Methodischer Exkurs I: Experimentelles Vorgehen, Experimentelle Designs, Deskriptive Statistik, und Signifikanzprüfung

pruefung@ifv.gess.ethz.ch

Wissenschaftliches Vorgehen

Ziel wissenschaftlicher Untersuchungen:

- Den Wahrheitsgehalt und die Aussagekraft von Erklärungen prüfen
- Entscheidung zwischen konkurrierenden Erklärungen

Wissenschaftliches Vorgehen

Todsünden in der Wissenschaft: Keine Unterscheidung zwischen

- Befund
- Erklärung
- Evidenz für Erklärung

Beispiel: Geschlechtsunterschiede in mathematischen und naturwissenschaftlichen Leistungen

- Befund: Wo genau zeigen sich Unterschiede (Mittelwert, Anteil an Spitzenleistungen)
- Erklärungen: 1) Unterschiede in den kognitiven Basisvoraussetzungen 2) Fehlede positive Modelle für Schülerinnen
- Evidenz für die beiden sich nicht ausschliessenden Erklärung erfordern weitere Untersuchunen

Wissenschaftliches Vorgehen

- Fragestellung
- Theorie (Kriterien für gute Theorien nach Popper: Falsifizierbarkeit, Interne Konsistenz, Explanatorischer Gehalt, Einfachheit)
- Hypothese

Wissenschaftliches Vorgehen

Nach der Ausformulierung der Fragestellung:

- Wahl des Untersuchungsdesigns
- Datenerhebung
- Datenauswertung und statistische Prüfung
- Interpretation
- Nicht selten: Präzisierung der Fragestellung und Planung neuer Untersuchung

Experimentelles Vorgehen

- Unabhängige Variable (UV)
  - Variable, die vom Wissenschaftler gewählt und variiert wird
  - Hypothetischer Einflussfaktor
  - Mindestens zwei Abstufungen oder Ausprägungen

- Abhängige Variable (AV)
  - Laut Hypothese von der UV beeinflusste Variable
  - Muss quantifizierbar sein

UV: Gruppen im experimentellen Design

- Unabhängige Gruppendesigns: Jeder Versuchsteilnehmer wird einer Gruppe nach dem Zufall zugeordnet und ein Messwert erhoben
- Messwiederholungsdesigns: Von jedem Versuchsteilnehmer werden zwei oder mehr Messungen unter verschiedenen Bedingungen (also den verschiedenen Ausprägungen der UV) erhoben
- Quasiexperimentelle Designs: UV wird nicht geschaffen, sondern vorgefunden (z.B. Geschlecht, Bildungshintergrund der Eltern, besuchter Schultyp)
  - Keine zufällige Zuordnung der Versuchsteilnehmer zu den Gruppen
  - GEFAHR DER KONFUNDIERUNG MIT ANDEREN VARIABLEN (z.B. Bildungsgrad und Einkommen der Eltern)
  - Kann statistisch herausgerechnet werden, wenn Variablen bekannt und erhoben

Formulierung der Fragstellung bzw. Hypothesen

- Bedingungen der Falsifizierung spezifizieren: Wie müssen Daten aussehen, wenn die Hypothese nicht beibehalten werden kann?
- ALTERNATIVERKLÄRUNGEN beachten (sonst tuns die Gegner)
- Gegen sich arbeiten, es also der Hypothese maximal schwer machen

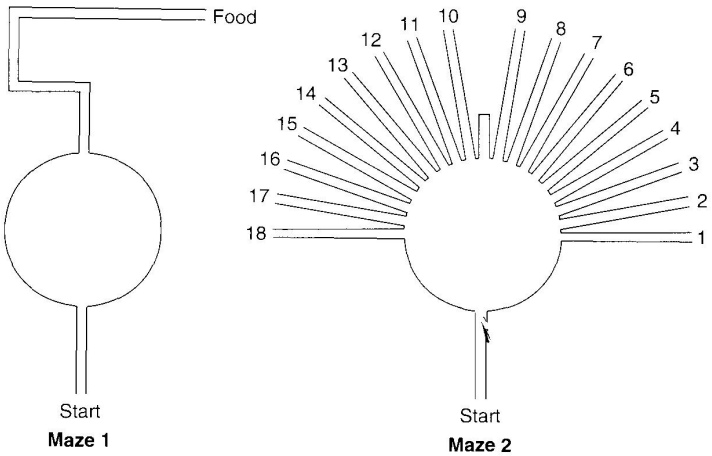


Figure 8–2
Mazes used by Tolman, Ritchie, and Kalish (1946).

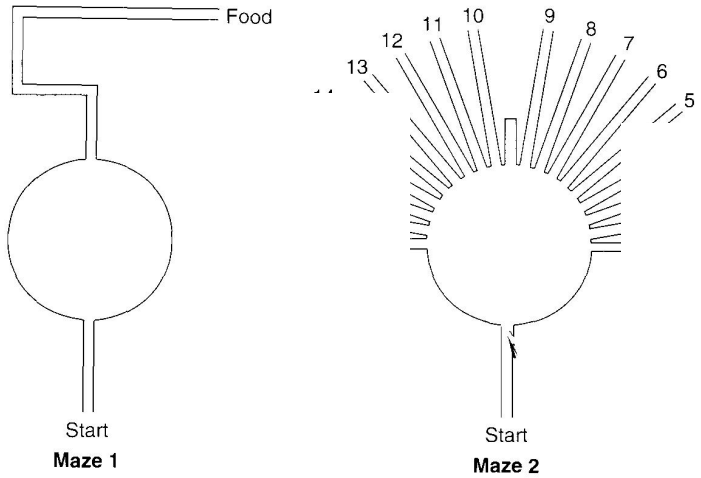


Figure 8–2
Mazes used by Tolman, Ritchie, and Kalish (1946).

AV: Operationalisierung

- Ein quantifizierbares messbares Merkmal muss festgelegt werden (z.B. Herzrate für Angst)
  Unabhängige Gruppendesigns: Jeder Versuchsteilnehmer wird einer Gruppe zugeordnet und ein Messwert erhoben
- Messwiederholungsdesigns: Von jedem Versuchsteilnehmer werden zwei oder mehr Messungen unter verschiedenen Bedingungen (also den verschiedenen Ausprägungen der UV) erhoben

Skalenniveau der Messwerte: Was darf man berechnen bzw. interpretieren?

- Nominalskalenniveau: Zahlen dienen nur der Differenzierung, die Grösse (Mächtigkeit): Geschlecht m=0, w=1
- Ordinalskalenniveau: Mit den Zahlen werden >< Beziehungen ausgedrückt, Abstände zwischen Zahlen dürfen nicht interpretiert werden (z.B. Schulnoten)
- Intervallskala: Zahlen drücken auch Abstände zwischen den Messwerten aus, aber keine Verhältnisse, da es keinen absoluten Nullpunkt gibt. Bespiel: Temperatur gemessen in Celcius oder Fahrenheit, Werte aus Tests, die nach wissenschaftlichen Kriterien konstruiert wurden (z.B. Intelligenztests, Schulleistungstests)
- Verhältnisskalenniveau: Zahlen drücken Verhältnisse aus, es gibt einen absoluten Nullpunkt (Masse, Grösse, Volumen, Geld)

Skalenniveau der Messwerte: Was darf man berechnen bzw. interpretieren?

- Nominalskalenniveau: Zahlen dienen nur der Differenzierung, die Grösse (Mächtigkeit): Geschlecht m=0, w=1
- Ordinalskalenniveau: Mit den Zahlen werden >< Beziehungen ausgedrückt, Abstände zwischen Zahlen dürfen nicht interpretiert werden (z.B. Schulnoten)
- **Intervallskala: Zahlen drücken auch Abstände zwischen den Messwerten aus, aber keine Verhältnisse, da es keinen absoluten Nullpunkt gibt. Bespiel: Temperatur gemessen in Celcius oder Fahrenheit, Werte aus Tests, die nach wissenschaftlichen Kriterien konstruiert wurden (z.B. Intelligenztests, Schulleistungstests)**
- Verhältnisskalenniveau: Zahlen drücken Verhältnisse aus, es gibt einen absoluten Nullpunkt (Masse, Grösse, Volumen, Geld)

Skalenniveau der Messwerte: Was darf man berechnen bzw. interpretieren?

- **Intervallskala: Zahlen drücken auch Abstände zwischen den Messwerten aus, aber keine Verhältnisse, da es keinen absoluten Nullpunkt gibt. Bespiel: Temperatur gemessen in Celcius (C) oder Fahrenheit (F)**
- F=1.8 C + 32

- **Werte aus Tests, die nach wissenschaftlichen Kriterien konstruiert wurden (z.B. Intelligenztests, Schulleistungstests)**

Bearbeiten von Messdaten: Aggregation der Einzelwerte zu interpretierbaren Parametern

- Parameter der zentralen Tendenz:
  - Modalwert: der am häufigsten vorkommende Wert (einzige Möglichkeit für nominalskalierte Daten)
  - Median: genau in der Mitte liegender Wert, wenn alle gemessenen Einheiten in eine Rangreihe gebracht werden (das höchste, was man mit ordinalskalierten Daten berechnen darf)
  - Mittelwert M (setzt Intervallskalenniveau voraus):

$$M = \frac{\sum_{i=1}^{n} i}{n}$$

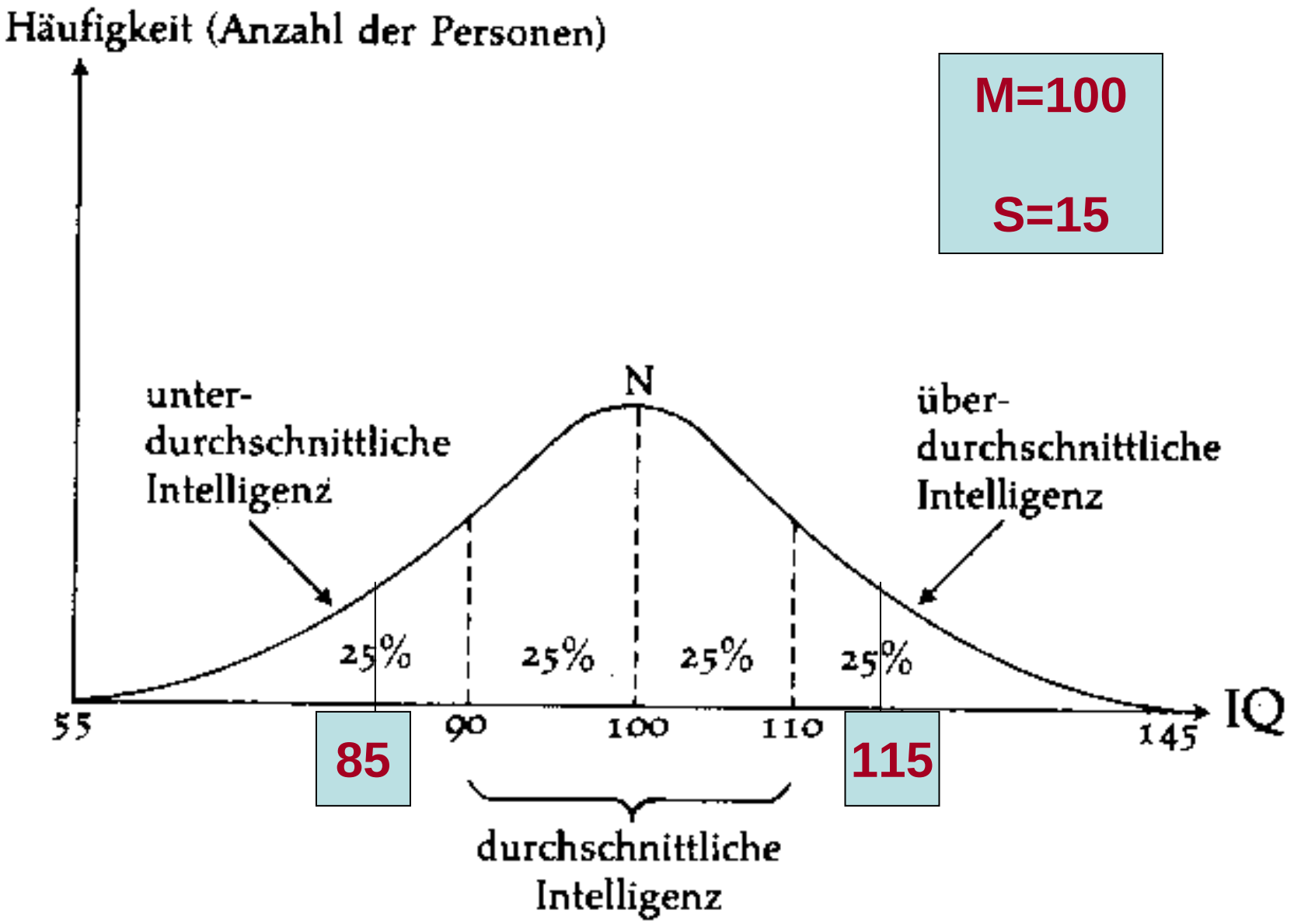
Dispersionsparameter:Gibt Aufschluss über die Heterogenität in der untersuchten Gruppe

Varianz (V)

$$V = \frac{\sum_{i=1}^{n} (x_i - M)^2}{n}$$

Standardabweichung (s)

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - M)^2}{n}}$$


Häufigkeit (Anzahl der Personen)
M=100
S=15
N
unter-
durchschnittliche
Intelligenz
über-
durchschnittliche
Intelligenz
25%
25%
25%
25%
55
85
90
100
110
115
145
IQ
durchschnittliche
Intelligenz

Warum erreichen manche Leistungstest nur Ordinalskalenniveau und andere Intervallskalenniveau?

- Wissenschaftlich fundierte Testkonstruktion setzt Skalierung der Aufgaben durch mathematische Modellierung voraus. Es muss sicher gestellt sein, dass jede Aufgabe (Item) mit gleichem Gewicht in den Gesamtwert eingeht
- Bei gut skalierten Tests mit einer grösseren Anzahl von Aufgaben (Items) zeigt sich eine Normalverteilung des Gesamtwertes

Formulierung der Hypothesen

- Inhaltliche Hypothese (Alternativhypothese, H1)
  - gerichtet (mehr-weniger) oder ungerichtet (anders)
  - spezifisch (Ausmass wird quantifiziert) oder unspezifisch (mehr-weniger)
- Nullhypothese (H0): die zur Alternativhypothese komplementäre Aussage trifft zu
- Nullhypothese keine Inhaltshypothese

Signifikanzaussagen

- Was heisst: Das Ergebnis ist signifikant?
- Für jeden Datensatz muss man herausfinden, mit welcher Wahrscheinlichkeit er auch durch Zufall hätte zustande kommen können
- Fehlerwahrscheinlichkeit (Signifikanzniveau)
- Konvention: Fehlerwahrscheinlichkeit wird auf höchstens 5% festgelegt
- WICHTIG: Signifikanzniveau sagt nichts über die Qualität und Wahrheitsgehalt der Hypothesen und Theorien aus

Warum kann man Ergebnisse nicht einfach so nehmen wie sie sind?

- Messungen im Verhaltensbereich (und nicht nur dort) sind immer mit Fehlern behaftet

- Ausserdem gibt es natürliche Variation der Merkmalsträger, deshalb Problem des Stichprobenfehlers

- Wenn man Aussagen über eine Gruppe von Merkmalsträgern (z.B. Menschen) machen möchte, muss die untersuchte Gruppe eine repräsentative Stichprobe

Signifikanzprüfung: Vergleich mit dem Zufall

- Hypothese: Eine Kupfermünze wurde gezinkt, indem unter der Zahlseite etwas Gold (also schweres Metall) angebracht wurde als unter der Kopflseite, damit „Kopf oben“ häufiger vorkommt als „Zahl oben“
- Die Münze soll nicht aufgesägt werden
- Es wird geprüft, ob „Kopf oben“ wirklich häufiger vorkommt als „Zahl oben“
- Frage: Wie häufig muss die Münze geworfen werden, damit davon ausgehen kann, dass diese gezinkt ist?

Prüfverteilung: zufälliges Auftreten von Ereignissen

- Frage: Wie viele Kombinationen von Ereignissen, die p Ausprägungen annehmen können, treten bei n Durchgängen auftreten?

$$p^n$$

Prüfverteilung: zufälliges Auftreten von Ereignissen

- Frage: Wie viele Kombinationen von Ereignissen, die p Ausprägungen annehmen können, treten bei n Durchgängen auf?

$$p^n$$

- Frage: Wie viele Kombinationen können bei n Münzwürfen auftreten?

$$2^n$$

Anzahl der Würfe, mögliche Kombinationen und deren Auftretenswahrscheinlichkeiten p

- 1: K,Z p=.50
- 2: KK,ZZ,ZK,KZ p=.25
- 3: KKK,KKZ,KZZ,ZZZ,ZZK,ZKK,ZKZ,KZK

P=.16

Wie wahrscheinlich ist s, dass bei 4 Würfen 4x Kopf auftritt? P=.08

Wie wahrscheinlich ist es, dass bei 5 Würfen 5x Kopf auftritt? P=.04

- Frage: Wie häufig muss die Münze geworfen werden, damit man davon ausgehen kann, dass sie gezinkt ist?
- Präzisierung: Je stärker der angenommene Effekt ist, um so seltener muss geworfen werden.
- Wird angenommen, dass die Münze so stark gezinkt ist, dass nur noch Kopf auftritt, reichen bereits 5 Würfe aus. Die Hypothese wäre bereits falsifiziert, wenn einmal Zahl aufträte.
- Die Wahrscheinlichkeit, dass zufällig 5x Zahl kommt, liegt unter 5%.

- Wird angenommen, dass die Münze nur leicht gezinkt ist und deshalb in 60% der Würfe Kopf kommt, müssen 100 Würfe durchgeführt werden, um zu zeigen, dass die Häufigkeit, mit der Kopf kommt, vom zufällig zu erwartenden Wert abweicht (Fehlerwahrscheinlichkeit <.05)
- Bei 50 Würfen wäre die Wahrscheinlichkeit, dass in 60% der Fälle (also bei 30) Kopf oben vorkommt 10%.

Signifikanzprüfung

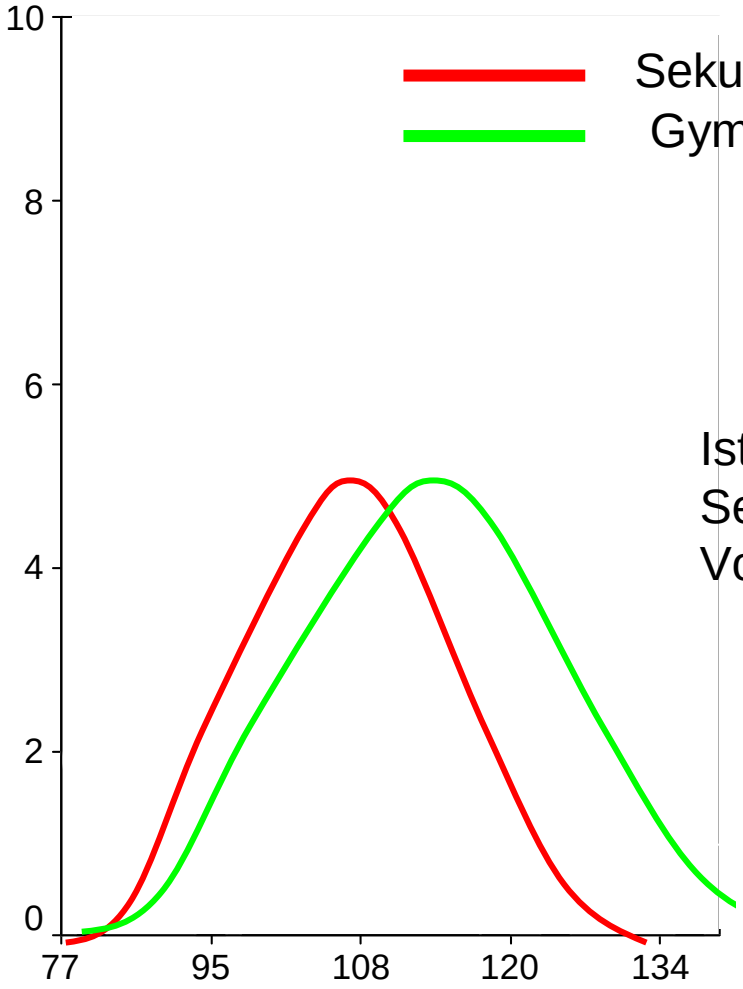
- Nochmal: Signifikanzprüfung heisst, die Wahrscheinlichkeit ermitteln, mit der ein gegebener Datensatz zufällig zustande gekommen ist und nicht durch den angenommenen Einflussfaktor (UV).
- Signifikanzprüfung heisst NICHT: Den richtigen Einflussfaktor ermittelt haben
- Beispiel Münze: ob es wirklich Gold ist, oder nicht doch Platin, das zum Zinken verwendet wurde, wird natürlich nicht geprüft.

Signifikanzprüfung bei Gruppenunterschieden (UV)

- Unterschied **zwischen** den Gruppen wird verglichen mit dem Unterschied **innerhalb** der Gruppen

- Die gleiche Gruppe wird mehrfach gestestet: Messwiederholungsdesign

  – Frage: Gibt es eine Veränderung?



Ist der durchschnittliche IQ von Sekundarschülern niedriger als Von Gymnasiasten

(hypothetische Daten)

Signifikanzprüfung bei Gruppenunterschieden (UV)

- Irrtumswahrscheinlichkeit (Signifikanzniveau): Wahrscheinlichkeit, dass die Gruppenunterschiede zufällig zustande gekommen sind, z.B. weil ein Stichprobenfehler vorlag
- Was ein signifikantes Ergebnis wahrscheinlich macht:
  - Grosse Anzahl von Beobachtungen (Versuchspersonen), da Stichprobenfehler unwahrscheinlicher wird
  - Geringe Unterschiede in der AV innerhalb der Gruppen
  - Grosse Unterschiede zwischen den Gruppen

Effektstärke

- Ein Gruppenunterschied kann signifikant sein, weil eine sehr grosse Stichprobe genommen wurde, oder weil der Ausgangswert sehr niedrig war. Dennoch können die Unterschiede so gering sein, dass sie praktisch nicht bedeutsam sind.
- Effektstärke ist eine statistische Kenngrösse, in die die Unterschiede zwischen den Gruppen (oder den unterschiedlichen Messzeitpunkten in Messwiederholungsdesigns) in Beziehung gesetzt wird zu den Unterschieden innerhalb der Gruppe
- Einheit: in Standardabweichungen
- Die Effektstärke gibt Auskunft über praktische Relevanz

Signifikanzprüfung in der Lehr-und Lernforschung

- Interventionen auf Lernwirksamkeit prüfen
- 2-Gruppendesign: Innovative Lernumgebung wird mit konventionellem Unterricht verglichen, unter der innovativen Lernumgebung zeigen sich signifikant bessere Leistungen

- Welche Schlüsse kann man NICHT ziehen?
  - Innovative Lernumgebung ist das beste, was es gibt
  - Die in der innovativen Lernumgebung realisierten Elemente wirklich entscheidend sind
  - **Dazu benötigt man eine weitere Kontrollgruppe, in der Variationen innerhalb einer Lernumgebung vorgenommen werden**.

Warum man sich in Interventionsstudien nicht auf eine abhängige Variable beschränken sollte

- Wenn man einen spezifischen Effekt der Intervention nachweisen möchte, sollte man auch AV haben, die keine oder geringere Effekte zeigen

Wo findet man die Wahrheit in der Lehr- und Lernforschung?

- Journals mit Reviewverfahren, z.B.

- Journal of Educational Psychology
- Learning and Instruction
- Cognition and Instruction
- Journal of the Learning Sciences