

Empirische Forschungsmethoden II-2

no need to p – Umgang mit nicht-signifikanten Ergebnissen

Auf geht's! Wir...

...(1) erkunden das Problem nicht-signifikanter p-Werte,

...und (2) was ein (nicht-signifikanter) p-Wert wirklich bedeutet,

...um (3) 6 Optionen abzuleiten, wie wir mit nicht-signifikanten Ergebnissen umgehen können (und diese informativ machen)

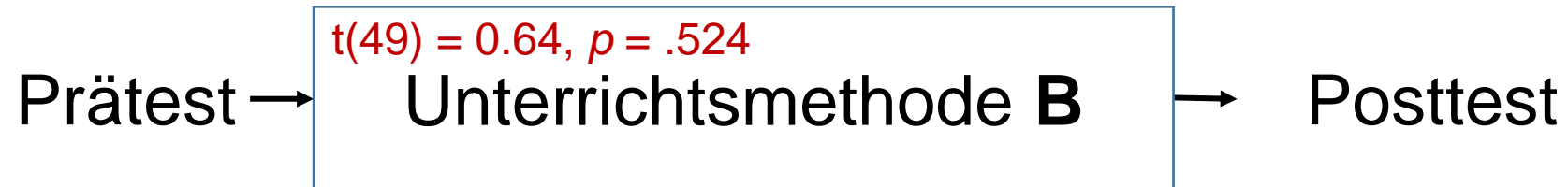
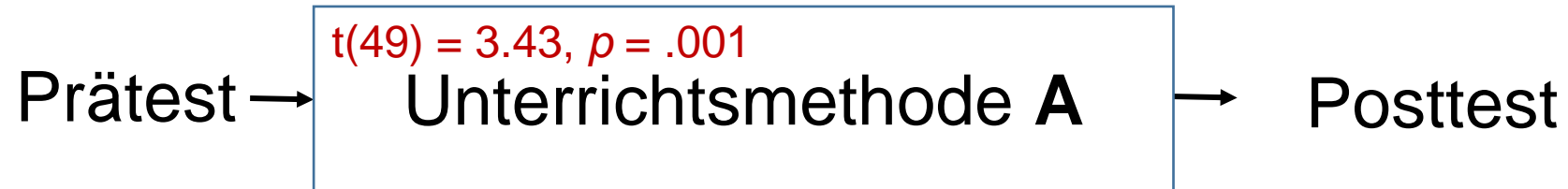
Abhängige Variable: Lernzugewinne (Posttest - Prätest)



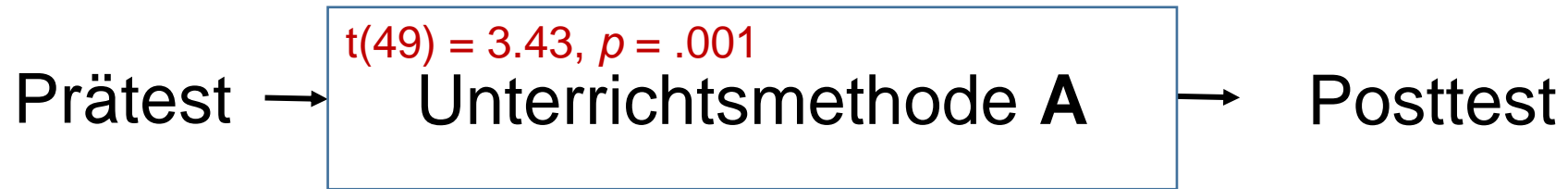
Was ist **deine** Interpretation?
Bearbeite die Aufgabe auf Qualtrics.



https://descil.eu.qualtrics.com/jfe/form/SV_6Jv15HjvkAFpRYy



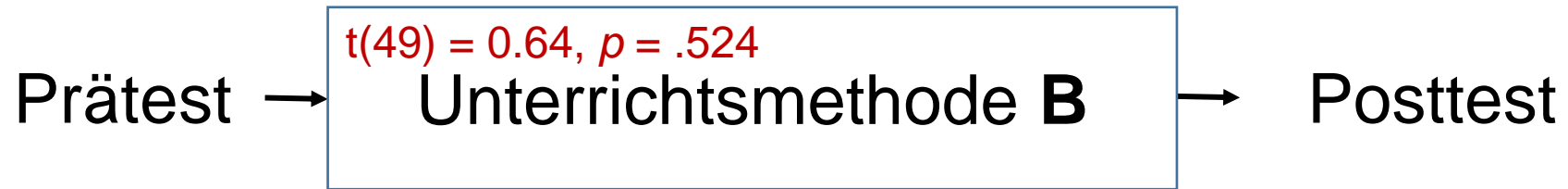
Was ist **deine** Interpretation?
Bearbeite die Aufgabe auf Qualtrics.



Die Forscherin fand dieses Ergebnis als sie einen t-Test für abhängige Stichproben einsetzte um zu überprüfen, ob unter Unterrichtsmethode **A** lernen stattfindet. *Notieren Sie Ihre Antworten zu den folgenden Fragen **anonym auf diesem Zettel**.*

Wie würden Sie den p-Wert interpretieren, welche Schlussfolgerungen erlaubt dieser?

Warum? Bitte erläutern Sie Ihre Antwort unter Bezug auf die Bedeutung des p-Wertes.



Die Forscherin fand dieses Ergebnis als sie einen t-Test für abhängige Stichproben einsetzte um zu überprüfen, ob unter Unterrichtsmethode **B** lernen stattfindet. *Notieren Sie Ihre Antworten zu den folgenden Fragen **anonym auf diesem Zettel**.*

Wie würden Sie den p-Wert interpretieren, welche Schlussfolgerungen erlaubt dieser?

Warum? Bitte erläutern Sie Ihre Antwort unter Bezug auf die Bedeutung des p-Wertes.

~~Es gab keine Lernzugewinne unter
Unterrichtsmethode B~~

~~Wir können nicht ausschließen, dass
die beobachteten Ergebnisse durch
Zufall zustande kamen.~~

~~Die Wahrscheinlichkeit, dass der
wahre Effekt 0 ist, ist $p = .524$~~

Ein nicht-signifikanter p-Wert läßt keine
informativen Schlussfolgerungen zu

Eine wichtige Frage ist somit aufgekomen:
Was zeigt ein p -Wert eigentlich an?

Webseite einfügen

Diese App ermöglicht Ihnen, sichere Webseiten, deren Adresse mit "https://" beginnt, in das Foliendeck einzufügen. Nicht sichere Webseiten werden aus Sicherheitsgründen nicht unterstützt.

Geben Sie unten die URL ein.

Hinweis: Viele beliebte Websites ermöglichen den sicheren Zugriff. Klicken Sie auf die Vorschaufläche, um zu überprüfen, ob auf die Webseite zugegriffen werden kann.

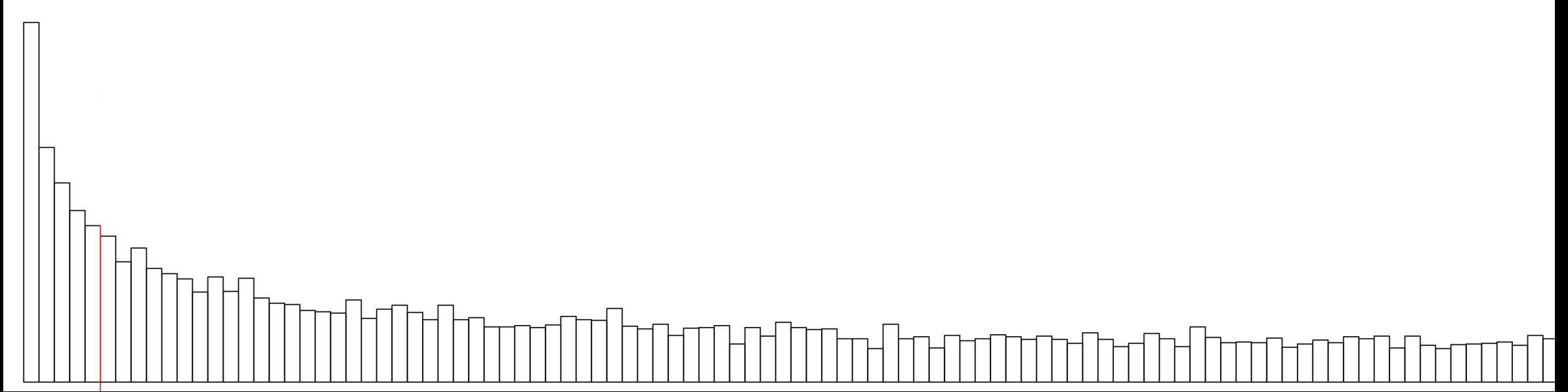
bit.ly/simplepvaluesim

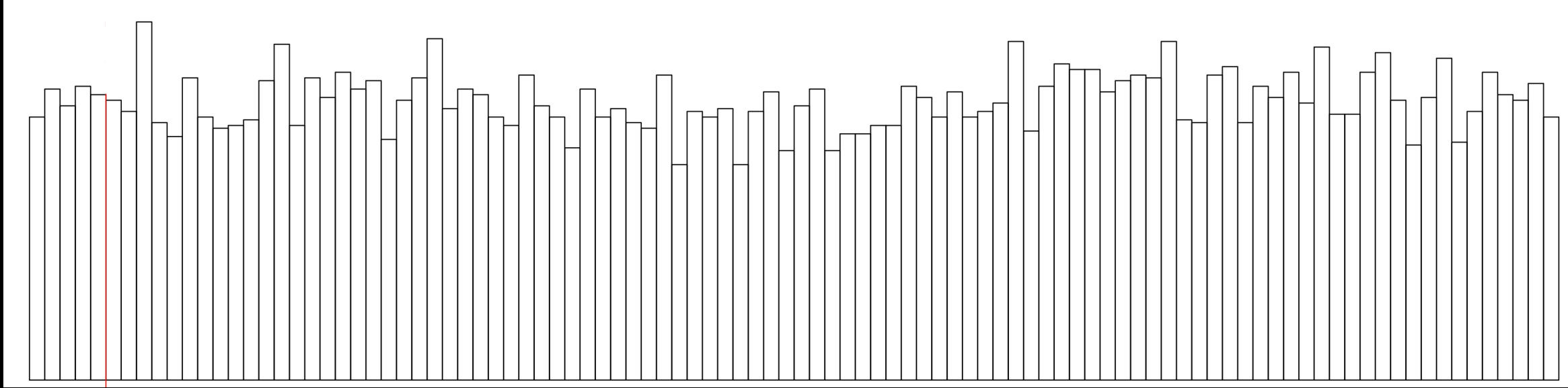
Ein p -Wert zeigt die *relative Häufigkeit* von Teststatistiken an, die von der H_0 mindestens so stark abweichen wie die beobachtete.

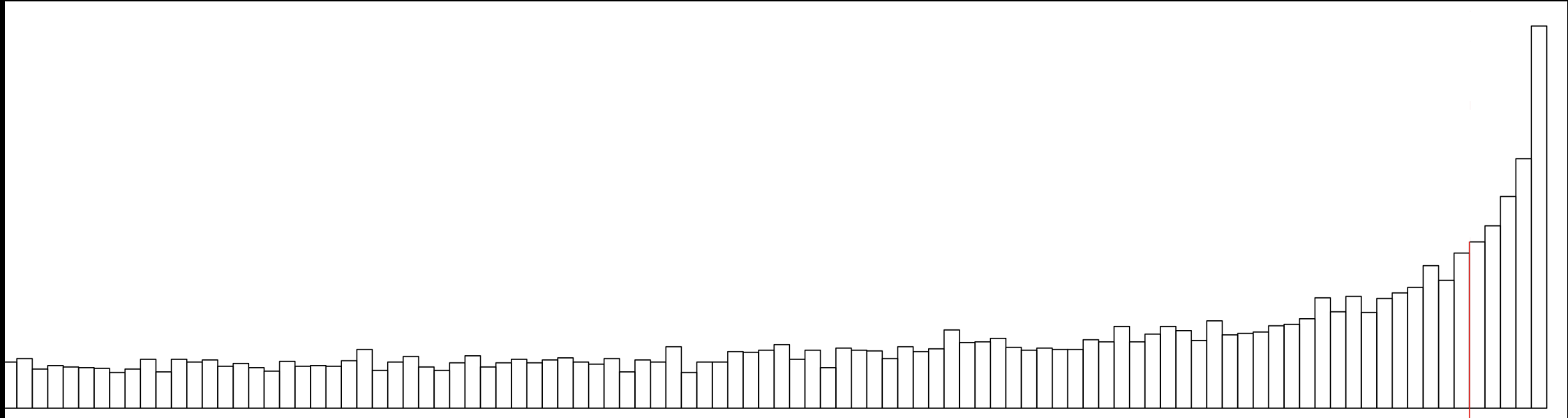
...er zeigt eine hypothetische *Seltenheit*.

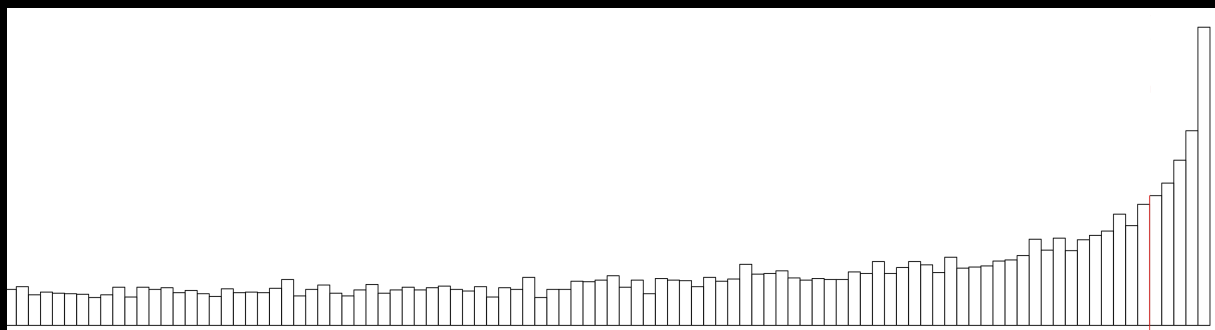
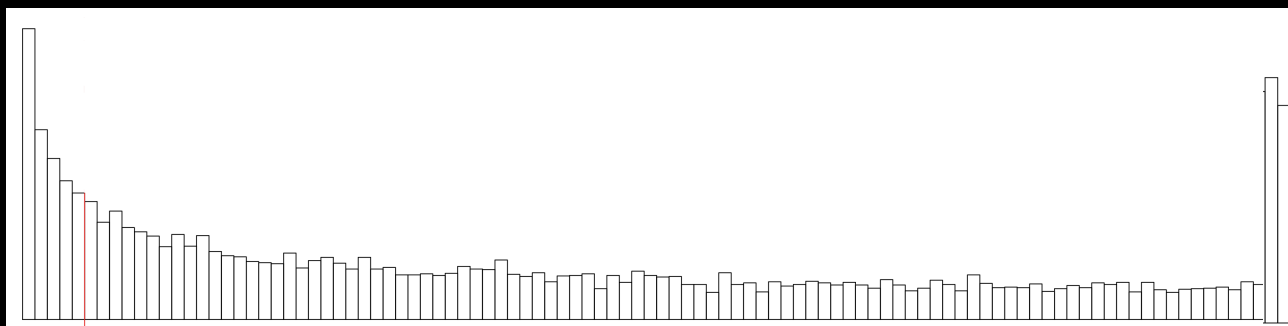
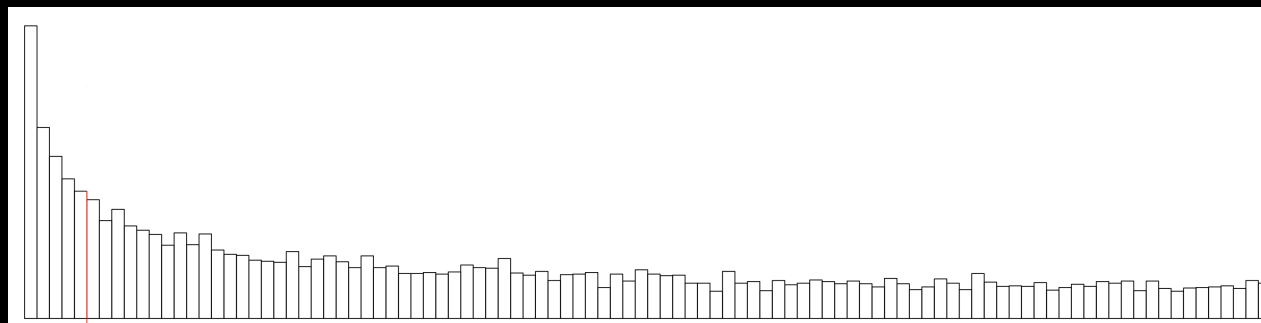
...er beschreibt keine Eigenschaft der beobachteten Daten, sondern von *hypothetischen Ergebnissen*.









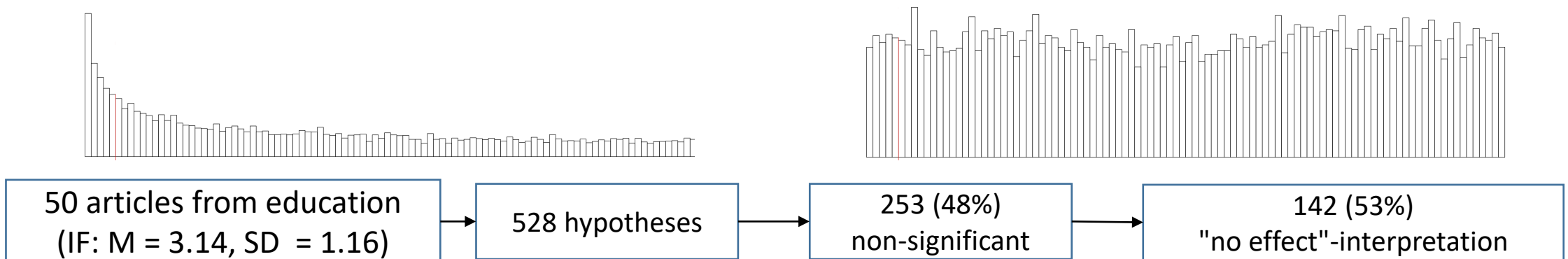


Ein nicht-signifikanter p-Wert alleine erlaubt gar keine Schlussfolgerungen, denn...

...bei Signifikanztestung **nehmen wir an, dass die H0 stimmt.**

...er könnte *entweder* zeigen, dass die H0 korrekt ist, *oder* einen Beta-Fehler (ein übersehener Effekt, der eigentlich vorhanden ist). **Wir wissen nicht, was davon der Fall ist.**

Wissen Forschende das? Edelsbrunner & Thurn (2024)



Zwei wichtige Fragen haben sich ergeben:

Was können wir mit nicht-signifikanten Ergebnissen tun?

Wie können wir einen Nulleffekt von einem zufällig nicht-signifikanten Ergebnis unterscheiden?

Was man mit nicht-signifikanten Ergebnisse tun kann: Fünf Möglichkeiten

- 1) Einfach angeben, dass die Evidenz uneindeutig ist
- 2) Effektgrößen (mit Konfidenzintervallen)
- 3) Äquivalenztestung
- 4) Bayes-Faktor
- 5) ROPE (region of practical equivalence – Region praktischer Äquivalenz)
- 6) Modellvergleiche

1) Einfach angeben, dass die Evidenz uneindeutig ist

“Der Lernzugewinn in unserer Intervention war nicht statistisch signifikant, $t(49) = 0.64$, $p = .524$, was bedeutet, dass die vorliegenden Daten keine Schlussfolgerungen über Hypothese B zulassen.”

...wir sind so klug als wie zuvor.

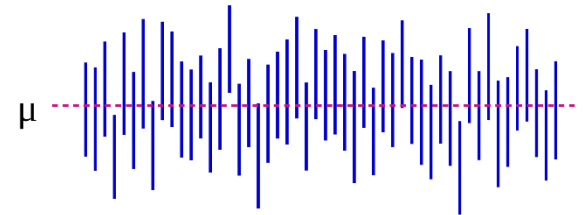
(wobei dies unter Nullhypothesen-Signifikanztestung die korrekte Vorgangsweise ist!)

2) Effektgrößen (und Konfidenzintervalle)

- Effektgrößen informieren uns über die Grösse eines Effektes
- Es gibt nicht nur Cohen's d , r und η^2 : <https://easystats.github.io/effectsize/articles/interpret.html>



Ein 95%-CI bedeutet:
In 95% der Fälle wird das
Konfidenzintervall den wahren
Populationsparameter beinhalten.



- Wir könnten schreiben:

"Unter Unterrichtsmethode B gab es keine statistisch signifikanten Lernzugewinne, $t(49) = 0.64$, $p = .524$, mit einer geschätzten Effektgrösse von Cohen's $d = 0.09$, $CI_{90}[-0.15; 0.33]$. Es handelt sich also um einen kleinen Effekt."

Wie können wir diese Effektgrösse interpretieren?

$d = 0.20$ ist ein kleiner Effekt, $d = 0.50$ ein mittelgrosser und $d = 0.80$ ein grosser Effekt.
Aber: Manchmal sind auch kleinere oder grössere Effekte zu erwarten!

2) Effektgrößen (und Konfidenzintervalle): Weitere Interpretationsrahmen (Edelsbrunner & Thurn, 2024)

Category	Frame of Reference
Relative comparisons	Typical effect sizes of educational interventions
	Comparison to effects from similar prior interventions
	Comparison to regular growth during an academic period
Impact	Developmental impact (effects on outcome variable of interest in the long run)
	comparison to policy-relevant achievement gaps between groups (e.g., impact of intervention on gender differences or decrease between at-risk students and not-at-risk students, absolute decrease in number of at-risk students)
	Percentile rank changes on a standardized test instrument
	Changes in probability to score above a certain relevant proficiency-threshold
Resources	Duration
	Implementaion effort
	Costs
Theoretical references	Cognitive processes involved (e.g. far transfer across contexts)
	Tests of the generalizability and robustness of effects
	The stability of effects/transfer over time
Study design	Nature of the control group (e.g., a waiting list-condition or an active control group differing only in a key element)
	Specifics of the assessment instruments (e.g., non-standardized or standardized instruments)
	Sample size: small-scale experiment or large-scale implementation
	The predicted scalability of an intervention to larger population

3) Äquivalentestellung

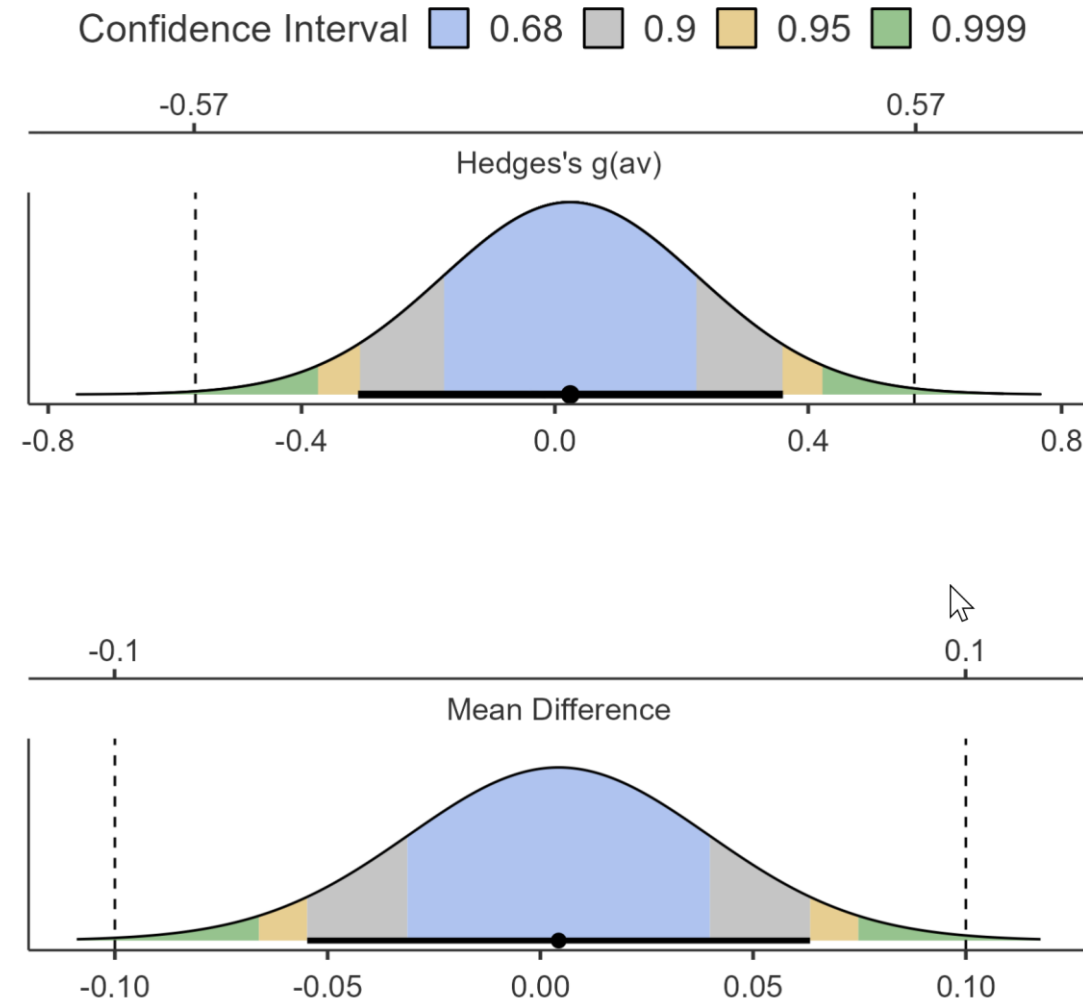
Äquivalenztestung

- ...wurde für Biopharmazeutische Studien entwickelt (erste Referenzen: Westlake 1972; Anderson & Hauck 1983; Selwyn & Hall 1984; Schuirmann 1987)
- ...ist nicht ein Test sondern eine Familie von Tests
- ...testet die Hypothese: «der Effekt ist null oder sehr klein»
- ...verwendet zwei t-Tests (TOST) oder 90%-Konfidenzintervalle
 - 1) definiere den kleinsten interessierenden Effekt δ (aka SESOI)
 - 2) definiere die Hypothese(n): $H_{01} : \mu_1 - \mu_2 \geq \delta; H_{02} : \mu_1 - \mu_2 \leq -\delta,$
 - 3) teste gegen diese Hypothesen

2) Äquivalenztestung: Interpretationsrahmen zur Festlegung des kleinsten int. Effektes

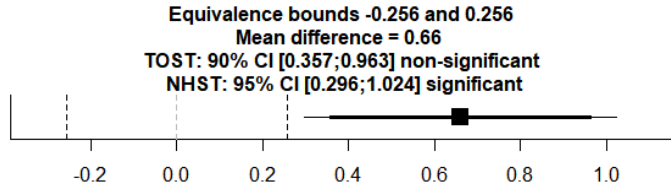
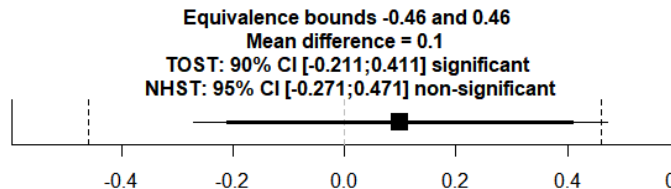
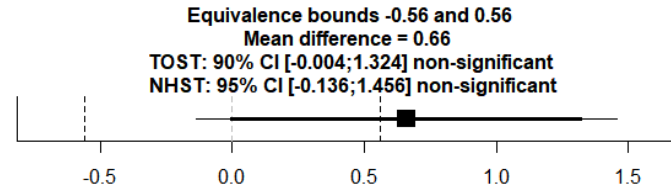
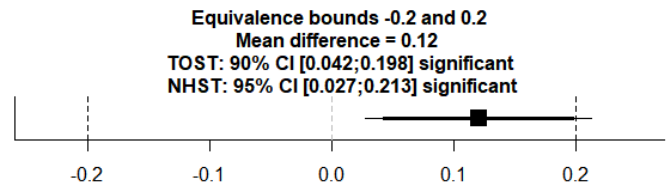
Category	Frame of Reference
Relative comparisons	Typical effect sizes of educational interventions
	Comparison to effects from similar prior interventions
	Comparison to regular growth during an academic period
Impact	Developmental impact (effects on outcome variable of interest in the long run)
	comparison to policy-relevant achievement gaps between groups (e.g., impact of intervention on gender differences or decrease between at-risk students and not-at-risk students, absolute decrease in number of at-risk students)
	Percentile rank changes on a standardized test instrument
	Changes in probability to score above a certain relevant proficiency-threshold
Resources	Duration
	Implementaion effort
	Costs
Theoretical references	Cognitive processes involved (e.g. far transfer across contexts)
	Tests of the generalizability and robustness of effects
	The stability of effects/transfer over time
Study design	Nature of the control group (e.g., a waiting list-condition or an active control group differing only in a key element)
	Specifics of the assessment instruments (e.g., non-standardized or standardized instruments)
	Sample size: small-scale experiment or large-scale implementation
	The predicted scalability of an intervention to larger population

Beispiel: R-Paket TOSTER



3) Äquivalenztestung

Mittels Hypothesentestung und Äquivalenztestung können wir vier Fälle unterscheiden:

Signifikanztest	Äquivalenz	Bedeutung	Beispielhafte Abbildung
Signifikant	Nicht signifikant	Evidenz gegen H_0	 <p>Equivalence bounds -0.256 and 0.256 Mean difference = 0.66 TOST: 90% CI [0.357; 0.963] non-significant NHST: 95% CI [0.296; 1.024] significant</p>
Nicht signifikant	Signifikant	Evidenz gegen H_1	 <p>Equivalence bounds -0.46 and 0.46 Mean difference = 0.1 TOST: 90% CI [-0.211; 0.411] significant NHST: 95% CI [-0.271; 0.471] non-significant</p>
Nicht signifikant	Nicht signifikant	Inkonklusive evidenz / uneindeutiges Ergebnis	 <p>Equivalence bounds -0.56 and 0.56 Mean difference = 0.66 TOST: 90% CI [-0.004; 1.324] non-significant NHST: 95% CI [-0.136; 1.456] non-significant</p>
Signifikant	Signifikant	Evidenz gegen H_0 aber der Effekt ist wahrscheinlich nicht bedeutungsvoll	 <p>Equivalence bounds -0.2 and 0.2 Mean difference = 0.12 TOST: 90% CI [0.042; 0.198] significant NHST: 95% CI [0.027; 0.213] significant</p>

3) Äquivalenztestung: Interpretation

Wir könnten schreiben:

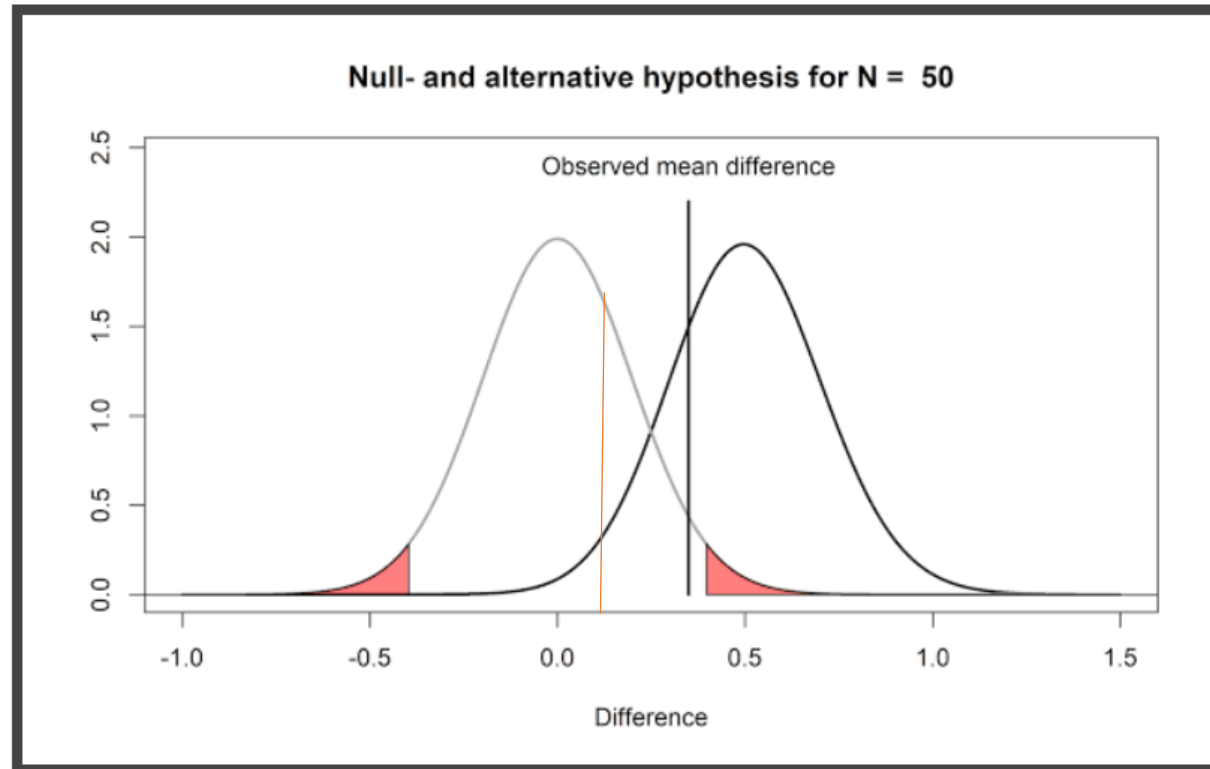
“Die Differenz vom Prä- zum Posttest war nicht statistisch signifikant $t(59) = 1.73$, $p = .089$, $d = .09$.
Für eine Äquivalenztestung wurden die Äquivalenzgrenzen $d = -0.45$ and $d = -0.45$ festgelegt.
Ein Äquivalenztest mit diesen Grenzen war signifikant. Wir können die Hypothese ablehnen, dass der Effekt größer als $d = 0.45$ oder kleiner als $d = -0.45$ ist und davon ausgehen, dass der wahre Effekt von vernachlässigbarer Größe ist.”

R Pakete für Äquivalenztestung:

- equivalence <https://cran.r-project.org/web/packages/equivalence/equivalence.pdf>
- TOSTER <https://cran.r-project.org/web/packages/TOSTER/TOSTER.pdf>
- EQUIVNONINF <https://cran.r-project.org/web/packages/EQUIVNONINF/EQUIVNONINF.pdf>
- emmeans <https://www.rdocumentation.org/packages/emmeans/versions/1.4.5/topics/emmeans>
- equivalencetests <https://github.com/cribbie/equivalencetests>

4) Bayes-Faktor

- Ein Bayes Faktor gibt an, wie stark Daten für ein Modell oder für ein anderes sprechen.



Der p -Wert zeigt uns nur, ob das Ergebnis sehr überraschend ist, wenn es in Wahrheit keinen Effekt gibt. Ein nicht-signifikanter p -Wert bedeutet nicht, dass die Null-Hypothese wahr ist. Dies könnte zwar sein, aber es ist auch möglich, dass die Daten, die wir erhalten haben, unter der Alternativhypothese wahrscheinlicher sind, als unter der Nullhypothese (wie in der Abbildung oben).

4) Bayes-Faktoren

- Bayes-Faktoren drücken die Plausibilität eines Modells im Vergleich zu einem anderen aus.

Independent Samples T-Test

		Statistic	±%	df	p
Eingebundenheit_Mittelwert	Student's t	0.121		100	0.904
	Bayes factor ₁₀	0.214	2.47e-4		

Note. $H_a: \mu_{\text{Eingebundenheit|Selbstkonzept}} \neq \mu_{\text{Selbstkonzept|Eingebundenheit}}$

BF	Evidence for	Evidence
30	H_1	Strong
10	H_1	Moderate
3	H_1	Anecdotal
1		
1/3	H_0	Anecdotal
1/10	H_0	Moderate
1/30	H_0	Strong

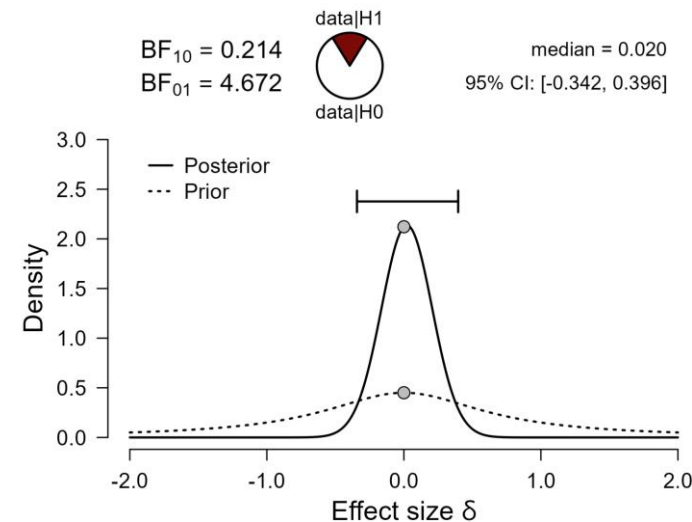
Hoyda et al., 2019

5) Region praktischer Äquivalenz (ROPE)

- Das Bayesianische Analog zum Konfidenzintervall ist das Kredititätsintervall (CI; *credible interval*)
- wenn das 90% (oder welche Genauigkeit wir bevorzugen) CI innerhalb der ROPE liegt, dann nehmen wir Äquivalenz an.



95% credible interval: Bereich in welchem mit 95%er Sicherheit der wahre Wert liegt.



6) (relative) Modellvergleiche

- Basieren auf Likelihood (Modellfit)
- Welches Modell erreicht die optimale Balance aus Fit und Sparsamkeit?



Es gibt so viele unterschiedliche relative Fit-Indizes, weil diese die Komplexität des Modells unterschiedlich einbeziehen und gewichten.

Sechs Optionen zum Umgang mit nicht-signifikanten Ergebnissen

1) Einfach uneindeutige Evidenz berichten

“The difference from pre- to posttest was not statistically significant $t(49) = 0.64, p = .524$, meaning that our data yielded inconclusive evidence regarding hypothesis B.”

?

2) Effektgrößen (mit Konfidenzintervallen)

“The difference from pre- to posttest was not statistically significant $t(49) = 0.64, p = .524, d = .45 [CI_{95} 0.18, 0.71]$, with the effect size estimate and its confidence interval indicating that the effect was of small to medium size.”

?

3) Äquivalenztestung

“The difference from pre- to posttest was not statistically significant $t(49) = 0.64, p = .524, d = .45$.
A TOST equivalence test with default alpha = 0.05 and equivalence bounds of $d = -0.45$ and $d = 0.45$ was significant. We can reject the hypothesis that the effect was larger than $d = 0.45$ ”

H_0

4) Bayes Faktor

“The difference from pre- to posttest was not statistically significant $t(49) = 0.64, p = .524, d = .45$.
The Bayes Factor of 0.57 using uninformative priors indicated that the data is about two times more likely under the null hypothesis than under the alternative hypothesis. This result represents anecdotal evidence for the null hypothesis.”

H_0

5) ROPE (region of practical equivalence)

“The difference from pre- to posttest was not statistically significant $t(49) = 0.64, p = .524, d = .45$.
The Bayesian Region of Practical Equivalence from -0.5 to .5 points learning gain covered only 54% of the posterior distribution and left inconclusive evidence whether the effect was equivalent to zero or not.”

H_0

6) Modellvergleiche

“Sowohl der AIC als auch der BIC sprachen dafür, dass das Modell ohne einen Effekt der Bedingung eine bessere Balance aus Modellfit und Komplexität darstellt und deshalb gegenüber dem Modell mit jenem Effekt zu präferieren ist.”

H_0