FORSCHEN.
FÖRDERN.
VERNETZEN.
DEUTSCHE GESELLSCHAFT FÜR FREMDSPRACHENFORSCHUNG

UNIVERSITÄT PADERBORN

DGFF-Kolleg
Statistik (Grundlagen)

23. Juni 2023

Prof. Dr. Dominik Rumlich

dominik.rumlich@uni-paderborn.de

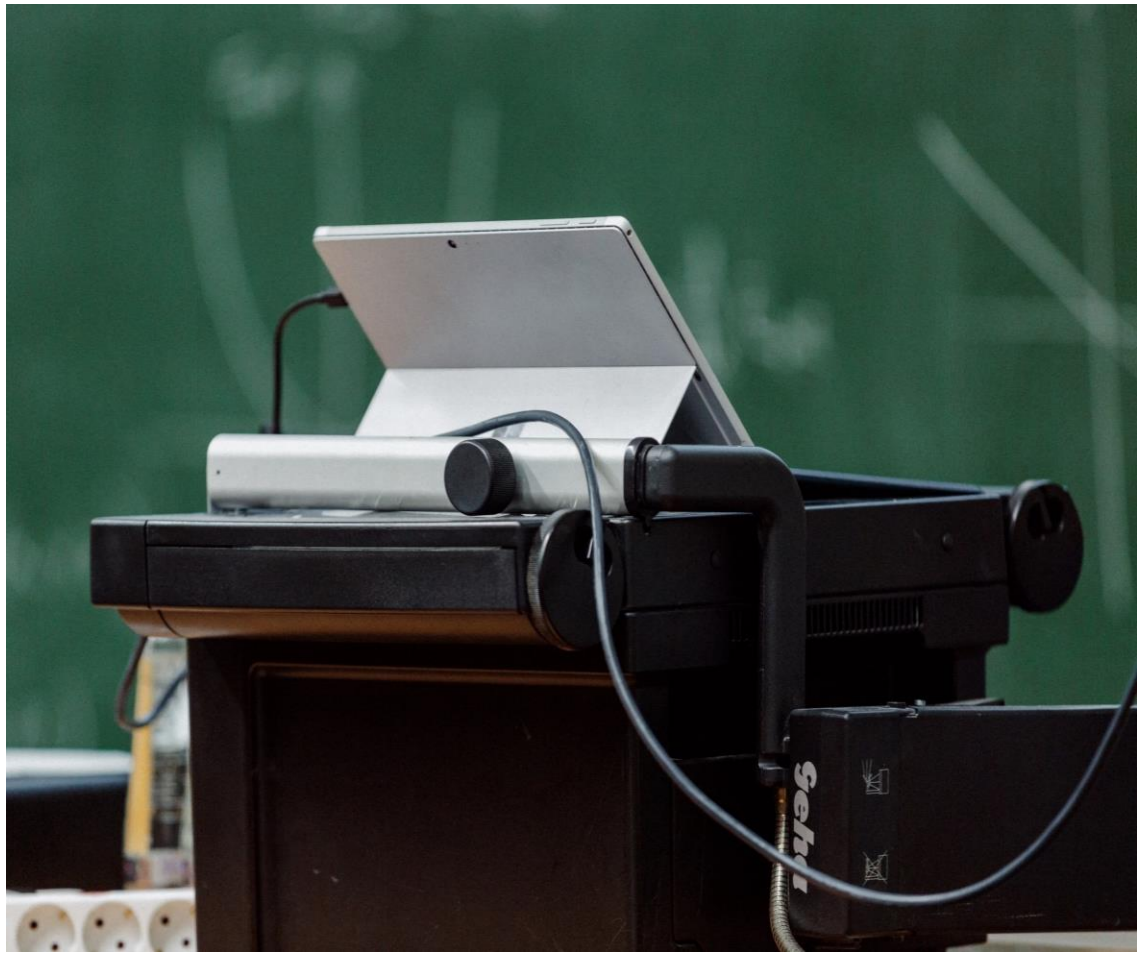


Foto: © Universität Bielefeld

Vorbereitende Lektüre

Theorie

- Field, Andy (2018): *Discovering statistics using IBM SPSS statistics* (5th edition; chapters 1-3). Thousand Oaks: Sage. [**sehr ausführlich**]
- Settinieri, Julia (2022): Deskriptiv- und Inferenzstatistik. In: Caspari, Daniela; Klippel, Friederike; Legutke, Michael & Schramm, Karen (Hrsg.): *Forschungsmethoden in der Fremdsprachendidaktik* (2. Aufl.). Tübingen: Narr Francke Attempto, 349–365. [**Wesentliches zusammengefasst**]

  [Ergänzende Empfehlung aufgrund ähnlicher Inhalte, aber anderer Beispiele/Perspektiven: Gültekin-Karakoç, Nazan & Feldmeier, Alexis (2014): Analyse quantitativer Daten. In: Settinieri, Julia; Demirkaya, Sevilen; Feldmeier, Alexis; Gültekin-Karakoç, Nazan & Riemer, Claudia (Hrsg.): *Einführung in empirische Forschungsmethoden für Deutsch als Fremd- und Zweitsprache.* Paderborn: UTB, 183–211.]

Praxisbeispiel

- Rumlich, Dominik (2018): Englischnoten und globale englische Sprachkompetenz in bilingualen Zweigen. *Zeitschrift für Erziehungswissenschaft* 21 [DOI 10.1007/s11618-017-0801-z].

Inhalte & Ziele

Statistische Kenntnisse und dazugehörige Systematik entwickeln, Bewusstsein schärfen, Hintergründe & Grenzen/Schwächen kennen

- Das „große Ganze“: Wissenschaftliche Ziele, Erkenntnisinteressen & Prinzipien
- Wie hilft Statistik?
- Umgang mit numerischen Daten: Ziele, Grundlagen & Prozeduren am Beispiel eines Zeitschriftenartikels (Rumlich 2018)
  - Forschungsfrage/Kontext der Studie
  - Einschub: Zustandekommen numerischer Daten (Forschungskreislauf)
  - Deskriptivstatistik
  - Inferenzstatistik
- Literaturangaben
- Anhang (Bedingungen für stat. Verfahren)

Wissenschaftliche Ziele & Erkenntnisinteressen

„Ausgehend von einer klar formulierten Fragestellung bieten statistische Verfahren fremdsprachen-didaktisch Forschenden systematische Prozeduren, mit denen in Hypothesenform ausgewiesene Forschungsfragen objektiv überprüft und zahlenmäßig erfasste Forschungsergebnisse beschrieben und interpretiert werden können. Statistische Verfahren sind somit ein unverzichtbares ‚kulturelles Werkzeug' (Vygotsky), um:

- Informationen über einen Untersuchungsgegenstand systematisch zu sammeln und darzustellen,
- aus Ergebnissen begründete Schlussfolgerungen zu ziehen sowie
- validierbare Verallgemeinerungen in Bezug auf das jeweilige Erkenntnisinteresse zu formulieren.

Etablierte statistische Verfahren (als mathematisch fundierte standardisierte Prozeduren) haben den großen Vorteil, in der Diskursgemeinschaft empirisch-quantitativ Forschender anerkannt und nachvollziehbar zu sein. Damit ist nicht nur eine höhere Qualität und größere Reichweite der aus den Forschungsergebnissen gezogenen Konsequenzen zu erzielen, sondern auch die Wiederholbarkeit der Analysen (unter analogen oder variierten Bedingungen) möglich, um so unser Wissen über Lehr-Lernprozesse von Fremdsprachen unter institutionell-unterrichtlichen Bedingungen kontinuierlich voranzubringen."

(Grum & Zydatiß 2022: 343)

Wissenschaftliche Ziele & Erkenntnisinteressen

„Ausgehend von einer klar formulierten Fragestellung bieten statistische Verfahren fremdsprachen-didaktisch Forschenden systematische Prozeduren, mit denen in Hypothesenform ausgewiesene Forschungsfragen objektiv überprüft und zahlenmäßig erfasste Forschungsergebnisse beschrieben und interpretiert werden können. Statistische Verfahren sind somit ein unverzichtbares ‚kulturelles Werkzeug' (Vygotsky), um:

- Informationen über einen Untersuchungsgegenstand systematisch zu sammeln und darzustellen,
- aus Ergebnissen begründete Schlussfolgerungen zu ziehen sowie
- validierbare Verallgemeinerungen in Bezug auf das jeweilige Erkenntnisinteresse zu formulieren.

Etablierte statistische Verfahren (als mathematisch fundierte standardisierte Prozeduren) haben den großen Vorteil, in der Diskursgemeinschaft empirisch-quantitativ Forschender anerkannt und nachvollziehbar zu sein. Damit ist nicht nur eine höhere Qualität und größere Reichweite der aus den Forschungsergebnissen gezogenen Konsequenzen zu erzielen, sondern auch die Wiederholbarkeit der Analysen (unter analogen oder variierten Bedingungen) möglich, um so unser Wissen über Lehr-Lernprozesse von Fremdsprachen unter institutionell-unterrichtlichen Bedingungen kontinuierlich voranzubringen."

(Grum & Zydatiß 2022: 343)

Wissenschaftliche Ziele & Prinzipien (u.a.)

- Validität (inhaltlich-methodische Passung/Methodenadäquatheit etc.), Reliabilität/Fehlerfreiheit bzw. -minimierung, Objektivität/Intersubjektivität
- Einbettung in Existierendes (Erkenntnisse, Methoden etc.) & Abgrenzung
- (angestrebte) Gültigkeit über Studienkontext hinaus/„Verallgemeinerbarkeit“
- Systematik, Nachvollziehbarkeit, Transparenz, Differenzierung, Präzision, Tiefe, Korrektheit, Eindeutigkeit
- Zwei Ebenen:
    1) Prozess der Erkenntnisgewinnung [„doing science“/Prozess] &
    2) Darstellung [„communicating science“/Produkt]

Wie hilft Statistik?

- i.d.R. große (unübersichtliche, kaum interpretierbare) Zahlenmenge
- Unterstützung wiss. Ziele/Prinzipien durch statistische Prozeduren
  - Übersichtlichkeit/Zugänglichkeit/Repräsentation
    (Zusammenfassungen/Umformungen/Berechnungen/Analysen: Zahlen, Tabellen, Grafiken…)
  - Tendenzen/Muster (hypothesenprüfend)
  - Modelle (Zusammenhänge, Unterschiede, Vorhersagen)
  - Interpretationshilfen (abs./rel. Größenabschätzungen, Vergleiche)

Umgang mit numerischen Daten: Ziele, Prozeduren & Grundlagen

(am Beispiel eines Zeitschriftenartikels)

Rumlich, Dominik (2018): Englischnoten und globale englische Sprachkompetenz in bilingualen Zweigen. *Zeitschrift für Erziehungswissenschaft* 21 [DOI 10.1007/s11618-017-0801-z].

FF, Kontext & Studiendesign

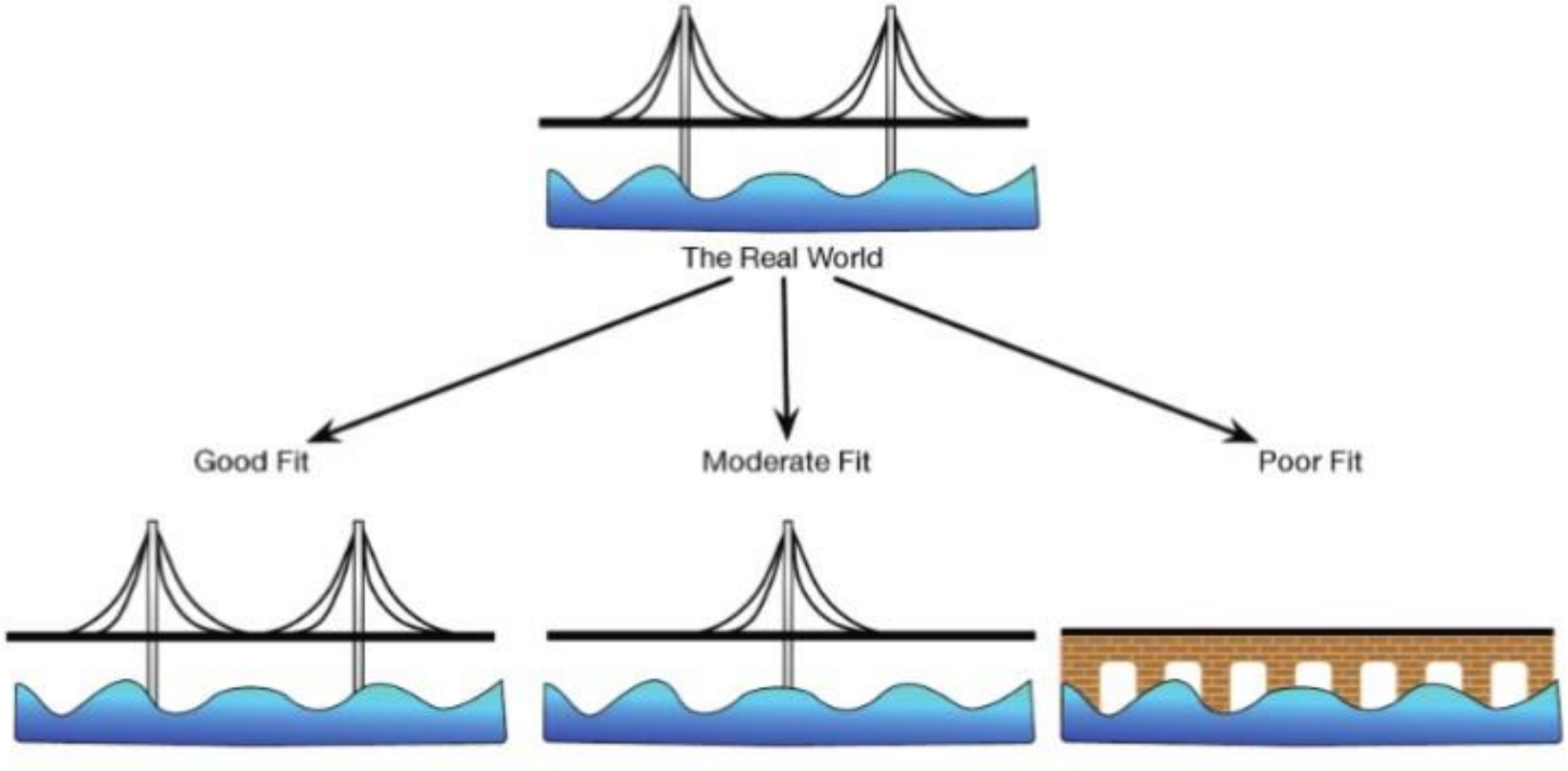
- Einfluss von Biling. SFU (BILI) auf globale engl. Sprachkomp. & Englischnoten
- Hypothesen:
    1) BILI verbessert globale englische Sprachkompetenz
    2) Leistungsstarkes Umfeld der BILI-Klassen beeinflusst Englischnoten
- Grundlegendes Studiendesign
  - (Quasi-)Experimentalgruppe: Klassen mit bilingualem Sachfachunterricht (BILI)
  - Vergleichsgruppe 1: Parallelklassen an gleichen Schulen ohne Bili (PARA)
    Vergleichsgruppe 2: Klassen von Regelschulen ohne Bili (REGEL)
  - Längsschnitt: Ende Klasse 6 (prä/vor BILI) & Ende Klasse 8 (post)
  - Instrumente: C-Test (Hamburger Schulleistungstest für 6. und 7. Klassen bzw. 8. und 9. Klassen; Amt für Schule, Hamburg 1998, 2000); SuS-Selbstauskunft zur Englischnote

Methodischer Teil: Stichprobe

- Grundlegendes Studiendesign + Wissen & Beurteilung: Auszuwertende Daten
  - Stichprobe (ggf. Ausschluss/Auswahl)

Gesamt	30 Kl.	*N = ~1000*	M_{Alter} = 11,87 Jahre	SD_{Alter} = 0,45	52,0% weiblich
BILI	16 Kl.	*n* = 428 [414]	M_{Alter} = 11,79 Jahre	SD_{Alter} = 0,49	62,2% weiblich
PARA	15 Kl.	*n* = 360	M_{Alter} = 11,91 Jahre	SD_{Alter} = 0,44	40,8% weiblich
REGEL	7 Kl.	*n* = 179	M_{Alter} = 11,98 Jahre	SD_{Alter} = 0,38	47,8% weiblich

Ziele, Prozeduren & Grundlagen: Methodischer Teil



Field 2013: 42

Ziele, Prozeduren & Grundlagen: Methodischer Teil

- Grundlegendes Studiendesign + Wissen & Beurteilung: Auszuwertende Daten
  - Instrumente: Wie sind die Konstrukte beobachtet/gemessen/erfasst worden?
    => Beurteilung ermöglichen, was die Daten ausdrücken (Validität/Rel./Obj.)
    - Wie sahen die Items aus? (Beispiele, Referenzieren bekannter Tests…)
    - Eigenschaften der Items: Anzahl und (statistische) Eigenschaften wie bspw. Lösungshäufigkeit/Schwierigkeit, Diskriminierung, Reliabilitäten...
      => z.T. von Art des Umgangs mit den Rohdaten abhängig => unter Datenauswertung
    - Theoretische & empirische Bestimmung der Gütekriterien in quant. Forschung

Methodischer Teil: Datenauswertung

- Grundlegendes Studiendesign + Wissen & Beurteilung: Auszuwertende Daten
  - Datenauswertung (Software, gewählte stat. Prozeduren)
    => Beurteilung ermöglichen, was die stat. Werte ausdrücken & ob das passt („numbers don‘t know where they came from“; Lord 1953: 751)
    - Rohwerte (bspw. Anzahl korrekt gelöster items): Längsschnittlich kaum interpretierbar
    - Eindimensionales Rasch-Modell in Conquest (Wu et al. 2007)
      (Stat. Modell mit paralleler Schätzung der Itemschwierigkeiten und Personenfähigkeiten)
    - weighted likelihood estimates (WLEs) = Schätzer der Personenfähigkeiten
    - „Kennwerte der Rasch-Items erwiesen sich als gut (0,8 < WMSQ < 1,25; EAP/PV Reliabilitäten > 0,90; Item-Separationsreliabilitäten > 0,99), wie es sich bei einem validierten Test für die entsprechende Schulstufe erwarten lässt.“ (Rumlich 2018: 38)

Methodischer Teil: Datenauswertung

- Grundlegendes Studiendesign + Wissen & Beurteilung: Auszuwertende Daten
  - Datenauswertung (Software, gewählte stat. Prozeduren)
    => Beurteilung ermöglichen, was die stat. Werte ausdrücken & ob das passt („numbers don‘t know where they came from“; Lord 1953: 751)
    - „Die Analysen der längsschnittlichen Leistungsentwicklung und der Referenzgruppeneffekte erfolgte auf Basis von Random-Intercept Mehrebenen-regressionen mit MPlus (Muthén und Muthén 2012), um der hierarchischen, geclusterten Struktur mit Daten auf Individual-, Klassen- und Gruppenebene gerecht zu werden. MLR und FIML ermöglichen dabei die Schätzung robuster Standardfehler bei gleichzeitiger Berücksichtigung fehlender Werte.“ (Rumlich 2018: 38)
    - Weitere Berechnungen: SPSS (Version 24; IBM 2016).

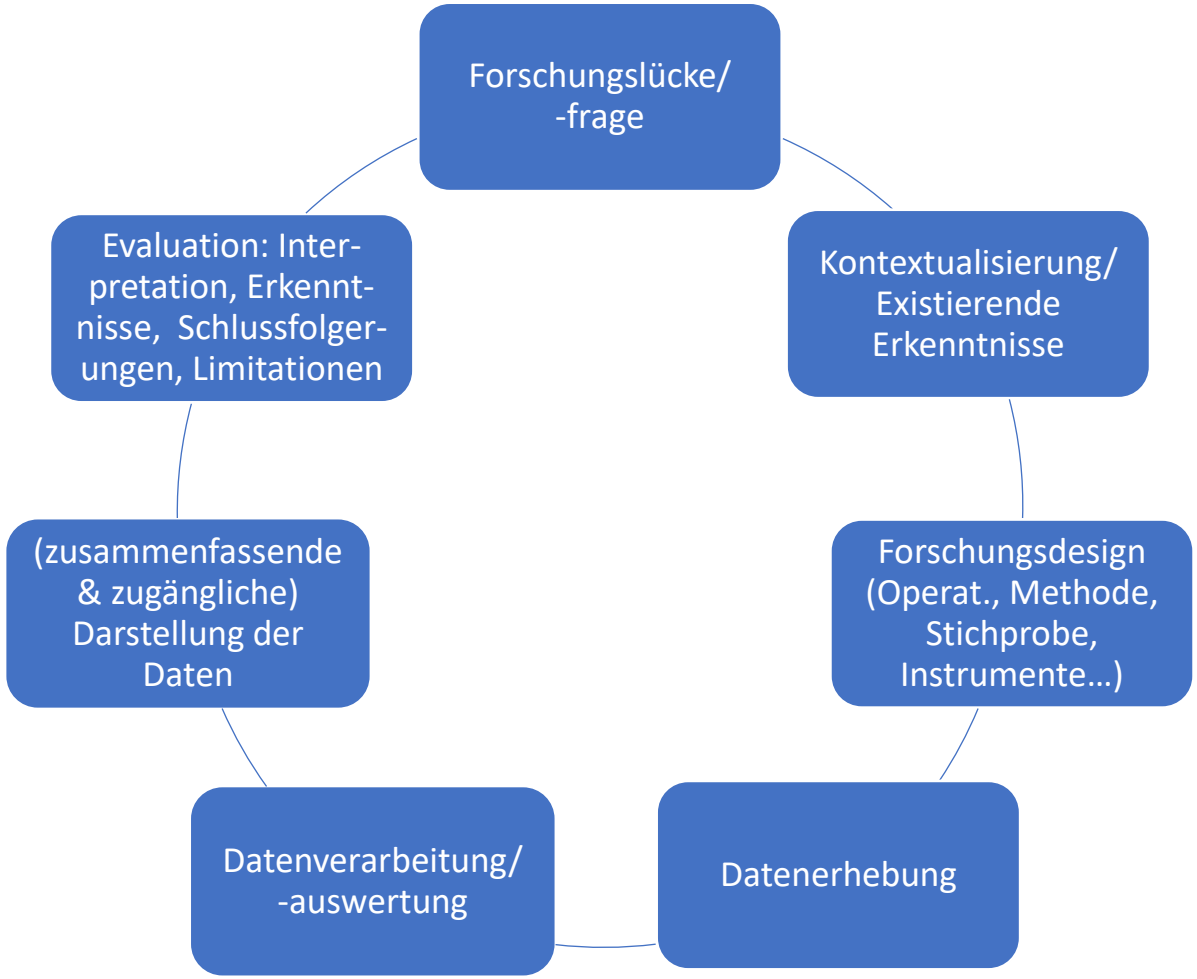
Methodischer Teil: Fehlende Werte & Demographie

- Grundlegendes Studiendesign + Wissen & Beurteilung: Auszuwertende Daten
  - Zusätzlich Relevantes, bspw. Umgang mit fehlenden Werten (Längsschnitt)
    - Prozentualer Anteil fehlender Werte aufgrund von Nicht-Teilnahme lag pro Subgruppe pro Messzeitpunkt zwischen 10 und 15%.
    - Drop-out Analysen (SuS i.d.R. nur 1x gefehlt): Missing at random (=> keine substantiellen Verzerrungen erwartbar)
    - „Da sich in den Gruppen z. T. signifikante Unterschiede bzgl. Alter, Geschlecht und Erstsprache ergaben, wurden die Einflüsse dieser Variablen statistisch kontrolliert. Da sich keine der Forschungsfragen auf diese Variablen bezieht, werden sie nachfolgend nicht explizit in die Auswertung eingebunden, um einen klaren Fokus des Ergebnis- und Diskussionsteils zu gewährleisten.“

Ziele, Prozeduren & Grundlagen: Methodischer Teil

- Grundlegendes Studiendesign + Wissen & Beurteilung: Auszuwertende Daten
  - Stichprobe & Untergruppen (ggf. Ausschluss/Auswahl)
  - Instrumente
  - Datenauswertung (Software, gewählte stat. Prozeduren)
  - Zusätzlich Relevantes, bspw. Umgang mit fehlenden Werten (Längsschnitt), Demographie

Zustandekommen numerischer Daten



(adaptiert von Seliger & Shohamy 1989)

Zustandekommen numerischer Daten

(adaptiert von Seliger & Shohamy 1989)

Grundlegende Deskriptivstatistik

- Ziel: Grundlegender erster Überblick & „Gefühl“ für Daten (allgemein & bzgl. FF)
  1. WLE-Schätzer (weighted likelihood estimate) für globale Englischkompetenz

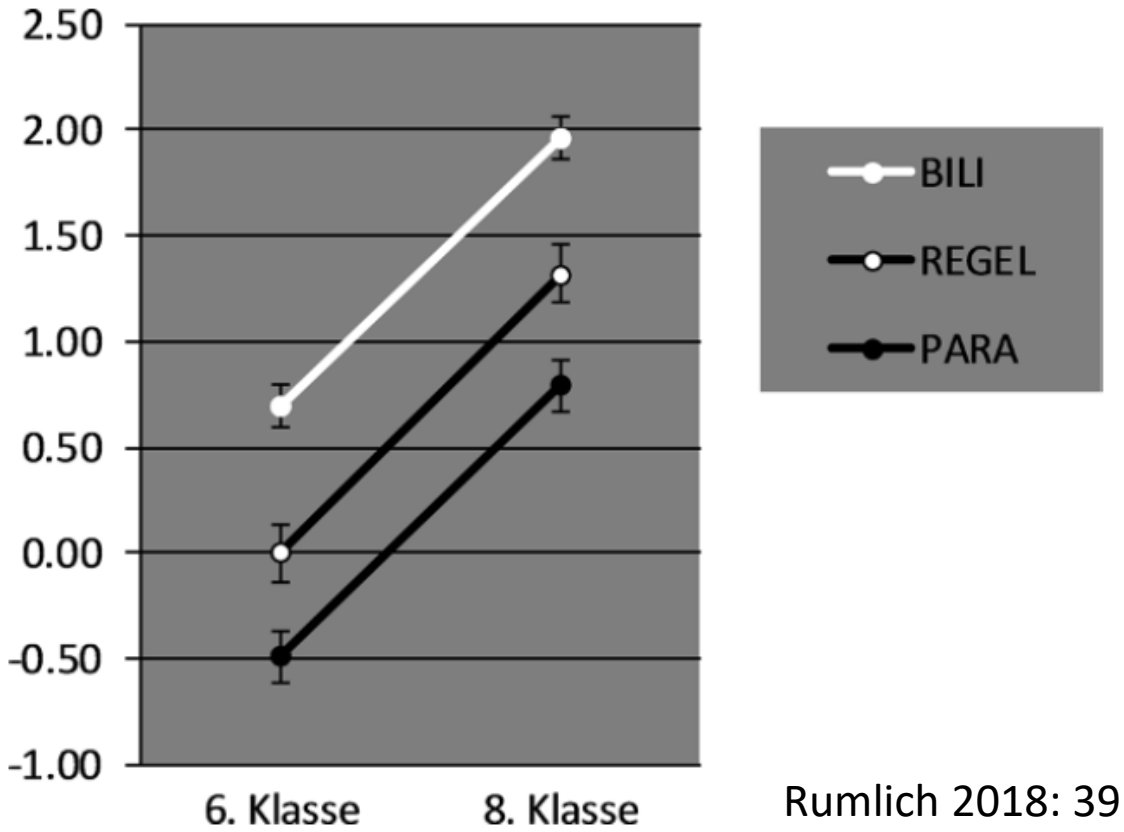
				Perzentile				
	N	*M*	*SD*	12,5%	25%	50%*	75%	87,5%
BILI	374 [414]	0,69	1,00	-0,44	0,04	0,70	1,33	1,73
PARA	318 [360]	-0,49	1,00	-1,70	-1,01	-0,55	0,16	0,70
REGEL	155 [179]	0,00	0,85	-1,01	-0,44	0,04	0,41	0,84

*Median

Tiefergehende Deskriptivstatistik

- Ziel: Grundlegender erster Überblick & „Gefühl“ für Daten (allgemein & bzgl. FF)

Abb. 1 Durchschnittliche Englischleistung (WLE-Scores) mit Konfidenzintervallen am Ende der 6. und 8. Klasse nach Gruppen



Rumlich 2018: 39

Aus dem Text: 0,85 ≤ *SD* ≤ 1,00

BILI vs. REGEL: Cohen‘s *d** = 0,84
BILI vs. PARA: Cohen‘s *d** = 1,18
PARA vs. REGEL: Cohen‘s *d** = 0,55

*Cohen‘s *d* (Effektstärke): $(M_1 - M_2)/SD$
(Mittelwertsunterschied ausgedrückt in *SD*)

Grundlegende Deskriptivstatistik

- Ziel: Grundlegender erster Überblick & „Gefühl“ für Daten (allgemein & bzgl. FF)
  2. Englischnoten

Tab. 2 Prozentuale Verteilung der Noten am Ende der 6. und 8. Klasse nach Gruppen

Gruppe	6. Klasse (Noten)					8. Klasse (Noten)				
	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
BILI	18,6	47,5	26,7	6,9	0,3	12,9	39,6	35,4	11,8	0,3
REGEL	7,3	31,1	37,7	21,9	2,0	5,8	24,4	35,9	30,8	3,2
PARA	10,3	45,6	35,7	8,5	0,0	5,5	32,9	37,4	20,8	3,5

„In der 6. Klasse berichtet die BILI-Gruppe im Durchschnitt die besten Englisch-Halbjahreszeugnisnoten (*M* = 2,22; *SD* = 0,84), gefolgt von der PARA- (*M* = 2,43; *SD* = 0,79) und REGEL-Gruppe (*M* = 2,80; *SD* = 0,93).“
(Rumlich 2018: 41)

Einschub: Eigenschaften numerischer Daten

- Skalenniveaus
  - Nominal/Kategorial (bspw. Erstsprache, Geschlecht)
    Zahlenwert steht für Name/Bezeichnung, d.h. ohne math. Bedeutung
    => math. Operationen stark eingeschränkt (absolute/relative Häufigkeiten, Modalwert [häufigster Wert einer Verteilung], gleich/ungleich)
  - Ordinal/Reihenfolgen- bzw. Rangskala (bspw. Noten, Einlaufreihenfolge Sprint)
    Zahlenwert erlaubt größer/kleiner-Relation, keine Gleichabständigkeit
    => eingeschränkte math. Operationen (absolute/relative Häufigkeiten, *Me*, Range, Min/Max, Perzentile/Quartile…)
  - Metrisch (bspw. Punkte in Test, Alter; i.d.R. auch Fragebogendaten)
    „vollwertiger Zahlenwert“ (Gleichabständigkeit: 16 Jahre = doppelt so viel wie 8 J.); d.h. alle math. Operationen sinnvoll möglich (bspw. *M, SD*…)

  => Skalenniveau + weitere Voraussetzungen (Normalverteilung, homogene Varianzen…): parametrische (*t*-Test, ANOVA, Korrelationskoeffizient *r* nach Pearson, lineare Regression…) oder nicht-parametrische Verfahren (*U*-Test, Kruskal-Wallis-Test, Chi²-Test, logistische Regression...)

Einschub: Eigenschaften numerischer Daten

Was ist „mathematisch geboten“ (1. Blick – 2. Blick) vs. „gängige Praxis“?

- FB-Daten oft als metrisch angesehen/behandelt (Bortz & Schuster 2010: 23; Rasch et al. 2006: 14)
- Schulnoten nicht math. verzerrt wie Drittelnoten (math. Gleichabständigkeit)
- Hauptargument gg. parametrische Verfahren: Verzerrung/falsche Ergebnisse, aber…
  - beide Verfahren oft gleiche Ergebnisse
  - „falsche Verfahren“ erschweren Entdeckung von Effekten (Bortz & Schuster 2010: 23)
  - math. Robustheit ggüber Verletzung von Voraussetzungen (Kubinger et al. 2009)
  - Fehlereintrag bei Erhebung i.d.R. größer als pot. stat. Verzerrung durch falschen Test
- M/SD liefern wertvolle zusammenfassende Informationen (Muster/Tendenzen)

=> bewusste, abgewogene & informierte (und gut kommunizierte) Entscheidungen

Grundlegende Deskriptivstatistik

- Ziel: Grundlegender erster Überblick & „Gefühl“ für Daten (allgemein & bzgl. FF)
  2. Englischnoten

Tab. 2 Prozentuale Verteilung der Noten am Ende der 6. und 8. Klasse nach Gruppen

Gruppe	6. Klasse (Noten)					8. Klasse (Noten)				
	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
BILI	18,6	47,5	26,7	6,9	0,3	12,9	39,6	35,4	11,8	0,3
REGEL	7,3	31,1	37,7	21,9	2,0	5,8	24,4	35,9	30,8	3,2
PARA	10,3	45,6	35,7	8,5	0,0	5,5	32,9	37,4	20,8	3,5

„In der 6. Klasse berichtet die BILI-Gruppe im Durchschnitt die besten Englisch-Halbjahreszeugnisnoten (*M* = 2,22; *SD* = 0,84), gefolgt von der PARA- (*M* = 2,43; *SD* = 0,79) und REGEL-Gruppe (*M* = 2,80; *SD* = 0,93).“
(Rumlich 2018: 41)

Überblick: Deskriptivstatistik

- Ziele:
  1) Grundlegender erster Überblick, Zusammenfassung & „Gefühl“ für Daten (allg. & FF)
  2) Hinweise (oder ggf. erste Antworten) bzgl. FF/Hypothesen

- Typische Verfahren
  - Numerische Kennwerte: *N/n,* abs. & rel. Häufigkeiten, Min/Max, Spannweite, *Md, Me, M, SD, SE,* Perzentile, Effektstärkemaße (Cohen‘s *d*)
  - Grafische Verfahren: Tabellen mit numerischen Kennwerten, alle Arten von Diagrammen (Histogramme, Boxplots, Balken-/Säulen-/Linien-/Punktdiagramme etc.)

Grundlagen der Inferenzstatistik

- Ziel: (komplexere) Muster/Tendenzen finden & passende Modelle entwickeln; dafür Vorhandensein & Stärke von Zusammenhängen/Unterschieden statistisch abschätzen
- Zu Grunde liegende Idee (vereinfachtes Beispiel)
  - Münzwurf (*n*=6): blau – blau – blau – gelb – blau – blau
    Münzwurf (Theorie): 3xblau – 3x gelb
  - Problem von wenigen absoluten Beobachtungen (kleine Stichprobe)
    => wenig stat. Power, (vorhandene) Effekte zu entdecken
    => einfacher, wenn Effekte größer werden (15x blau, 1x gelb)
  - 1 zu 5 vs. 1.000 zu 5.000 (absolut unterschiedlich, relativ gleich)
    [Gesetz d. großen Zahlen: beobachtete Häufigkeit => theoret. Wahrscheinlichkeit]

https://www.supergurumi.de/kreisrunden-spiral-untersetzer-haekeln

- Zu Grunde liegende Idee (vereinfachte Theorie)
  - Vergleich zweier Verteilungen: Wie groß ist Wahrscheinlichkeit, dass sie aus der gleichen Grundgesamtheit stammen (= beobachtete Unterschiede sind zufällig/“nicht signifikant“)?
  - Antwort = *p*-Wert/Irrtumswahrscheinlichkeit: $p < 0{,}05$ => Unterschiede „signifikant“ (Konvention)
  - *p*-Wert durch Größe des Effekts & Stichprobe beeinflusst (s.o.)

Inferenzstatistik: Nullhypothesen-Signifikanztestung (NHST)

- Ziel: (komplexere) Muster/Tendenzen finden & passende Modelle entwickeln; dafür Vorhandensein & Stärke von Zusammenhängen/Unterschieden statistisch abschätzen

Tab. 2 Prozentuale Verteilung der Noten am Ende der 6. und 8. Klasse nach Gruppen

Gruppe	6. Klasse (Noten)					8. Klasse (Noten)				
	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
BILI	18,6	47,5	26,7	6,9	0,3	12,9	39,6	35,4	11,8	0,3
REGEL	7,3	31,1	37,7	21,9	2,0	5,8	24,4	35,9	30,8	3,2
PARA	10,3	45,6	35,7	8,5	0,0	5,5	32,9	37,4	20,8	3,5

„In der 6. Klasse berichtet die BILI-Gruppe im Durchschnitt die besten Englisch-Halbjahreszeugnisnoten (M = 2,22; SD = 0,84), gefolgt von der PARA- (M = 2,43; SD = 0,79) und REGEL-Gruppe (M = 2,80; SD = 0,93).

Signifikante Unterschiede ergeben sich zwischen
BILI- und REGEL- (Wald $\chi^2(1) = 12{,}78$; $p < 0{,}001$; $d = 0{,}65$) sowie zwischen
PARA- und REGEL-Gruppe (Wald $\chi^2(1) = 5{,}55$; $p < 0{,}05$; $d = 0{,}48$), nicht jedoch zwischen
BILI- und PARA-Gruppe (Wald $\chi^2(1) = 2{,}89$; $p = 0{,}09$; $d = 0{,}23$).“

(Rumlich 2018: 40)

Nullhypothesen-Signifikanztestung: Hintergründe

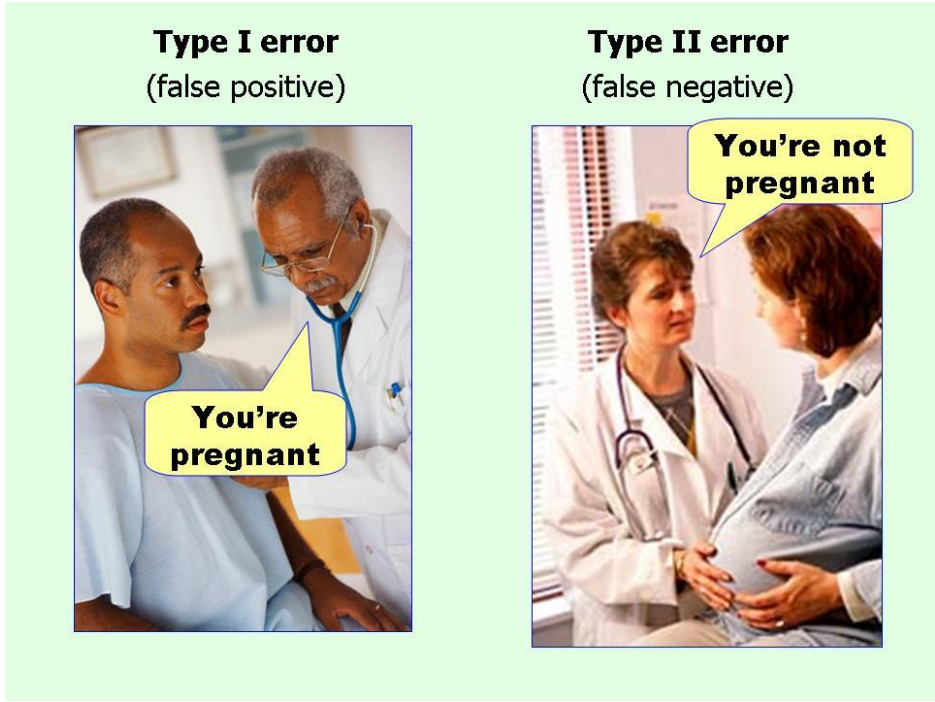
- Ziel: (komplexere) Muster/Tendenzen finden & passende Modelle entwickeln; dafür Vorhandensein & Stärke von Zusammenhängen/Unterschieden statistisch abschätzen
- Probleme…

Table 9.2

Significance Versus Covariation

N	Significant at .05	Variance in Common (%)
5	.80	64
10	.55	30
20	.38	14
100	.17	3
250	.10	1
1,000	.05	0.25

http://www.hawaii.edu/powerkills/UC.HTM#*

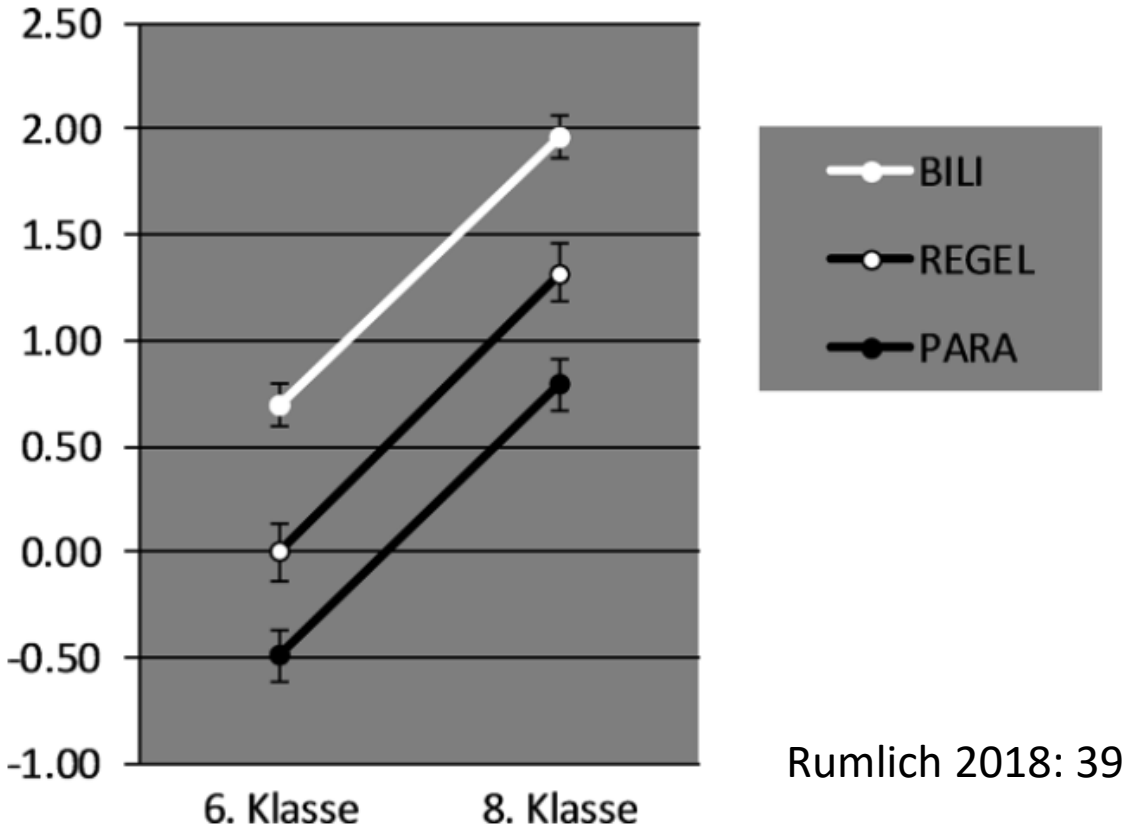


https://www.statisticsblog.com/wp-content/uploads/2014/05/Type-I-and-II-errors1-625x468.jpg

Inferenzstatistik: Konfidenzintervalle

- Ziel: (komplexere) Muster/Tendenzen finden & passende Modelle entwickeln; dafür Vorhandensein & Stärke von Zusammenhängen/Unterschieden statistisch abschätzen

Abb. 1 Durchschnittliche Englischleistung (WLE-Scores) mit Konfidenzintervallen am Ende der 6. und 8. Klasse nach Gruppen



Rumlich 2018: 39

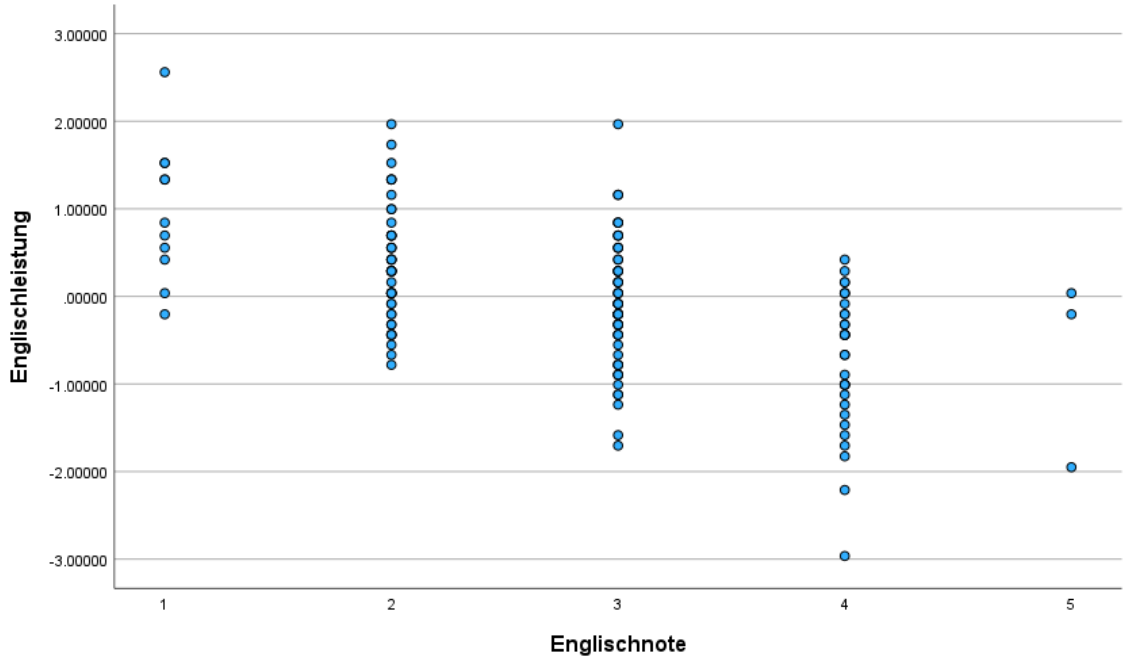
Aus dem Text: $0{,}85 \leq SD \leq 1{,}00$

BILI vs. REGEL: Cohen‘s d^* = 0,84
BILI vs. PARA: Cohen‘s d^* = 1,18
PARA vs. REGEL: Cohen‘s d^* = 0,55

*Cohen‘s d (Effektstärke): $(M_1 - M_2)/SD$
(Mittelwertsunterschied ausgedrückt in SD)

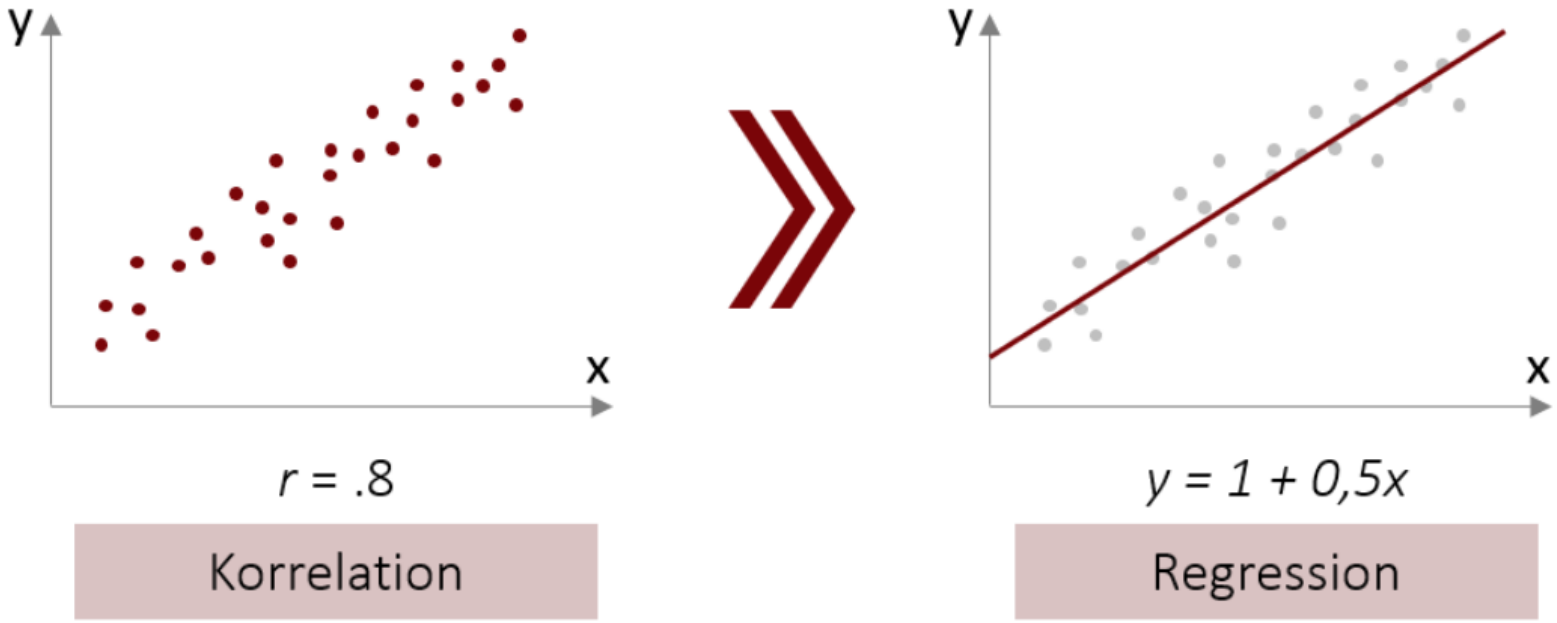
Inferenzstatistik: Zusammenhänge (Korrelation)

- Ziel: (komplexere) Muster/Tendenzen finden & passende Modelle entwickeln; dafür Vorhandensein & Stärke von Zusammenhängen/Unterschieden statistisch abschätzen



Inferenzstatistik: Zusammenhänge (Korrelation)

- Ziel: (komplexere) Muster/Tendenzen finden & passende Modelle entwickeln; dafür Vorhandensein & Stärke von Zusammenhängen/Unterschieden statistisch abschätzen

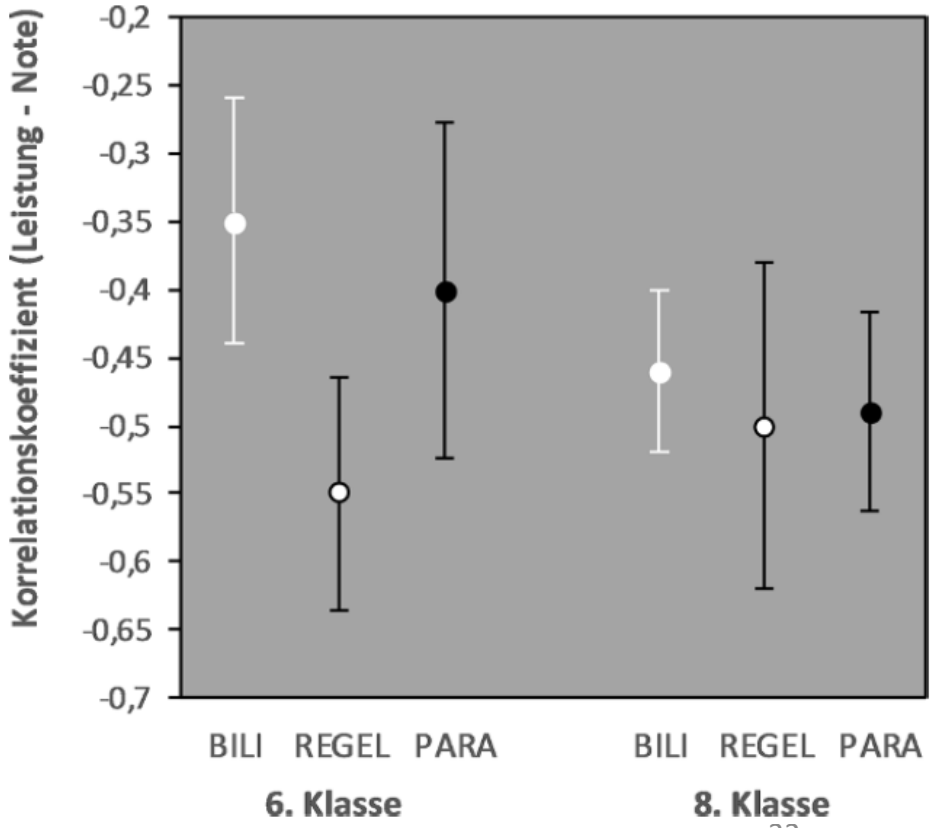


Planing (o.d.), https://statistikgrundlagen.de/ebook/chapter/regression/

Inferenzstatistik: Zusammenhänge (Korrelation)

- Ziel: (komplexere) Muster/Tendenzen finden & passende Modelle entwickeln; dafür Vorhandensein & Stärke von Zusammenhängen/Unterschieden statistisch abschätzen

Abb. 2 Korrelationen zwischen Noten und Testleistung mit Konfidenzintervallen am Ende der 6. und 8. Klasse nach Gruppen



Tab. 1 Korrelationen zwischen Noten und Testleistung am Ende der 6. und 8. Klasse nach Gruppen

	BILI	REGEL	PARA	Gesamt
6. Klasse	–0,35***	–0,55***	–0,40***	–0,44***
8. Klasse	–0,46***	–0,50***	–0,49***	–0,50***

*** $p < 0{,}001$

Rumlich 2018: 40

Inferenzstatistik: Zusammenhänge (Regression)

- Ziel: (komplexere) Muster/Tendenzen finden & passende Modelle entwickeln; dafür Vorhandensein & Stärke von Zusammenhängen/Unterschieden statistisch abschätzen
- Regression: (Lineares) Modell (= Gerade) mit Faktoren/Prädiktoren entwickeln, das unabhängige/outcome Variable vorhersagt
- Stärke des Einflusses: Regressionsgewicht *b* (unstandardisiert)/β (standardisiert)
- R^2 = „Prozentuale Varianzaufklärung der Outcome-Variable durch Prädiktor“

Tab. 3 Standardisierte Regressionsgewichte für Prädiktoren der Englischnote in Klasse 6 und 8 nach Gruppen

		Englischnote 6. Klasse		
Gruppe	Prädiktoren	β	*SE*	R^2
BILI	Individuelle Leistung	–0,48***	0,04	0,23
	Mittlere Klassenleistung	0,67***	0,11	0,45

Rumlich 2018: 43

Überblick: Inferenzstatistik

- Ziel: möglichst klare, valide, reliable, objektive (und „generalisierbare“ bzw. verallgemeinernde, d.h. über die Stichprobe hinausgehende) Schlüsse/Hinweise bzgl. Muster & Tendenzen im Hinblick auf FF/Hypothesen

- Typische Verfahren
  - Unterschiede: Konfidenzintervalle, Nullhypothesen-Signifikanztests (*t*-/*U*-/*Chi*²-Test, Kruskal-Wallis-Test, ANOVA…)
  - Zusammenhänge: Korrelationskoeffizienten, Regression
  - Weitere Prozeduren: Faktorenanalyse

Vielen Dank
für Ihre Aufmerksamkeit!

Literaturangaben

- Amt für Schule, Hamburg. Behörde für Schule, Jugend und Berufsbildung. (Hrsg.) (1998): *Der Hamburger Schulleistungstest für sechste und siebte Klassen – SL-HAM 6/7.* Hamburg.
- Behörde für Schule, Jugend und Berufsbildung. Amt für Schule, Hamburg (Hrsg.) (2000): *Der Hamburger Schulleistungstest für achte und neunte Klassen – SL-HAM 8/9.* Hamburg.
- Bortz, Jürgen & Schuster, Christof (2010): *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Heidelberg: Springer.
- Bühner, Markus & Ziegler, Matthias (2009): *Statistik für Psychologen und Sozialwissenschaftler.* München: Pearson.
- Field, Andy (2013): *Discovering statistics using IBM SPSS statistics* (4th edition). Thousand Oaks: Sage.
- Field, Andy (2018): *Discovering statistics using IBM SPSS statistics* (5th edition). Thousand Oaks: Sage.
- Grum, Urška & Zydatiß, Wolfgang (2022): Statistische Verfahren – Einleitung. In: Caspari, Daniela; Klippel, Friederike; Legutke, Michael & Schramm, Karen (Hrsg.): *Forschungsmethoden in der Fremdsprachendidaktik* (2. Aufl.). Tübingen: Narr Francke Attempto, 349–365.
- Kubinger, Klaus D.; Rasch, Dieter & Moder, Karl (2009): Zur Legende der Voraussetzungen des t-Tests für unabhängige Stichproben. *Psychologische Rundschau* 60 [DOI 10.1026/0033-3042.60.1.26].
- Lord, F. M. (1953): On the statistical treatment of football numbers. *American Psychologist* 8, 750f.
- Muthén, Linda K. & Muthén, Bengt O. (1998-2012): *Mplus user's guide* (7. Aufl.). Los Angeles, CA: Muthén & Muthén.
- Porte, Graeme K. (2010): *Appraising research in second language learning* (2. Aufl.). Amsterdam: Benjamins.
- Rasch, Björn; Friese, Malte; Hofmann, Wilhelm & Naumann, Ewald (2006): *Quantitative Methoden* (2. Aufl., Band 1). Heidelberg: Springer.
- Rumlich, Dominik (2016): *Evaluating bilingual education in Germany: CLIL students' general English proficiency, EFL self-concept and interest.* Frankfurt am Main, Germany: Lang.
- Rumlich, Dominik (2018): Englischnoten und globale englische Sprachkompetenz in bilingualen Zweigen. *Zeitschrift für Erziehungswissenschaft* 21 [DOI 10.1007/s11618-017-0801-z].
- Seliger, Herbert W. & Shohamy, Elana (1989): *Second language research methods.* Oxford: Oxford University Press.
- Wu, M. L.; Adams, R. J. & Wilson, M.R (2007): *ConQuest Manual.* Camberwell, Australia: Acer Press.

Vorbereitende Lektüre

Theorie

- Field, Andy (2018): *Discovering statistics using IBM SPSS statistics* (5th edition; chapters 1-3). Thousand Oaks: Sage. [sehr ausführlich]
- Settinieri, Julia (2022): Deskriptiv- und Inferenzstatistik. In: Caspari, Daniela; Klippel, Friederike; Legutke, Michael & Schramm, Karen (Hrsg.): *Forschungsmethoden in der Fremdsprachendidaktik* (2. Aufl.). Tübingen: Narr Francke Attempto, 349–365. [Wesentliches zusammengefasst]

  [Ergänzende Empfehlung aufgrund ähnlicher Inhalte, aber anderer Beispiele/Perspektiven:
  Gültekin-Karakoç, Nazan & Feldmeier, Alexis (2014): Analyse quantitativer Daten. In: Settinieri, Julia; Demirkaya, Sevilen; Feldmeier, Alexis; Gültekin-Karakoç, Nazan & Riemer, Claudia (Hrsg.): *Einführung in empirische Forschungsmethoden für Deutsch als Fremd- und Zweitsprache.* Paderborn: UTB, 183–211.]

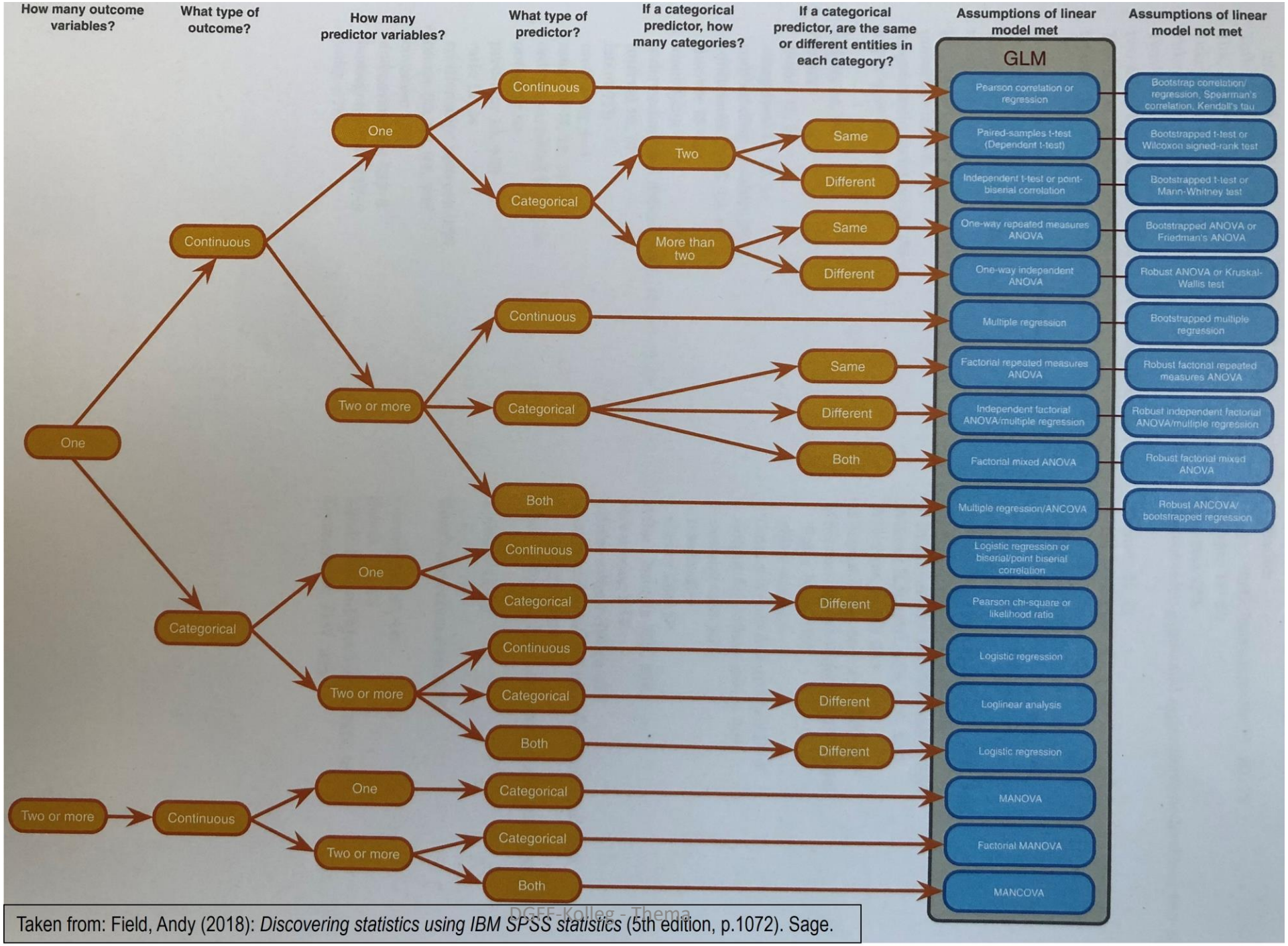
Praxisbeispiel

- Rumlich, Dominik (2018): Englischnoten und globale englische Sprachkompetenz in bilingualen Zweigen. *Zeitschrift für Erziehungswissenschaft* 21 [DOI 10.1007/s11618-017-0801-z].

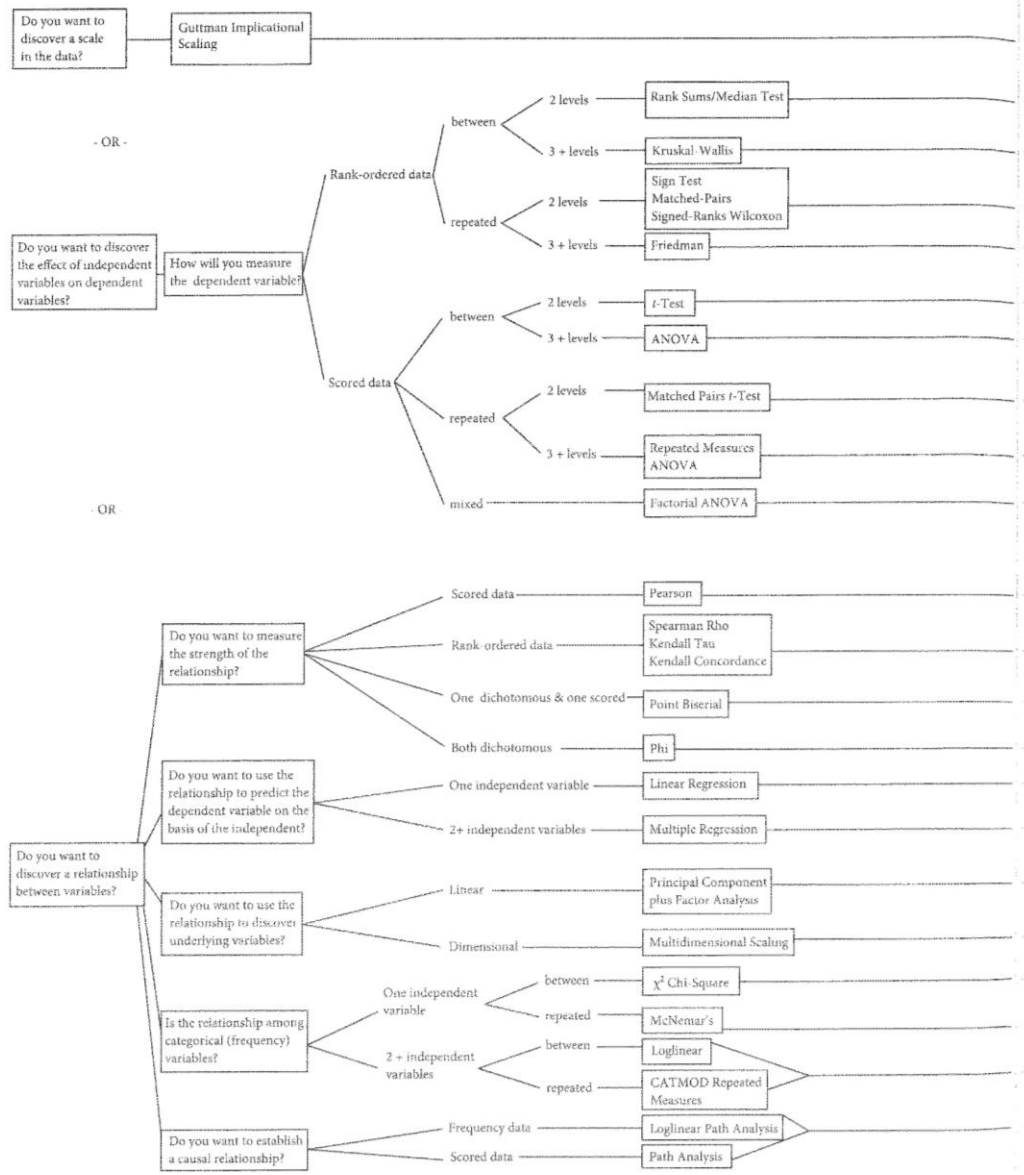
Anhang

Voraussetzungen für statistische Verfahren

1) Field, Andy (2018): *Discovering statistics using IBM SPSS statistics* (5th edition, p.1072). Thousand Oaks: Sage.
2) Porte, Graeme K. (2010): *Appraising research in second language learning* (2nd ed., p.292-295). Amsterdam: Benjamins.



Taken from: Field, Andy (2018): *Discovering statistics using IBM SPSS statistics* (5th edition, p.1072). Sage.



How will you interpret the results?	
Check that the coefficient of scalability is over .60	Interpret scale in light of reasonableness of the cutoff point, number of instances required, missing values, and context used to elicit forms.
Compare the groups Check *z* value	If the *z* score is significant, then one group has more *Ss* in higher ranks than the other. Use eta^2 for strength of association.
Compare the groups Check *H* value	If *H* is significant, the groups differ; use Ryan's procedure to locate which groups differ. Use eta^2 for strength of association.
Compare the groups Check *R* or *z* value	If the Sign test *R* or Wilcoxon *z* is significant, there is a change from time 1 to time 2. Use eta^2 for strength of association forWilcoxon.
Compare the groups Check χ^2	If χ^2 is significant, there is a change over several time points (or msrs.) To locate the difference more precisely report the results of the Nemenyi's procedure. Use eta^2 for strength of association.
Compare the two means Check *t* value	If the *t* value is significant, the two groups differ. Use eta^2 for strength of association.
Compare 3+ means Check *F* ratio	If the *F* ratio is significant, the groups differ. To locate the differences more precisely, interpret the multiple-range test (Scheffé, Tukey, or Newman-Keuls). Use $omega^2$ or eta^2 for strength of association.
Compare the two means Check *t* value	If the *t* value is significant, there is a difference in the means for the two times or measures. Use eta^2 for strength of association.
Compare 3+ means Check *F* ratio	If the *F* ratio is significant, the same (or matched) *Ss* perform differently on repeated measures. Use a multiple-range test to locate precise differences. Use eta^2 for strength ofassociation.
Compare the means Check *F* ratios	Step 1. If the interaction is significant, chart the means to show the interaction and interpret it. Interpret main effects in light of the interaction. Step 2. If the interaction is not significant, interpret the difference in the main effects. Use a multiple-range test to locate precise differences. Use eta^2 toshow the strength ofassociation.
Check the strength of each correlation	r^2 shows the amount of overlap between each pair of variables. Be sure to correct for attenuation if measures are not of equel reliability.
Check the probability of the correlation	If the correlation is significant, it shows that the *H* of no relation can be rejected. Interpret the value "sensibly" in terms of strength of relationship.
Check the value ofthe correlation	Explain the correlation in a "sensible" way.
Check χ^2 for significance	Explain the correlation in a "sensible" way.
Report predicted scores Check the SEE	The stronger the correlation and the smaller the SEE, the better the prediction will be.
Check each added variable	Identify the first independent variable, then the overlap ofthe second with the first to see how much each contributes (as well as their joint contribution) to explain variance in the dependent variable. Explain how much additional information is given by each succeeding independent variable.
Check each factor loading	If possible, once the number off actors has been determined, label each factor by consulting variables with high loadings vs. variables with low loadings on each. Else, label them as factorA, B, C, etc.
Check solutions and stress	Once a solution (about number of dimensions) has been identified or selected, label each dimension by consulting items in the cluster and those distant from cluster. Else, label them as dimension A, B, C, etc.
Check χ^2 value & $(O-E)^2/E$	If χ^2 is significant, the distribution differs from the expected distribution. Show which cells differ most from expected cell frequency or do a Ryan's procedure to locate the difference more precisely. Use Phi or Cramers V for strength of association.
Check *z* value values	If *z* is significant, conclude there is a change in proportion of *Ss* from time 1 to time 2.
Check parameter estimates to reduce model, compare models	The parameter estimates show which interactions and main effects are significant. To pare the model, compare various models with the saturated model. Decisions should be based on statistical and substantive arguments.
Check the paths to see which can be trimmed from the model	Use the analysis to trim paths from the model. Interpret the findings on both statistical and substantive grounds.

Taken from: Porte, G. K. (2010). Appraising research in second language learning (2nd ed.). Amsterdam, NL: Benjamins.

Appendix II Table of assumptions for popular statistical tests

Adapted from Brown, J.D. (1992), Statistics as a foreign language: part 2, *Tesol Quarterly, 26(4)*, 629-664.

Statistical procedure/ Assumptions	Independence of groups	Independence of observations	Normality	Equal variances	Linearity	Non-multicolinearity	Homoscedasticity	Other assumptions
Correlation								
Pearson r		•	•		•		•	
Spearman rho		•						
Kendall tau		•						
Kendall W		•						
Point-biserial correlation		•			•			
Phi coefficient		•	•		•			
Correlation/prediction								
Simple regression		•	•		•		•	
Multiple regression		•	•		•	•	•	
Loglinear analysis		•						No more than 20% of expected frequencies less than or equal to 5
Group differences								
z statistic (large samples)	•	•	•	•				
t test (any samples)	•	•	•	•				
One-way ANOVA	•	•	•	•				
One-way ANCOVA	•	•	•	•	•	•		
Matched pairs t-test		•	•	•				
Repeated measures ANOVA			•	•				
Repeated measures ANCOVA			•	•	•	•		
n-way ANOVA	•	•	•	•				
n-way ANCOVA	•	•	•	•	•	•		
n-way repeated measures ANOVA			•	•				
n-way repeated measures ANCOVA			•	•	•	•		

Statistical procedure/ Assumptions	Independence of groups	Independence of observations	Normality	Equal variances	Linearity	Non-multicolinearity	Homoscedasticity	Other assumptions
Multivariate ANOVA	•	•	•	•	•			
Multivariate ANCOVA	•	•	•	•	•	•		
Multivariate n-way ANOVA	•	•	•	•	•			
Multivariate n-way ANCOVA	•	•	•	•	•	•		
Median test	•	•						
Mann U/Wilcoxon	•	•						
Kruskal-Wallis	•	•						
Sign text	•	•						
Friedman One-way ANOVA	•	•						
Frequencies								
Chi-square								Expected frequencies greater or equal to 5 if the df is greater or equal to 2; greater than or equal to 10 if the df equals 1.
	•	•						
McNemar test								Differences all in same direction (same sign)
Fisher's exact test	•	•						
n-way chi-square	•	•						
Exploratory statistics								
Principal component analysis			•		•		•	Factorability of R
Factor analysis			•		•	•	•	Factorability of R
Multidimensional scaling			•		•	•		
Cluster analysis			•		•	•		
One-way discriminant analysis			•		•	•		Homogeneity of variance-covariance matrices
n-way discriminant analysis			•		•	•		Homogeneity of variance-covariance matrices
Guttman scaling								Scalable and reproducible
Path analysis			•		•	•	•	All relevant variables included; variables are causal

Taken from: Porte, G. K. (2010). Appraising research in second language learning (2nd ed.). Amsterdam, NL: Benjamins.