NORTHERN ARIZONA UNIVERSITY

THE STATISTICAL ANALYSIS AND RESULTS SECTIONS OF A MANUSCRIPT

The Essentials for Reviewers

Monica Lininger PhD, ATC, LAT

OUTLINE FOR WORKSHOP

1. **STATISTICAL ANALYSIS**
   - CHECKLIST
   - VARIABLES
   - ASSUMPTION TESTING
   - MISSING DATA
2. **RESULTS**
   - CHECKLIST
   - EFFECT SIZES
   - CONFIDENCE INTERVALS

ESSENTIALS OF THE STATISTICAL ANALYSIS SECTION

NORTHERN ARIZONA UNIVERSITY

CHECKLIST [1]

STATISTICAL ANALYSIS

- ❑ Description of the independent and dependent variable(s) including covariates
- ❑ Assumptions underlying the statistical tests being used
- ❑ Statistical power and sample size estimation is reported
  - May be seen early in Participants section
- ❑ Methods of handling missing data are discussed
- ❑ Descriptive statistics being utilized to summarize data
- ❑ Analytical techniques to assess differences, relationships, associations, prediction, etc.
- ❑ Post-hoc analyses, when appropriate
- ❑ Criterion to assess statistical significance
- ❑ Name and version of software package

VARIABLES
INDEPENDENT AND DEPENDENT

- **Explicitly stated including:**
  - **Levels of independent variable(s)**
  - **Fixed and random effects**
    - Fixed – generalizations about specific levels
    - Random – generalizations back to an entire population
  - **Scales of measurement for each variable**
    - Nominal
    - Ordinal
    - Interval
    - Ratio
  - **Within or between subject factors**

ASSUMPTION TESTING

- Violations of assumptions can influence Type I and Type II errors

- “The applied researcher who routinely adopts a traditional procedure without giving thought to its associated assumptions may unwittingly be filling the literature with nonreplicable results.”[2]

ASSUMPTION TESTING

HOEKSTRA ET AL.[3]

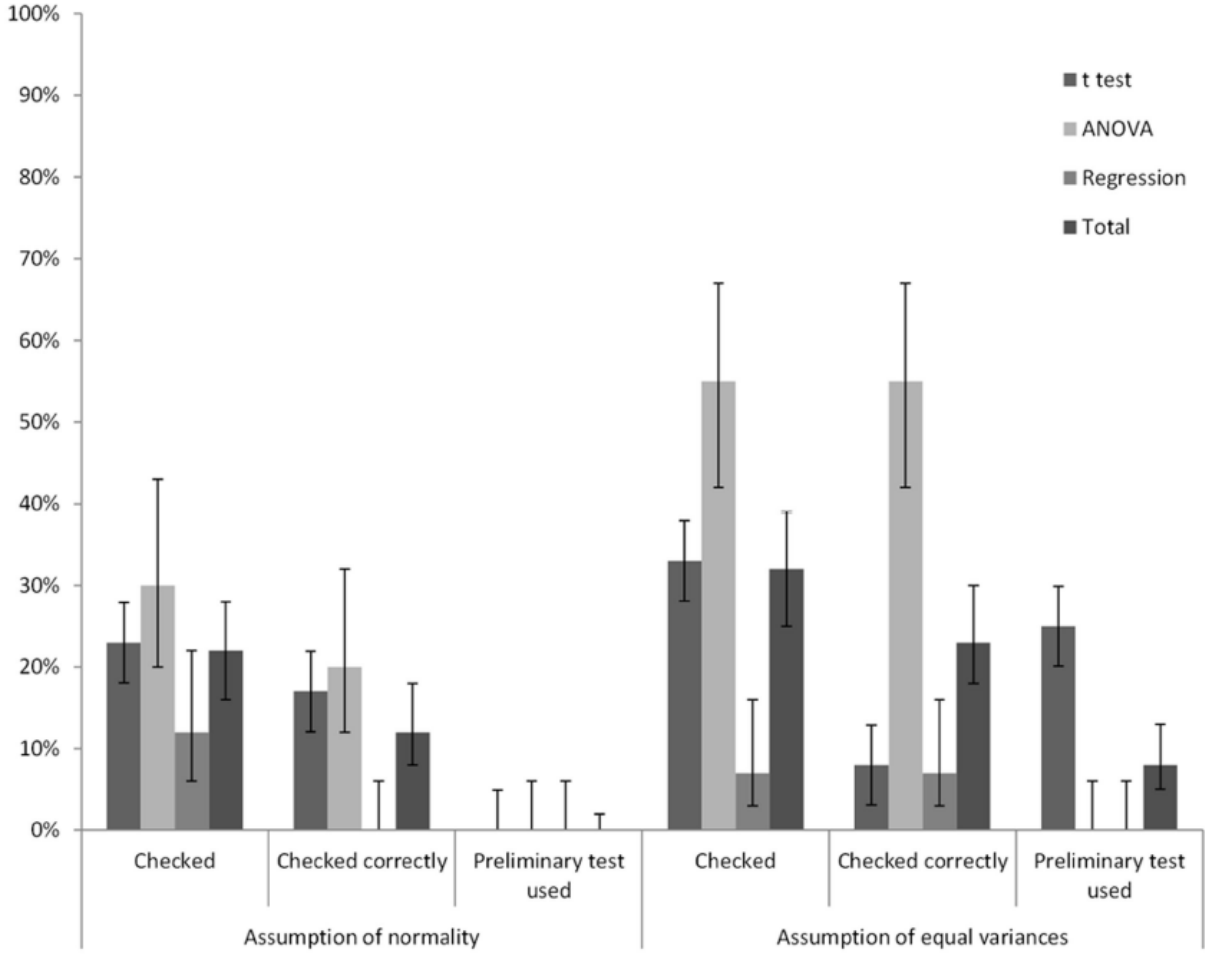
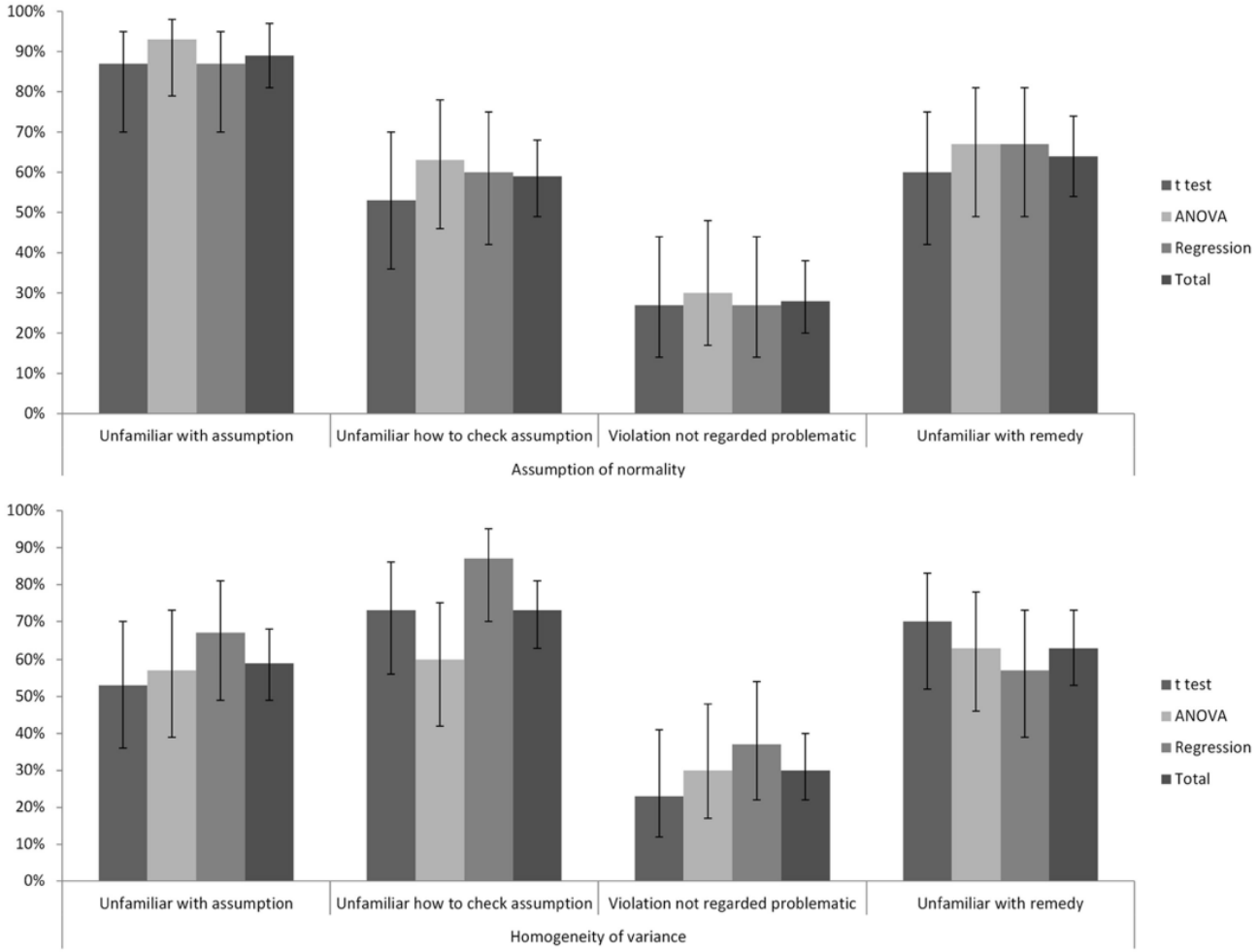


FIGURE 2 | The frequency of whether two assumptions were checked at all, whether they were checked correctly, and whether a preliminary test was used for three often used techniques in percentages of the total number of cases. Between brackets are 95% CIs for the percentages.

ASSUMPTION TESTING

HOEKSTRA ET AL.[3] CONT.



GURE 3 | Percentages of participants giving each of the explanations for not checking assumptions as a function of assumption and technique.
ror bars indicate 95% CIs.

ASSUMPTION TESTING

OVERVIEW OF JOURNALS FOR 2016

	Research Articles	Quantitative Articles	Mentioned Assumption Testing
ATEJ	20	12	1
JAT	98	88	23

- ***ATEJ* – 8%**
- ***JAT* – 26%**
- **Most commonly tested**
  - Normality and homogeneity of variance

ASSUMPTION TESTING

INDEPENDENCE

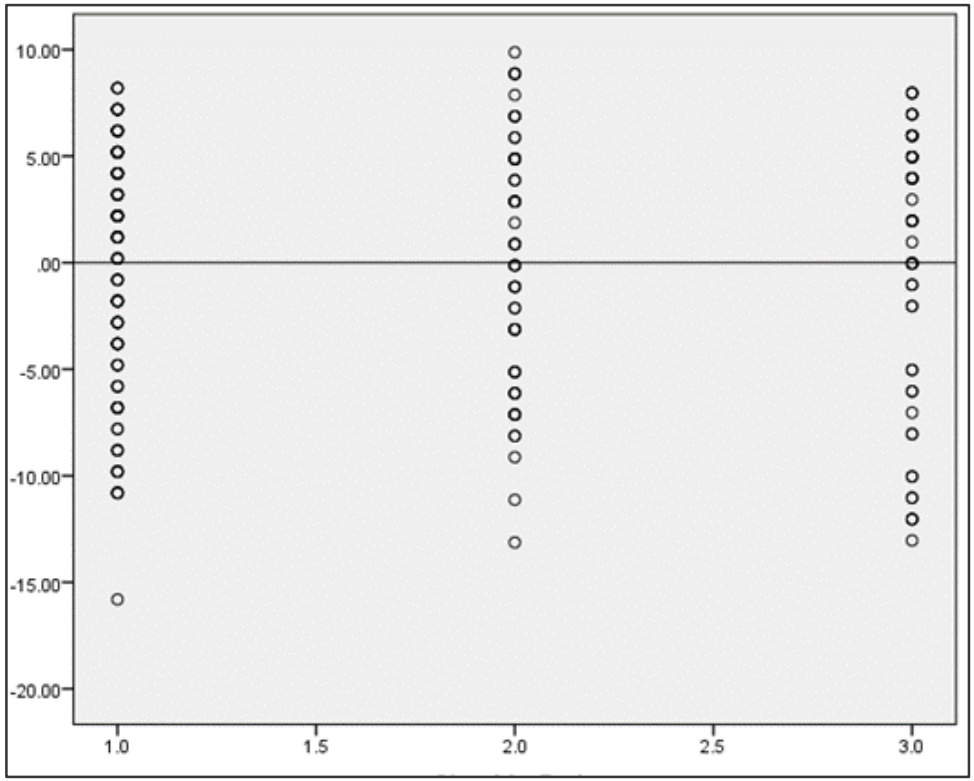
- Each sample is randomly selected from a population
- Methods
  - Very challenging to assess through statistics
  - Examine residuals by group
    - Should maintain a ‘random display’[4]
    - Durbin-Watson statistic assesses autocorrelation

- Violations
  - Serious implications especially to the F ratio[4]
  - Impacts standard errors of the sample means
- Options
  - Not many since violation truly takes place in the design phase
  - Randomize whenever possible

ASSUMPTION TESTING

INDEPENDENCE EXAMPLE

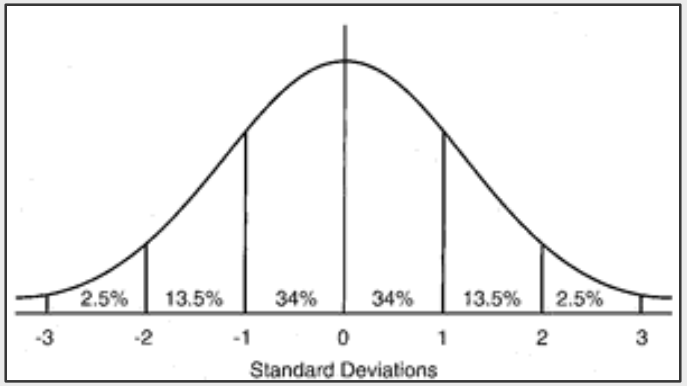
- Example of ‘random display’[4]



ASSUMPTION TESTING

NORMALITY

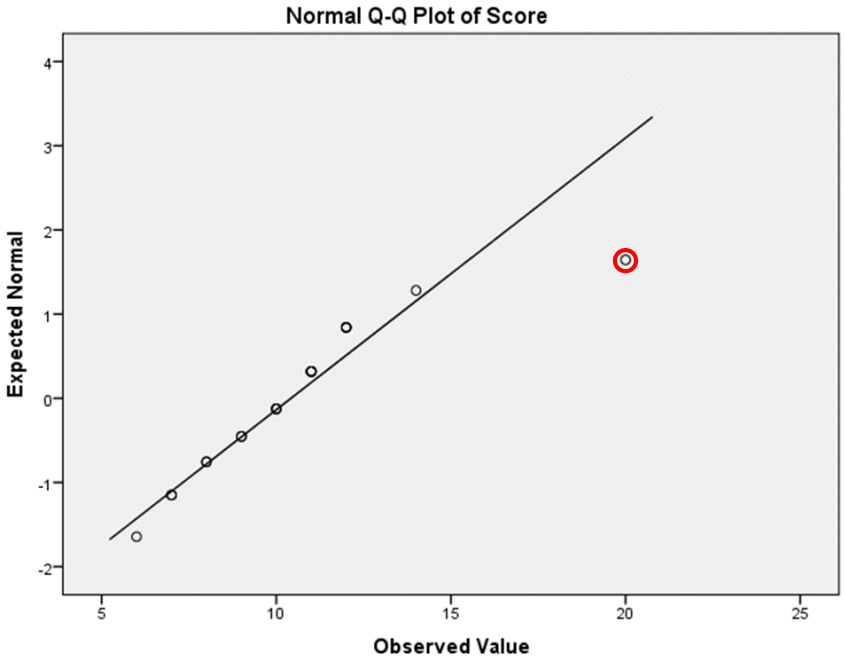
- Normal distribution with a mean of zero and a standard deviation of one
- Methods
  - Skewness and kurtosis
  - Q-Q plot
  - Shapiro-Wilk's W test
  - Kolmogorov-Smirnov test



- Violations
  - Most *F*-tests are robust to violations
- Options
  - Investigate outliers
  - Nonparametric analyses
  - Transformations
    - Log
    - Square root
      - Counts that follow a Poisson distribution
    - Angular
      - Proportions or percentages that follow a binomial distribution

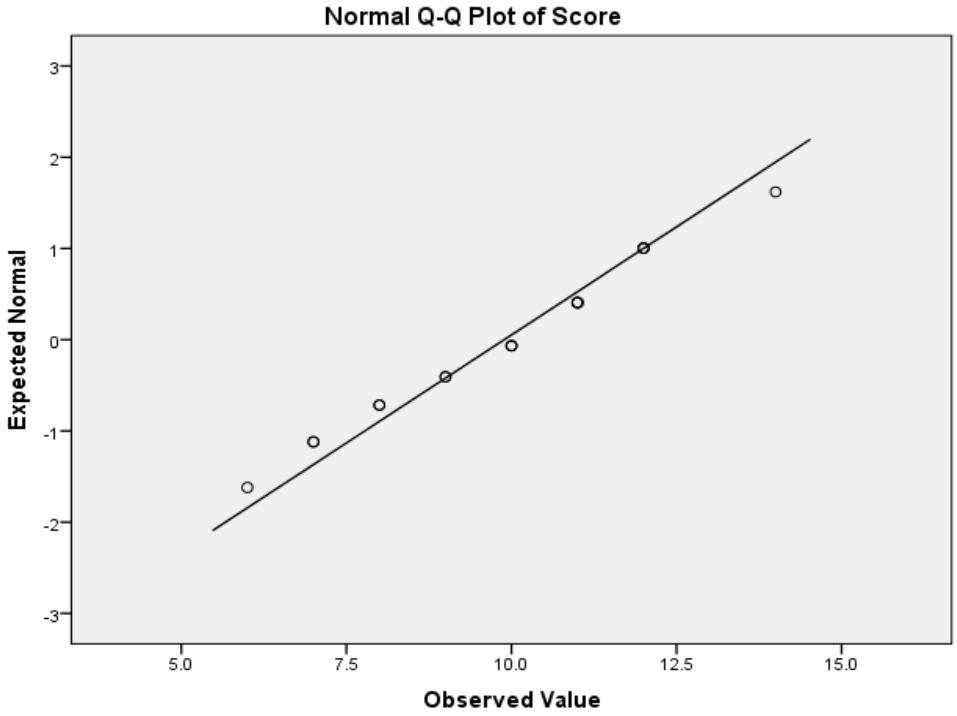
ASSUMPTION TESTING

NORMALITY EXAMPLE



ASSUMPTION TESTING

NORMALITY EXAMPLE



ASSUMPTION TESTING

NORMALITY EXAMPLE

Descriptives

			Statistic	Std. Error
Score	Mean		10.42	.710
	95% Confidence Interval for Mean	Lower Bound	8.93	
		Upper Bound	11.91	
	5% Trimmed Mean		10.13	
	Median		10.00	
	Variance		9.591	
	Std. Deviation		3.097	
	Minimum		6	
	Maximum		20	
	Range		14	
	Interquartile Range		4	
	Skewness		1.540	
	Kurtosis		4.304	

Descriptives

			Statistic	Std. Error
Score	Mean		9.89	.498
	95% Confidence Interval for Mean	Lower Bound	8.84	
		Upper Bound	10.94	
	5% Trimmed Mean		9.88	
	Median		10.00	
	Variance		4.458	
	Std. Deviation		2.111	
	Minimum		6	
	Maximum		14	
	Range		8	
	Interquartile Range		3	
	Skewness		-.132	.536
	Kurtosis		-.465	1.038

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Score	.200	19	.044	.874	19	.017

a. Lilliefors Significance Correction

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Score	.145	18	.200*	.966	18	.711

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

ASSUMPTION TESTING

HOMOGENEITY OF VARIANCE

- Equal variances across samples
- Methods
  - Levene's test
  - Bartlett's test
    - Uses chi-square statistic and based on meeting assumption of normality
  - Box's M test
    - Multivariate homogeneity

Test of Homogeneity of Variances

Pain

Levene Statistic	df1	df2	Sig.
.510	2	27	.606

- Violations
  - Bias error term
    - Small sample sizes and violation leads to increase in Type I error (incorrectly rejecting the null hypothesis)
- Options
  - Brown-Forsythe procedure
  - Welch procedure
  - Decrease alpha

ASSUMPTION TESTING

LINEARITY

- The relationship between X and Y is linear
  - Mainly for ANCOVA and regression models
- Methods
  - Plot of Y versus X

- Violations
  - General linear model
  - Under-estimate the true relationship
- Options
  - Transformations
  - Polynomial regression

ASSUMPTION TESTING

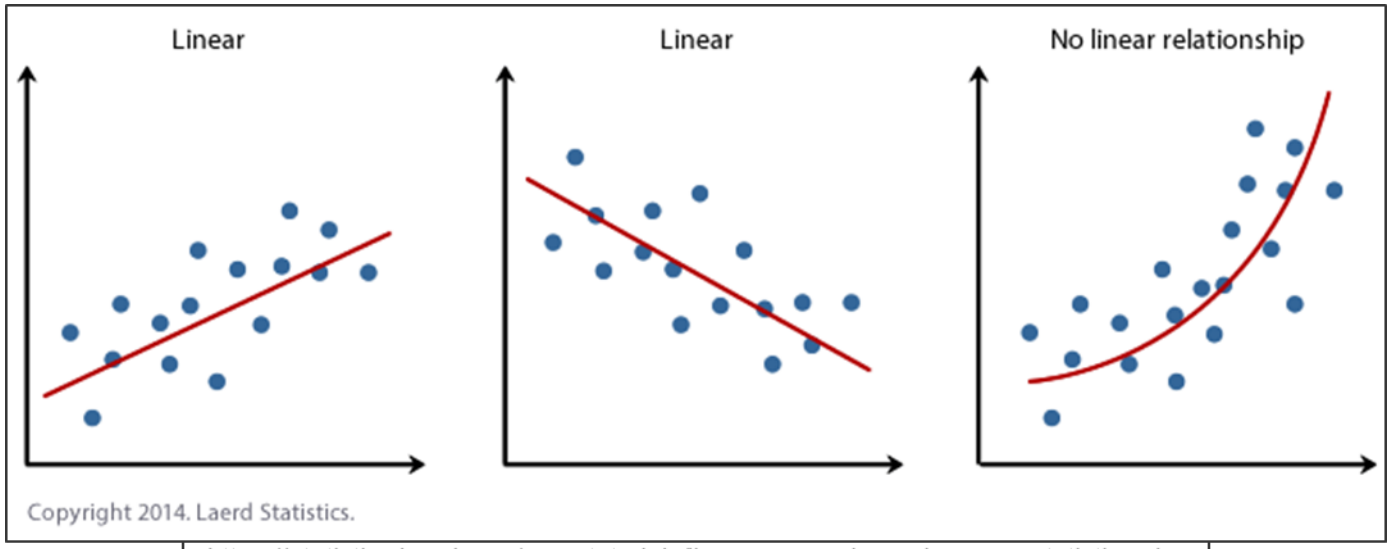
LINEARITY

- The relationship between X and Y is linear
  - Mainly for ANCOVA and regression models
- Methods
  - Plot of Y versus X

- Violations
  - General linear model
  - Under-estimate the true relationship
- Options
  - Transformations
  - Polynomial regression

ASSUMPTION TESTING

LINEARITY EXAMPLE



https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php

Journal of Athletic Training 2016;51(11):936–945
doi: 10.4085/1062-6050-51.10.13

www.natajournals.org

Evaluation of 2 Heat-Mitig
Trainees

**JoEllen M. Sefton, PhD, ATC*; J. S. M
Lohse, PhD*; Robert L. Banda, ME
Andrew R. Cherrington, MEd, ATC*;**

*Warrior Research Center, School of Kinesiology, Aub

Context: Heat injury is a significant threat to milit
trainees. Different methods of heat mitigation are in use acr
military units. Mist fans are 1 of several methods used in the
and humid climate of Fort Benning, Georgia.

Objectives: To determine if (1) the mist fan or the cool
towel effectively lowered participant core temperature in
humid environment found at Fort Benning and (2) the mist fan
the cooling towel presented additional physiologic or saf
benefits or detriments when used in this environment.

Design: Randomized controlled clinical trial.

Setting: Laboratory environmental chamber.

Patients or Other Participants: Thirty-five physically act
men aged 19 to 35 years.

Intervention(s): (1) Mist fan, (2) commercial cooling tow
(3) passive-cooling (no intervention) control. All treatme
lasted 20 minutes. Participants ran on a treadmill at 6
$\dot{V}O_2$max.

Main Outcome Measure(s): Rectal core temperature, he
rate, thermal comfort, perceived temperature, perceived w
ness, and blood pressure.

Tests of Statistical Assumptions. We created Q-Q normal plots for each group at each time point and judged all the distributions to be approximately normal. A Shapiro-Wilk test confirmed no significant deviations from normality, except for core temperature in the cooling-towel group at time 0 of the control condition ($W = 0.81$, $P < .01$). Given that analysis of variance (ANOVA) is robust to violations of the normality assumption, we chose not to transform all cases of the dependent variable to adjust for this relatively minor violation of normality.

To test the effects of the different treatment conditions on core temperature and heart rate, we first conducted a group (mist fan versus cooling towel) × condition (passive-cooling control versus active-cooling experimental) × time (0 versus 20 minutes) mixed-factorial ANOVA with repeated measures on condition and time. Due to a significant 3-way interaction involving both core temperature and heart rate, we conducted follow-up condition × time ANOVAs separately for each group. For the analysis of blood pressure, we used a similar group × condition × time mixed-factorial ANOVA but with the additional repeated measure of cycle (diastolic versus systolic pressure). For analysis of the survey measures (thermal comfort, perceived temperature, and perceived wetness), a group × condition × time mixed-factorial ANOVA was conducted separately for each outcome.

MISSING DATA [5]

What is a missing value?

- **Missing completely at random (MCAR)**
  - Missing value doesn't depend on other variables
- **Missing at random (MAR)**
  - Missing value does not depend on variable of interest, after accounting for observed data
- **Missing not at random (MNAR)**
  - Probability of a missing value depends on the variable that is missing

What should I do as a reviewer?

Journal of Athletic Training 2016;51(7):540–549
doi: 10.4085/1062-6050-51.9.01

www.natajournals.org

Use of Cold-Water Immer
Damage and Delayed-Ons
Preserve Muscle Power i

Líllian Beatriz Fonseca, MS*; Ciro J. PhD*; Marzo Edir Silva-Grigoletto, PhD*; Emerson Franchini, PhD‡

*Postgraduation Program of Physical Education, Fe
Education, Federal University of Juiz de Fora, Minas G
School of Physical Education and Sport, University of

Context: Cold-water immersion (CWI) has been appl widely as a recovery method, but little evidence is available support its effectiveness.

Objective: To investigate the effects of CWI on mus damage, perceived muscle soreness, and muscle pov recovery of the upper and lower limbs after jiu-jitsu training.

Design: Crossover study.

Setting: Laboratory and field.

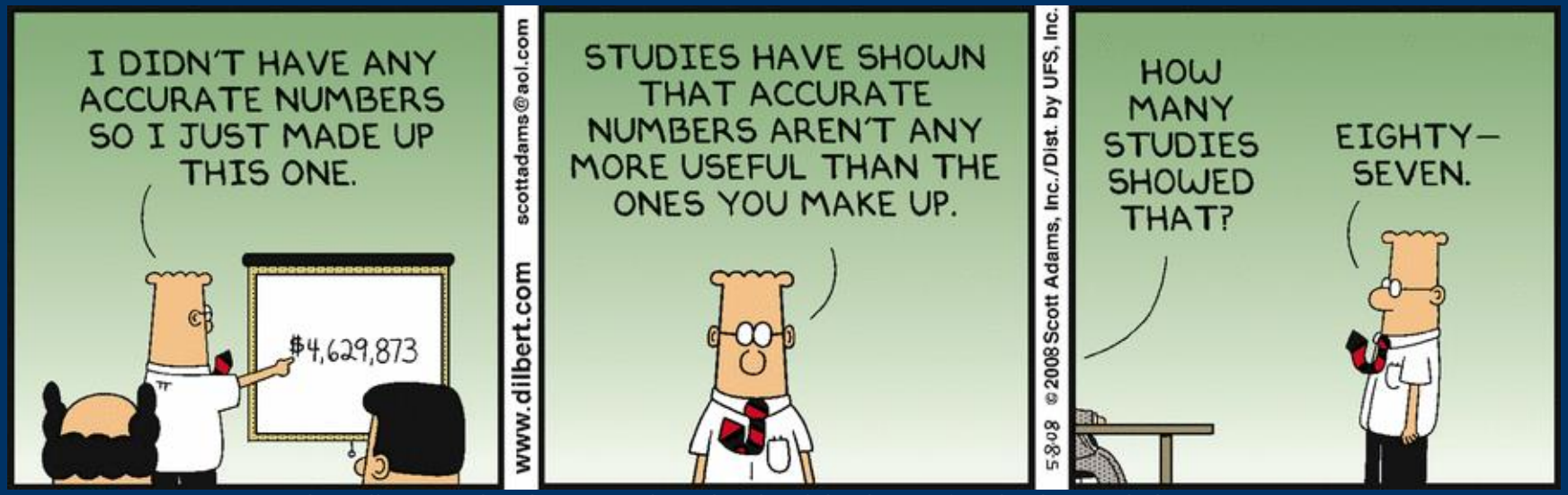
Patients or Other Participants: A total of 8 highly trair male athletes (age = 24.0 ± 3.6 years, mass = 78.4 ± 2.4 percentage of body fat = 13.1% ± 3.6%) completed all stu phases.

Intervention(s): We randomly selected half of the sam for recovery using CWI (6.0°C ± 0.5°C) for 19 minutes; the ot participants were allocated to the control condition (pass recovery). Treatments were reversed in the second sess (after 1 week).

Main Outcome Measure(s): We measured serum levels creatine phosphokinase, lactate dehydrogenase (LDH), asp tate aminotransferase, and alanine aminotransferase enzym perceived muscle soreness; and recovery through vis

Statistical Analysis

Exploratory data analysis was performed for identification and correction of extreme values, which was necessary only for CK. Normality and homoscedasticity were tested using the Kolmogorov-Smirnov test and the Bartlett criterion, respectively. We used analysis of variance with 2 factors (recovery × measurement time) to establish mean differences. For validation of repeated measurements, we used the Mauchly sphericity test and, when necessary, applied the Greenhouse-Geisser correction. If we observed a difference in the analysis of variance, we used a post hoc Bonferroni test. When a main effect and interaction were found, only the interaction effect was reported. The magnitude of treatment effects was calculated using the η^2 effect size. The upper and lower 95% confidence intervals (CIs) were calculated for corresponding mean variations. The standardized effect size (Cohen d)[28] analysis was used to interpret the magnitude of differences among measurements. To examine the strength of association among variables, we used the Pearson product moment correlation. The α level was set at .05 for all analyses. We used SPSS (version 15.0; SPSS Inc, Chicago, IL) to analyze the statistics.



https://stats.stackexchange.com/questions/423/what-is-your-favorite-data-analysis-cartoon

ESSENTIALS OF THE RESULTS SECTION

RESULTS SECTION CHECKLIST [1]

- **Sufficient information about the results of the test of significance including test statistics and degrees of freedom.**

- **Need to move past only reporting *P*-value as well as < 0.05**
  - There are problems with reporting only the *P*-value of a hypothesis test[6,7]
  - "We teach it because it's what we do; we do it because it's what we teach." [8]
  - Helpful Links for Authors of the *JAT*

- **Adequate statistical information to facilitate interpretation of results.**
  - Means with standard deviations
  - Effect sizes
  - Confidence intervals

RESULTS SECTION CHECKLIST [1] CONT.

- **Put into normal language and support with statistical evidence.**
  - There was a statistical difference between the treatment and the control group ($t_{23} = 5.321$, $P = 0.025$).
  - Student-athletes had higher tests scores (45.6 ± 2.32) with the new method compared to the student-athletes in the control group (42.2 ± 2.20) ($t_{23} = 5.321$, $P = 0.025$, 95%CI: 2.85, 3.95, Cohen's $d = 1.50$).

EFFECT SIZES

NORTHERN ARIZONA UNIVERSITY

EFFECT SIZES

- **Indicator of the practical importance of the research results.**
  - Magnitude of the observed effect or relationship

- **No direct relationship between a *P*-value and the magnitude of the effect.[10]**
  - **Williams (2003) compared the percent of time that faculty members spent teaching with the percent of time they would prefer to spend teaching.**
    - $t_{154} = 2.20$, $P = 0.03$, Cohen's $d = 0.09$

- **Nearly 70 different effect size indexes.[11]**
  - Goodman-Kruskal's lambda

TYPES OF EFFECT SIZES

- **Unstandardized**
  - **Means of variables with meaningful units that can be directly interpreted**
    - Treatment increase of 6°
    - Control increase of 2°

- **Standardized**
  - **Results expressed on a unitless scale**
  - ***d* family**
    - Differences between groups
  - ***r* family**
    - Measure of association or relationship

NORTHERN ARIZONA UNIVERSITY

d FAMILY

TWO INDEPENDENT SAMPLES

- **Cohen's *d* [12]**
  - Similar standard deviations

$$d = \frac{\bar{x}_1 - \bar{x}_2}{S_{pooled}}$$

$$S_{pooled} = \sqrt{\frac{s_1^{\,2}(n_1 - 1) + s_2^{\,2}(n_2 - 1)}{n_1 + n_2 - 2}}$$

ALTERNATIVES TO COHEN'S *d*

- **Hedges' *g*** [13]
  - *Small sample size*
  - *Weights the pooled standard deviation*

$$g = \frac{\bar{x}_1 - \bar{x}_2}{S^*_{pooled}}$$

- **Glass's Δ** [14]
  - *Treatment impacts standard deviation*
  - *Uses the standard deviation of the control group*

$$\Delta = \frac{\bar{x}_1 - \bar{x}_2}{S_{control}}$$

ODDS RATIO

- **The odds of injury for members of the treatment group were 4 times higher than odds for members of the control group**
  - NOT four times the number of injuries

	Injury	No Injury
Treatment Group	A	B
Control Group	C	D

$$\frac{AD}{BC}$$

- **Relative risk**
  - Probability of an event occurring in one group compared to the probability of the same event in another group

NORTHERN ARIZONA UNIVERSITY

r FAMILY

ONE-WAY ANOVA
BETWEEN SUBJECTS

- **Eta Squared**
  - Proportion of the variation in Y explained by X
  - Positively biased
    - Sample variance only, uncorrected
    - Less biased for larger samples (> 30) [4]

$$\eta^2 = \frac{SS_{between}}{SS_{total}}$$

- **Epsilon Squared**
  - Less biased than eta squared
    - Subtracting MSE

$$\varepsilon^2 = \frac{SS_{between} - (J-1)(MSE)}{SS_{total}}$$

- **Omega Squared**
  - Equal sample sizes
  - Less biased than epsilon squared
    - Adding MSE to SST in denominator

$$\omega^2 = \frac{SS_{between} - (J-1)(MSE)}{SS_{total} + MSE}$$

FACTORIAL ANOVA

- **Partial eta squared**
  - Proportion of variation in Y explained by the effect of interest
  - Default in SPSS
  - Results are the same for eta squared in one-way ANOVA

$$\eta_A^2 = \frac{SS_A}{(SS_A + SS_{within})} \qquad \eta_B^2 = \frac{SS_B}{(SS_B + SS_{within})} \qquad \eta_{AB}^2 = \frac{SS_{AB}}{(SS_{AB} + SS_{within})}$$

- **Partial omega squared**
  - Less biased estimator

$$\omega_A^2 = \frac{SS_A - (J-1)MS_{within}}{SS_{total} + MS_{within}} \qquad \omega_B^2 = \frac{SS_B - (K-1)MS_{within}}{SS_{total} + MS_{within}}$$

$$\omega_{AB}^2 = \frac{SS_{AB} - (J-1)(K-1)MS_{within}}{SS_{total} + MS_{within}}$$

REPEATED MEASURES

- **Entirely different set of effect sizes for repeated measures designs.[15]**
  - Olejnik S and Algina J. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psych Methods*. 2003;8(4):434-447.

RELATIONSHIPS

- **Pearson-product-moment correlation coefficient**
  - Two continuous variables

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{S_X S_Y}$$

- **Point-biserial correlation coefficient**
  - One dichotomous variable
  - One continuous variable

$$r_{pb} = \frac{\overline{X}_1 - \overline{X}_0}{S_n} \sqrt{\frac{n_1 n_0}{n^2}}$$

- **Spearman's rank correlation coefficient**
  - Two ordinal variables

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

REGRESSION

- **Coefficient of determination**
  - r^2
  - Simple linear regression
  - Amount of variance shared between the two variables

- **Coefficient of multiple determination**
  - R^2
  - Multiple linear regression
  - Amount of variance shared between the dependent variable and the set of independent variables

RECOMMENDATIONS FOR EFFECT SIZES

- **Choose the most suitable effect size based on the purpose, design, and outcome(s) of the study.** [16]

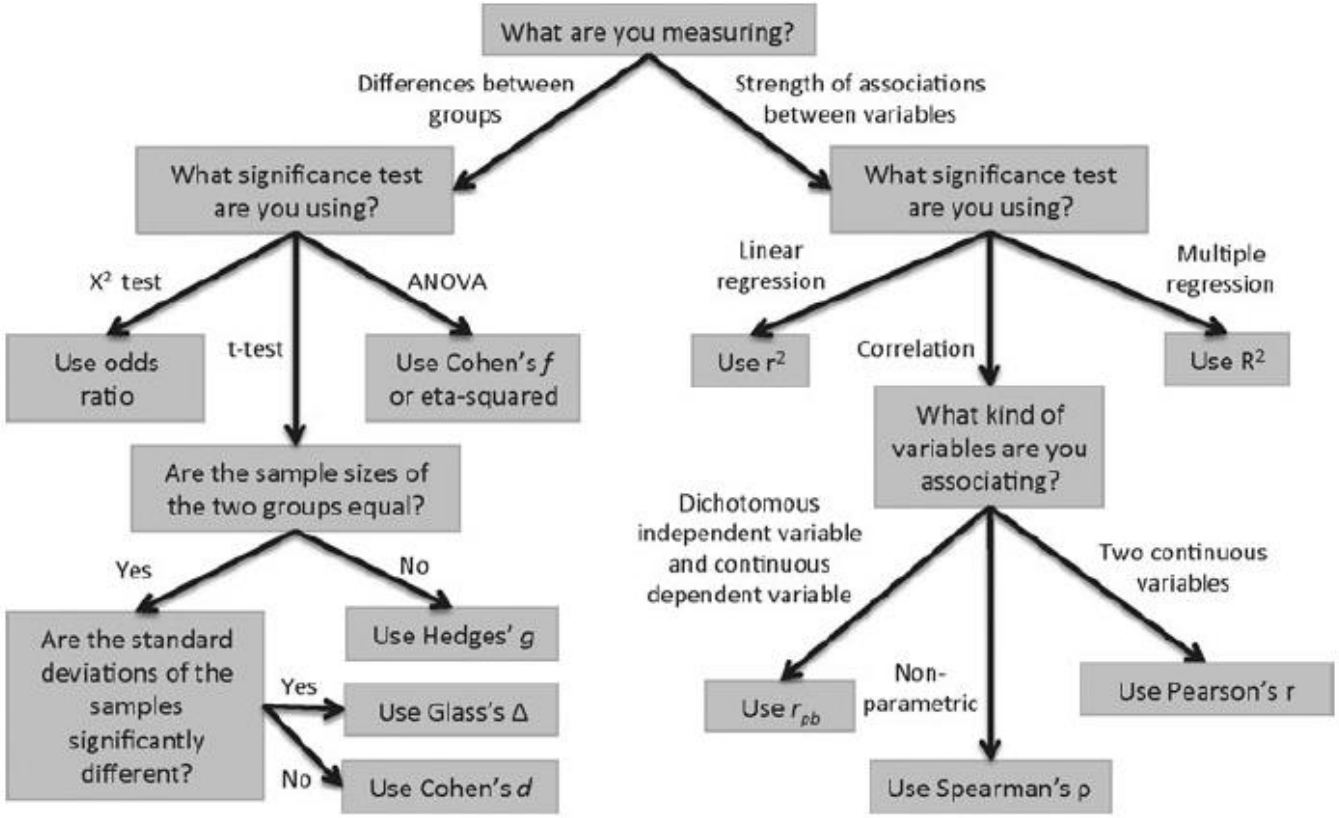
$$\eta^2 > \varepsilon^2 > \omega^2$$

- **Provide effect sizes whether or not a statistically significant finding is obtained.**

- **Specify exactly how effects were calculated.**

- **Caution when interpreting against a rigid benchmark because context matters so much.** [17]
  - Glass's caution to not classify effects into 't-shirt sizes' [18]
  - Rhea new classifications for strength training research [19]
    - < 0.35 trivial, 0.35-0.80 small, 0.80-1.50 moderate, and > 1.5 large

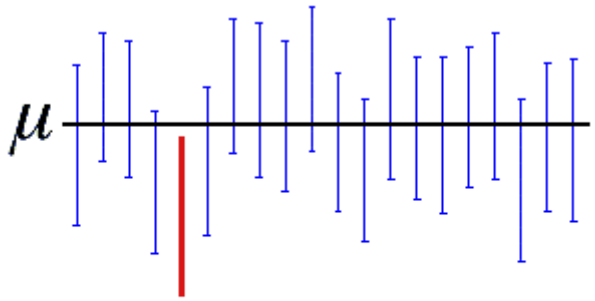
J. Middlemis Maher *et al.*



CONFIDENCE INTERVALS

CONFIDENCE INTERVALS

- **Many replications of the study, we would expect 95% of these intervals to include the population mean, or another parameter being estimated.**



A 95% confidence interval indicates that 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the population parameter.

http://blog.minitab.com/blog/adventures-in-statistics-2/understanding-hypothesis-tests%3A-confidence-intervals-and-confidence-levels

- **Interval estimate of a population parameter allowing us to determine the accuracy of the sample estimate.**
  - This interval is a set of values that are plausible for μ. Values outside the interval are relatively implausible but not impossible.[4]

CONFIDENCE INTERVALS

- If the CI contains zero → **no** statistically significant difference
- If the CI does not contain zero → statistically significant difference

- So much more information
  - **Precision of a population estimate**
    - Smaller the interval
      - Less sampling error
  - **Location of a population estimate**
    - Interpret from scale used in study
  - **Provide interpretation**

CI INTERPRETATION

- Difference in AROM (ankle-dorsiflexion) improvement following a 3-week intervention[9]
  - **95% CI (0.07°, 2.13°)**

- There is a statistically significant difference between groups.

- The difference for the population means could be as small as 0.07°, or as large as 2.13°, at the 95% confidence level. Due to the narrow CI, there was a smaller impact of sampling error.

- The researcher would have to decide if a possible difference of less than 1° improvement in the population is worth the extra time and expense involved in using the intervention.

DURING THE REVIEW PROCESS

- ***Were the variables clearly defined?***
- ***Did the author perform assumption testing?***
- ***How were the missing data handled?***
- ***For <u>all</u> hypothesis testing, where the degrees of freedom, test statistic, associated P-value, confidence interval, and effect size (with how this was calculated) presented?***
- ***What was the interpretation of the confidence interval(s) and effect size(s)?***

NORTHERN ARIZONA UNIVERSITY

https://stats.stackexchange.com/questions/423/what-is-your-favorite-data-analysis-cartoon?page=2&tab=votes#tab-top

ANY QUESTIONS?

Thank you!

monica.lininger@nau.edu

REFERENCES

1. Klockars AJ. Analysis of variance: Between groups designs. In Hancock GR and Mueller RO (Eds.). *The Reviewer's Guide to Quantitative Methods in the Social Sciences.* New York, NY: Routledge Taylor & Francis Group; 2010.
2. Keselman HJ, Huberty CJ, Lix LM, et al. Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA. *Rev Educ Res*. 1998;68:351.
3. Hoekstra R, Kiers HAL, Johnson A. Are assumptions of well-known statistical techniques checked, and why (not)? *Front Psych.* 2012;3:137-146.
4. Lomax RG and Hahs-Vaughn D. *An Introduction to Statistical Concepts*. 3rd ed. New York, NY: Routledge Taylor & Francis Group; 2012.
5. Schafer JL and Graham JW. Missing data: Our view of the state of the art. *Psychol Methods*. 2002;7(2):147-177.
6. Sterne, JAC and Davey Smith. Sifting the evidence: what's wrong with significance test *BMJ.* 2001;222:226-231.
7. Maher JM, Markey JC, and Ebert-May D. The others half of the story: effect size analysis in quantitative research. *CBE Life Sci Edu.* 2013;12:345-351.
8. Cobb G. In Wasserstein RL and Lazar NA. *The ASA's statements on P-values: Context, process, and purpose*. The American Statistician. 2016;70(2):129-133.
9. Riemann BL and Lininger MR. Statistical primer for athletic trainers: The difference between statistical and clinical meaningfulness. *J Athl Train*. 2015;50(12):1223-1225.
10. Durlack JA. How to select, calculate, and interpret effect sizes. *J Ped Psychol*. 2009;34(9):917-928.

REFERENCES CONT.

11. Kirk RE. The important of effect magnitude. In SF Davis (Ed.) *Handbook of Research Methods in Experimental Psychology*. Oxford, UK: Blackwell; 2003.
12. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1998:274-288.
13. Hedges LV. *Distribution theory for Glass's estimator of effect size and related estimators*. *J Edu Stat*. 1981;6(2):107-128.
14. Glass GV. Primary, secondary, and meta-analysis of research. *Edu Res*. 1976;5(10):3-8.
15. Olejnik S and Algina J. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychol Methods*. 2003;8(4):434-447.
16. Cumming G and Fidler F. Effect sizes and confidence intervals. In Hancock GR and Mueller RO (Eds.). *The Reviewer's Guide to Quantitative Methods in the Social Sciences.* New York, NY: Routledge Taylor & Francis Group; 2010.
17. Thompson. Research synthesis: Effect sizes. In Green J, Camilli G, Elmore PB (Eds.). *Handbook of complementary methods in education research.* Washington, DC: American Educational Research Association; 2006.
18. Glass GV, McGaw B, and Smith ML. *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage Publishers; 1981.
19. Rhea MR. Determining the magnitude of treatment effects in strength training research through the use of the effect size. *J Strength Cond Res*. 2004;18(4):918-920.