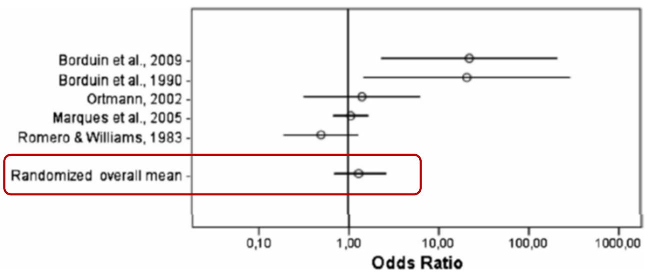
- Synthesize results from different studies by computing average effect size
- In our area, the average effect size reported is usually correlation, Cohen's *d*, or odds ratio

Carleton UNIVERSITY
Canada's Capital University

78

Nunes K. L., & Choy, A. (2018, October). *Understanding and evaluating research studies: The basics*. Pre-conference seminar given at the 37th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers (ATSA), Vancouver, British Columbia, Canada.

Schmucker & Lösel (2015) Meta-Analysis



79

Meta-Analysis Statistics

- Q
  - Significance test of heterogeneity of effect sizes
  - If $p < .05$, then statistically significant amount of variability
- I^2
  - Estimate of extent of heterogeneity of effect sizes
  - The larger the value, the greater the variability
- $Q_{between}$
  - Significance test of a potential moderator variable
  - If $p < .05$, then statistically significant difference in average effect size by the moderator variable

80

McPhail et al. (2013) Meta-Analysis

Table 5
Meta-Analyses of the Relationship Between Emotional Congruence With Children and Sexual Recidivism

Variable	Mean d	95% CI	Z	Q (df)	$Q_{between}$ (df)	I^2	k (N)
Overall	0.39	[0.21, 0.57]	4.29***	13.72 (7)		48.97	8 (5,217)
Measurement method					1.51 (2)		
Risk assessment	0.45	[0.14, 0.76]	2.85**	0.28 (1)		0.00	2 (397)
Self-report	0.25	[−0.06, 0.56]	1.59	6.12* (2)		67.32	3 (3,924)
Typology	0.55	[0.15, 0.96]	2.67**	4.53 (2)		55.89	3 (375)
SOC subgroup					8.10**(1)		
SOC-E[a]	0.58	[0.31, 0.85]	4.16***	5.86 (4)		31.78	5 (639)
SOC-I	−0.15	[−0.58, 0.27]	−0.71	0.14 (2)		0.00	3 (893)

Note. CI = confidence interval; SOC = sexual offenders against children; SOC-E = extrafamilial sexual offenders against children; SOC-I = intrafamilial sexual offenders against children.
[a] With outlier: $d = 0.47$, 95% CI [0.15, 0.79], $Z = 2.85$, $p = .004$, $Q(5) = 16.54$, $p = .005$, $I^2 = 69.77$, $Q_{between}(1) = 5.21$, $p = .023$, $k = 6$, $N = 1,615$.
* $p < .05$. ** $p < .01$. *** $p < .001$.

81

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.*

Cohen, J. (1992). A power primer. *Psychological Bulletin, 122, 155–159.*

Harris, G. T., & Rice, M. E. (2015). Progress in violence risk assessment and communication: Hypothesis versus evidence. *Behavioral Sciences and the Law, 33,* 128-145.

McPhail, I. V., Hermann, C. A., & Nunes, K. L. (2013). Emotional congruence with children and sexual offending against children: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 81*, 737-749.

82

Nunes, K. L., Hanson, R. K., Firestone, P., Moulden, H. M., Greenberg, D. M., & Bradford, J. M. (2007). Denial predicts recidivism for some sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 19,* 91-105.

Nunes, K. L., Pedneault, C., Filleter, W. E., Maimone, S., Blank, C., & Atlas, M. (2017). “I know correlation doesn’t prove causation, but…”: Are we jumping to unfounded conclusions about the causes of sexual offending? *Sexual Abuse: A Journal of Research and Treatment.* Advance online publication.

Olver, M. E., Nicholaichuk, T. P., Kingston, D. A., & Wong, S. C. P. (2014). A multisite examination of sexual violence risk and therapeutic change. *Journal of Consulting and Clinical Psychology, 82,* 312-324.

83

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen’s *d*, and *r*. *Law and Human Behavior, 29,* 615-620.

Schmucker, M., & Lösel, F. (2015). The effects of sexual offender treatment on recidivism: An international meta-analysis of sound quality evaluations. *Journal of Experimental Criminology, 11,* 597-630.

Shadish, W. R., Cook. T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Belmont, CA: Wadsworth.

84