Kapitel 8

Stichproben- und datenadäquate Auswertungsansätze für die Unterrichtsforschung

Andreas Zendler

8.1 Einleitung

Für die experimentelle Unterrichtsforschung werden üblicherweise relativ große homogene Stichproben (*N* = large) vorausgesetzt, insbesondere dann, wenn die *Testung* von Hypothesen zur Klärung von Ursache-Wirkungszusammenhängen im Vordergrund des Interesses steht, hohe Repräsentativität verlangt und ein hoher Grad an externer Validität beansprucht werden (Gerring, 2007; George & Bennett, 2005; Blatter & Haverland, 2012). Für die Versuchsplanung und statistische Auswertung liegen dafür zahlreiche einschlägige Verfahren bereit. Sie sind für die experimentelle Unterrichtsforschung teilweise aufgearbeitet und exemplifiziert (Lewis, 1968; Winer, 1971; Cooper, 1988). Seit einiger Zeit wird darüber hinaus das hierarchische lineare Modell (HLM) für die experimentelle Unterrichtsforschung propagiert (Bryk & Raudenbush, 1992; Hox, 2000; de Leeuw & Meijer, 2008), insbesondere dann, wenn der Klassen- und/oder Schulkontext mit einer sehr großen Anzahl von Stichproben (Raudenbush, 2008) berücksichtigt werden soll und aus ökonomischer Sicht auch entsprechende Ressourcen dafür bereitgestellt sind.

Ist das Ziel experimenteller Unterrichtsforschung aber derart, dass die *Generierung* von Hypothesen und der Einblick in *mögliche* Ursache-Wirkungszusammenhänge im Fokus stehen sowie spezifische Repräsentativität verlangt und ein hoher Grad an interner Validität beansprucht werden, sollte mit kleinen Stichproben (*N = small*) geplant werden. Dabei spielt die Fallselektion (Ragin, 2004; George & Bennett, 2005; Blatter & Haverland, 2012) eine herausragende Rolle. Gerring (2007) charakterisiert dafür neun

Techniken: *Typical Case, Diverse Case, Extreme Case, Deviant Case, Influential Case, Crucial Case, Pathway Case, Most-similar Case* und *Most-different Case*. [1]

Bei Anwendung dieser Techniken erhält man in der Regel Stichproben, die relativ klein sind. Für die Auswertung der Daten aus solchen Stichproben sind statistische Verfahren erforderlich, die für kleine Fallzahlen anwendbar sind. Darunter fallen auch so genannte nicht-parametrische Verfahren, die auf Rangdaten operieren.

Statistische Verfahren, die Rangdaten analysieren, sind in fast jedem Statistikbuch beschrieben. Die dort beschriebenen Verfahren beschränken sich jedoch auf wenige Verfahren (z. B. Wilcoxon-, Mann-Whitney-, Kruskal-Wallis-, Friedman-Test), die nur für einfaktorielle Versuchspläne brauchbar sind. Dieses Kapitel liefert einen Beitrag zur Schließung dieser Lücke und schlägt deshalb geeignete Auswertungsansätze für die experimentelle Unterrichtsforschung vor, um Daten in Bezug zu Versuchsplänen auswerten zu können, die für die Unterrichtsforschung prädestiniert sind.

In der Nomenklatur von Kirk (2012) handelt es sich dabei um Versuchspläne der Typen RB-*p* (Randomized Block designs), CR-*p* (Completey Randomized designs), CRF-*p*×*q* (Completey Randomized Factorial designs), SPF-*p*•*q* (Split-Plot Factorial designs) und SPF-*p*×*r*•*q* (Split-Plot Factorial designs). Diese Versuchspläne besitzen für die experimentelle Unterrichtsforschung einen besonderen Stellenwert, besonders zur Untersuchung von Lernerfolgen bei Schülern. Die Versuchspläne sind ausführlich in Kapitel 12 beschrieben.

Die nächsten beiden Abschnitte befassen sich mit methodischen Anforderungen an die Unterrichtsforschung aus Sicht der statistischen Auswertung sowie mit Lösungen für den Umgang mit fehlenden Daten.

Der dann folgende Abschnitt empfiehlt statistische Auswertungsansätze (parametrischer Ansatz, parametrischer Ansatz nach Rangtransformation, nicht-parametrischer Ansatz, *Ranking after Alignment*) in Bezug zu den Versuchsplänen der Typen RB-*p*, CR-*p*, CRF-*p*×*q*, SPF-*p*•*q* und SPF-*p*×*r*•*q*. Der letzte Abschnitt gibt einen Ausblick.

8.2 Anforderungen an die statistische Auswertung

Keselman et al. (1998) untersuchten die statistische Auswertung in 226 Artikeln aus bekannten Zeitschriften zur Unterrichtsforschung (z. B. American Educational Research Journal, Journal of Educational Computing Research, Journal of Experimental

[1] Zur Unterscheidung von quantitativer und qualitativer Forschung in Bezug zur Stichprobengröße sei auf Collier, Bradfy, & Seawright (2004) verwiesen: "A second approach is to identify the qualitative-quantitative distiction with the contrast betwenn small-N and large-N research."

Education). Sie stellen fest: "In particular, applied statisticians have devoted a great deal of effort to understanding the operating characteristics of statistical procedures when the distributional assumptions that underlie a particular procedure are not likely to be satisfied. It is common knowledge that, under certain data-analytic conditions, statistical procedures will not produce valid results. 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." (Keselman et al., 1998, S. 351)

8.2.1 Skalendignität von Produktvariablen

Diese Anforderung besagt, dass zu klären ist, welches Skalenniveau Daten (z. B. Daten zum Lernzuwachs, Daten zur Behaltensleistung) aus empirischen Untersuchungen zur Unterrichtsforschung besitzen, um ein adäquates statistisches Verfahren wählen zu können.

Besitzen Daten nur ordinale Infomationen, sind nicht-parametrische Verfahren bei der Auswertung einzusetzen (Cooper, 1988; Gibbons & Chakraborti, 2003; Kvam & Vidakovic, 2007).

Ist messtheoretisch (Krantz, Luce, Suppes, & Tversky, 1971; Roberts, 1979; Narens, 2014) geklärt, dass Daten intervallskaliert sind, so können parametrische Verfahren eingesetzt werden, wenn weitere Voraussetzungen zu treffen.

8.2.2 Verteilungsannahmen

Diese Anforderung betrifft die Verteilung der Daten in einer Untersuchung. Zu prüfen ist hier im Allgemeinen, ob Daten normalverteilt und varianzhomogen sind.

Aus Reviews zum Einsatz statistischer Verfahren (Goodwin & Goodwin, 1985; West, Carmody, & Stallings; Lix, Keselman, & Keselman, 1996; Keselman et al., 1998; Zendler & Pfeiffer, 2009) ist allerdings bekannt, dass parametrische Verfahren (*t*-test, *F*-test, u.a.) zur Datenauswertung verwendet werden, ohne Verteilungsannahmen (Normalverteilung, Varianzhomogenität) berücksichtigt zu haben.

Eine interessante Alternative stellen nicht-parametrische Verfahren dar, weil ihre Anwendungen im Allgemeinen nur stetig verteilte abhängige Variablen und vielfach nur die Homomerität der Populationsverteilung (Verteilungen der untersuchten Populationen müssen die gleiche Form) voraussetzen (Gibbons & Chakraborti, 2003).

8.3 Zum Umgang mit fehlenden Daten

Fehlende Daten sind in empirischen Untersuchungen normal, auch in der Unterrichtsforschung sind sie kaum zu vermeiden. Für den Umgang mit fehlenden Daten werden in der Praxis oft relativ einfache, aber ungünstige Verfahren eingesetzt (Lüdtke, Robitzsch, Trautwein & Köller, 2007), obwohl mit *Multiple Imputation* eine flexible Methode zur Verfügung steht (Little & Rubin, 2002; Van Buuren, 2012).

8.3.1 Einfache Verfahren

Die einfachen Verfahren umfassen u.a. die listenweise und die paarweise Löschung von Werten sowie die Mittelwert- und die Regressionswertzuschreibung. Abbildung 8.1 veranschaulicht die Prinzipien der Verfahren an einem einfachen Datensatz mit sieben Fällen (s_1, s_2, ..., s_7) und drei Variablen (Y_1, Y_2, Y_3) mit fehlenden Werten (*NA*). Während die listenweise und die paarweise Löschung zu einer Reduzierung der Fälle führt (s_2 und s_5 bzw. s_5 werden gelöscht), bleibt bei Mittelwert- und Regressionswertzuschreibung die Fallzahl komplett.

	Y_1	Y_2	Y_3
s_1	1	2	2
~~s_2~~	~~*NA*~~	~~2~~	~~2~~
s_3	5	1	1
s_4	1	5	1
~~s_5~~	~~2~~	~~*NA*~~	~~*NA*~~
s_6	4	2	2
s_7	6	3	4
$\bar{x}$	3.2	2.6	2.0
n	5	5	5

a

	Y_1	Y_2	Y_3
s_1	1	2	2
s_2	*NA*	2	2
s_3	5	1	1
s_4	1	5	1
~~s_5~~	2	*NA*	*NA*
s_6	4	2	2
s_7	6	3	4
$\bar{x}$	3.2	2.8	2.0
n	6	6	6

b

	Y_1	Y_2	Y_3
s_1	1	2	2
s_2	3.2	2	2
s_3	5	1	1
s_4	1	5	1
s_5	2	2.6	2.0
s_6	4	2	2
s_7	6	3	4
$\bar{x}$	3.2	2.6	2.0
n	7	7	7

c

	Y_1	Y_2	Y_3
s_1	1	2	2
s_2	**4**	2	2
s_3	5	1	1
s_4	1	5	1
s_5	2	**3**	**3**
s_6	4	2	2
s_7	6	3	4
$\bar{x}$	3.3	2.6	2.1
n	7	7	7

d

Abbildung 8.1 Prinzipien der listenweise (*a*) und paarweisen Löschung (*b*), Mittelwert- (*c*) und Regressionswertzuschreibung (*d*)

Listenweise Löschung

Die listenweise Löschung *(listwise deletion*) ist das voreingestellte Verfahren in den meisten Statistikpaketen (z. B. SPSS). Bei diesem Verfahren werden alle Fälle mit einem oder mehreren fehlenden Werten in Bezug zu abhängigen Variablen eliminiert (vgl. Abbildung 8.1a). Der Vorteil des Verfahrens ist seine Einfachheit, der Nachteil der verschwenderische Umgang mit Daten.

Paarweise Löschung

Die paarweise Löschung (*pairwise deletion*) versucht die Nachteile der listenweisen Löschung zu kompensieren. Bei diesem Verfahren werden Mittelwerte und Kovarianzen für beobachtete Daten berechnet. Die Berechnung der Mittelwerte erfolgt über die vorhandenen Daten je Variable (vgl. Abbildung 8.1b), die Berechnung der Kovarianzen erfolgt über die vorhandenen Datenpaare je zweier Variablen.

Mittelwertzuschreibung

Beim Verfahren der Mittelwertzuschreibung (*mean imputation*) werden fehlende Werte durch den Mittelwert ersetzt, getrennt nach Variablen (vgl. Abbildung 8.1c). Das Verfahren ist unkompliziert, allerdings wird die Varianz der Variablen unterschätzt, zudem zerstört es Beziehungen zwischen Variablen.

Regressionswertzuschreibung

Beim Verfahren der Regressionswertzuschreibung (*regression imputation*) werden fehlende Werte in einer (abhängigen) Variable durch berechnete Werte mit Hilfe von Regressionsrechnung ersetzt (vgl. Abbildung 8.1d), wobei ein Modell zwischen abhängiger Variable und einer oder mehreren unabhängigen Variablen zugrunde gelegt sein muss.

8.3.2 *Multiple Imputation*

Das Verfahren der multiplen Imputation geht auf Rubin (1976) zurück. Das Verfahren erzeugt $m > 1$ komplette Datensätze, damit Unsicherheiten beim Ersetzen fehlender Werte berücksichtigt werden können. Abbildung 8.2 veranschaulicht das prinzipielle Vorgehen: (1) Aus einem Datensatz mit fehlenden Werten werden z. B. $m = 3$ Datensätze mit imputierten (zugeschriebenen) Werten erzeugt. Die erzeugten Datensätze sind identisch für die empirisch beobachteten Werte; sie unterscheiden sich in den imputierten Werten. (2) Die erzeugten Datensätze werden statistischen Analysen unterzogen und (3) zu einem endgültigen Datensatz gepoolt.

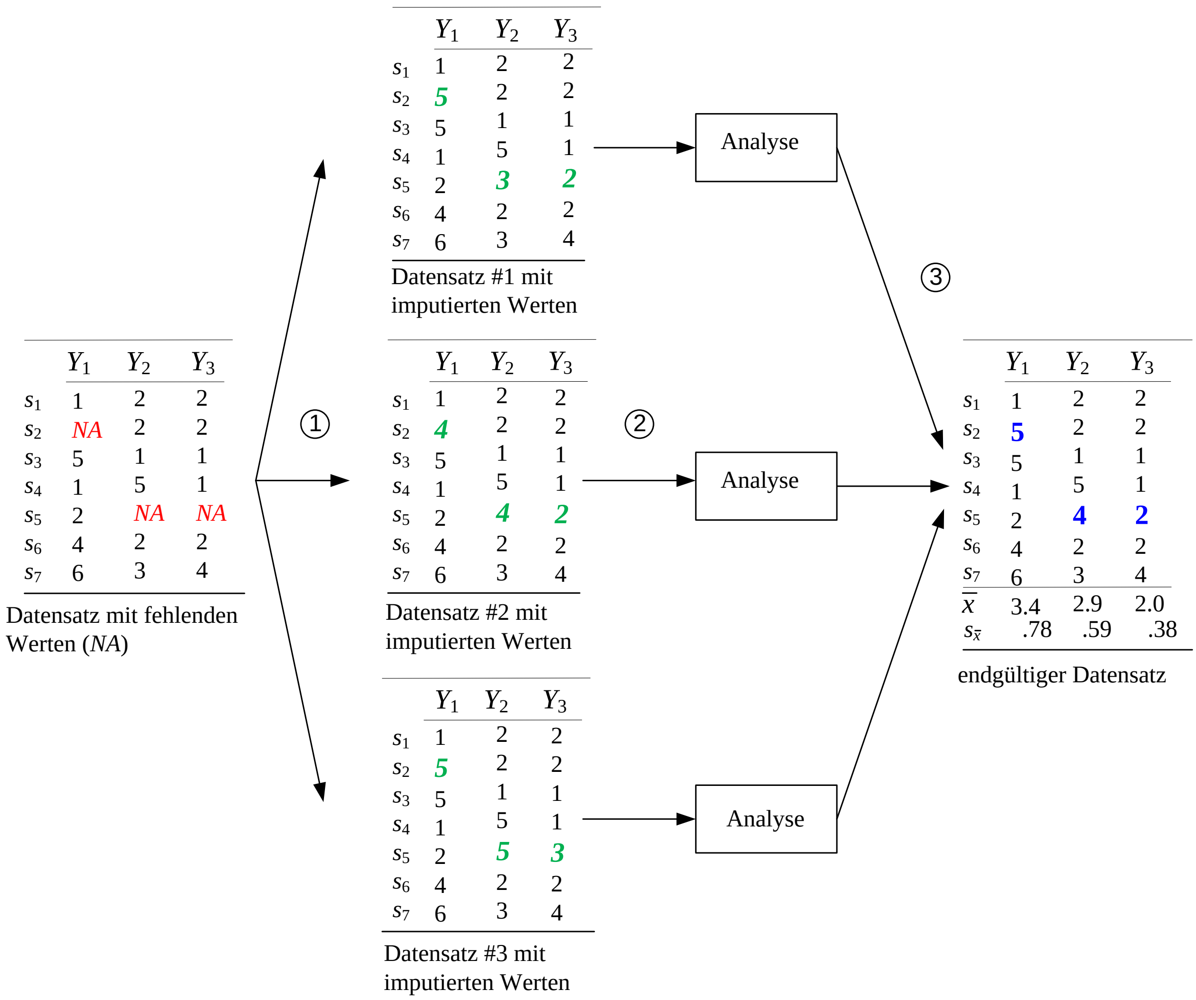

Abbildung 8.2 Verfahren der multiplen Imputation

Multiple Imputation löst zum einen das Problem fehlender Daten. Zum anderen liefert es die Lösung, damit komplette Datensätze (Datensätze mit voller Fallzahl) in die weitere statistische Verarbeitung eingehen können. Zudem löst es das Problem von zu großen und zu kleinen Standardfehlern, die bei listenweiser bzw. paarweiser Löschung, Mittelwert- und Regressionswertzuschreibung auftreten und zu Verfälschungen in nachgeschalteten statistischen Analysen (z.B. Varianzanalysen) führen können.

Das Verfahren der multiplen Imputation steht in SPSS seit Version 17 zur Verfügung. Sehr flexibel einsetzbar ist das Paket *mice* unter R (Van Buuren & Groothuis-Oudshoorn, 2011; Van Buuren, 2012; Kabacoff, 2015), das eine Vielzahl von Funktionen bereitstellt, welche

- die Inspektion von Mustern fehlender Daten,
- die Erzeugung von Datensätzen mit imputierten Werten,
- die Diagnose imputierter Werte,

- die Analyse kompletter Datensätze,
- das Pooling wiederholter Analysen,
- das Speichern und Exportieren von Datensätzen mit imputierten Werten,
- die Generierung von Datensätzen,
- den Einbau benutzerspezifischer Imputationsmethoden

erlauben.

Bereitgestellte Vignetten im *mice-Paket* unter R führen das Verfahren der multiplen Imputation Schritt für Schritt ein, mit Vignetten für einfache Datensätze bis hin zu Datensätzen aus komplexen (multilevel) Versuchsplänen.

8.4 Statistische Auswertungsansätze

Dieser Abschnitt hat statistische Auswertungsansätze zum Inhalt, die für die Datenauswertung der experimentellen Unterrichtsforschung von zentraler Bedeutung sind. Abbildung 8.3 zeigt in einer Entscheidungstabelle, unter welchen Bedingungen die Auswertungsansätze zu wählen sind: (1) Der parametrische Ansatz ist zu wählen, wenn die Daten intervallskaliert, normalverteilt und varianzhomogen sind, unabhängig davon, ob eine oder mehrere abhängigen Variablen vorhanden sind. (2) Der parametrische Ansatz nach Rangtransformation ist zu wählen, wenn die Daten intervallskaliert sind, diese jedoch nicht normalverteilt oder varianzhomogen sind. (3) Der nicht-parametrische Ansatz ist zu wählen, wenn die Daten ordinalskaliert sind. Sind mehrere abhängige Variablen vorhanden, sind diese (im Allgemeinen) separat auszuwerten. (4) *Ranking after Alignment* ist zu wählen, wenn die Daten ordinalskaliert sind und mehrere abhängige Variablen gemeinsam berücksichtigt werden sollen.

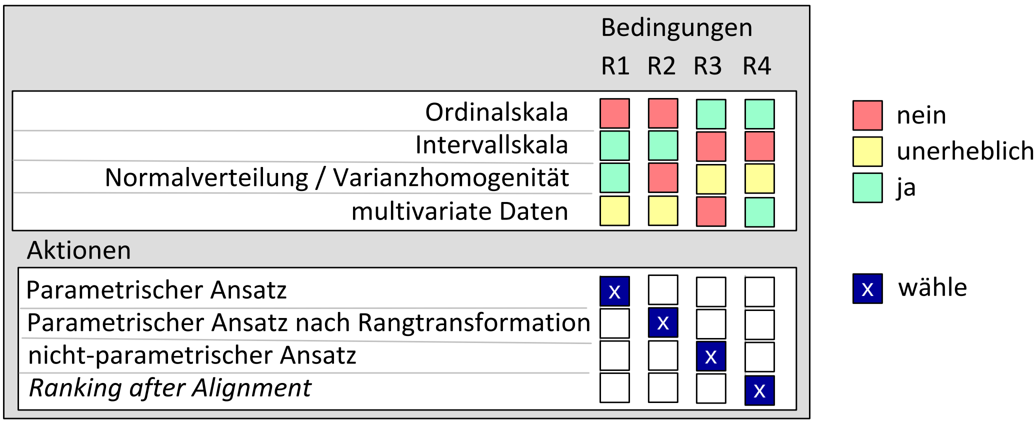

Abbildung 8.3 Entscheidungstabelle für Auswertungsansätze

Die Auswertungsansätze werden am Beispiel eines CRF-3×2 Versuchsplans mit p = 3 Faktorstufen für den Faktor Unterrichtsmethode (Computersimulation, Lernen durch Lehren, direkte Instruktion) und q = 2 Faktorstufen für den Faktor Schulklasse (Klasse #1, Klasse #2) exemplifiziert.

Mit CRF-p×q Versuchsplänen lässt sich allgemein überprüfen, wie eine abhängige Variable von zwei Faktoren A und B beeinflusst wird. Der Faktor A sei p-fach, der Faktor B q-fach gestuft. CRF-p×q Versuchspläne sind für die experimentelle Unterrichtsforschung wichtig, denn sie erlauben neben der Überprüfung zweier Haupteffekte, die Überprüfung von pädagogisch interessanten Interaktionseffekten. Somit lassen sich Interaktionen wie Unterrichtsmethode × Schulklasse, Unterrichtsmethode × Schulklassenmerkmal oder Unterrichtsmethode × Schülermer kmal prüfen.

Für die experimentelle Unterrichtsforschung sind CRF-p×q Versuchspläne durch vier Aspekte charakterisiert: (1) Unter dem Faktor A lassen sich $p \geq 2$ Bedingungen des Lehrens und Lernens gruppieren; für p = 2 bietet es sich an Halbklassen zu verwenden. (2) Unter dem Faktor B können $q \geq 2$ Schulklassen (bzw. Schulklassenmerkmale oder Schülermerkmale) berücksichtigt werden. (3) Lerner werden zufällig den Faktorstufenkombinationen zugeordnet. (4) es werden pq Stichproben mit Lernern benötigt.

Abbildung 8.4 veranschaulicht das Layout eines CRF-3×2 Versuchsplans mit p = 3 Unterrichtsmethoden (Computersimulation, Lernen durch Lehren, direkte Instruktion) und q = 2 Faktorstufen für den Faktor Schulklasse (Klasse #1, Klasse#2); insgesamt werden 3 × 2 = 6 Stichproben benötigt.

a_1	a_1	a_2	a_2	a_3	a_3
b_1	b_2	b_1	b_2	b_1	b_2
S_1	S_2	S_3	S_4	S_5	S_6

A = Unterrichtsmethode
a_1 = Computersimulation
a_2 = Lernen durch Lehren
a_3 = Direkte Instruktion

B= Schulklasse
b_1 = Klasse #1
b_2 = Klasse #2

Abbildung 8.4 Layout eines CRF-3×2 Versuchsplans

Tabelle 8.1 enthält den Datensatz für einen CRF-3×2 Versuchsplan mit jeweils 7 Lernern. Von den Lernern wird 1 Datensatz zum Lernerfolg [Note] erhoben. In der Tabelle sind auch die Mittelwerte und die Standardfehler der Mittelwerte berechnet; sie werden benötigt für die Berechnung von Konfidenzintervallen und Prüfgrößen.

Tabelle 8.1 Datensatz zum CRF-3×2 Versuchsplan

	a_1	a_1	a_2	a_2	a_3	a_3
	b_1	b_2	b_1	b_2	b_1	b_2
	3.0	3.0	2.5	2.0	2.0	1.0
	2.5	3.0	2.5	3.0	1.0	1.0
	3.0	2.5	3.0	2.0	3.0	3.0
	2.5	2.5	2.5	3.0	3.0	3.0
	3.5	3.5	2.0	2.0	3.0	2.0
	3.0	3.0	3.0	2.5	3.0	2.0
	2.5	4.0	2.0	2.5	3.0	2.0
$\overline{x}$	2.9	3.1	2.5	2.4	2.6	2.0
$s_{\overline{x}}$	.22	.22	.24	.21	.22	.22

Abbildung 8.5 visualisiert die Datensätze (Boxplots, Mittelwert, 95% Konfidenzintervalle) aus Tabelle 8.1. Aus der Abbildung lässt sich entnehmen, dass sich die Unterrichtsmethoden hinsichtlich ihrer Lernwirksamkeit unterscheiden: Mit der Unterrichtsmethode der Computersimulation wird der niedrigste, mit Lernen durch Lehren der zweithöchste, mit der direkten Instruktion der höchste Lernerfolg bei den Lernern erzielt – und zwar in beiden Schulklassen.

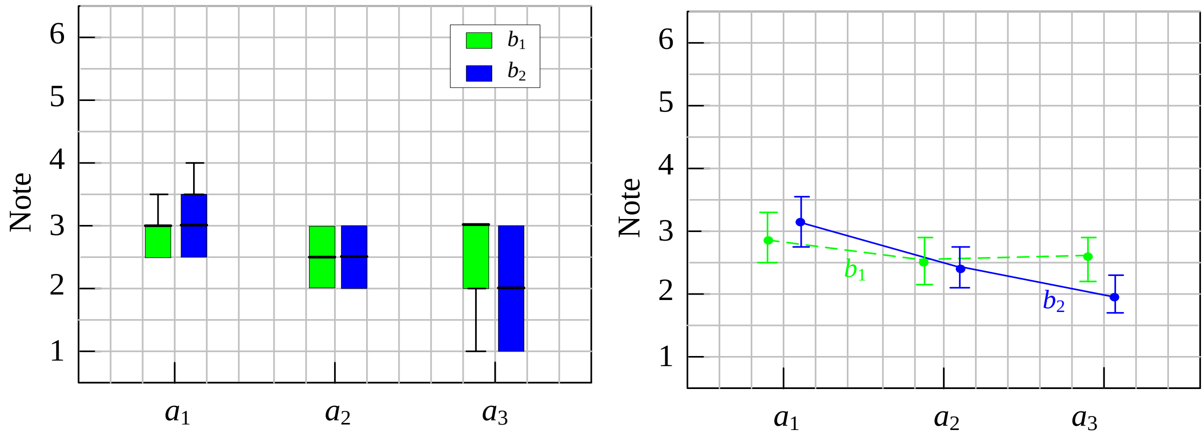

Abbildung 8.5 Boxplots, Mittelwerte und 95%-Konfidenzintervalle

8.4.1 Parametrischer Ansatz

Der parametrische Ansatz umfasst statistische Auswertungsverfahren (*t*-Tests, Varianzanalysen, u.a.), die Annahmen über die Wahrscheinlichkeitsverteilung von Beobachtungsdaten (Messwerten) machen, wobei die Verteilungen (z. B. Normalverteilung) durch Parameter (z. B. Erwartungswert, Varianz) bestimmt sind.

Die wichtigsten Voraussetzungen für die Anwendbarkeit parametrischer Verfahren sind:

1. Skalendignität der Daten: Intervallskala

2. Verteilung der Daten: Normalverteilung, Varianzhomogenität

Die Voraussetzung, ob Daten intervallskaliert sind, ist messtheoretisch zu untersuchen (Krantz, Luce, Suppes, & Tversky, 1971; Robert, 1971; Narens, 2014). Ob Daten normalverteilt sind, lässt sich mit dem Shapiro-Wilk-Test (Royston, 2006) prüfen – schon für Stichproben $N > 3$. Ob Daten die Voraussetzung der Varianzhomogenität erfüllen, wird etwa mit dem Levene-Test (Glaser, 2006) untersucht.

Parametrische Verfahren sind für die Auswertung univariater und multivariater Daten in der gängigen Statistikliteratur (Bortz & Schuster, 2010; Winer, 1971; Tabachnick & Fidell, 2014) dokumentiert und in Statistikpaketen wie SPSS (Field, 2013), R (Field & Miles, 2012) und SAS (Field & Miles, 2010) sehr leicht zugänglich.

Für die Auswertung von Daten der genannten Versuchspläne haben Varianzanalysen mit und ohne Messwiederholungen eine besondere Bedeutung. Ergebnisse von Varianzanalysen werden in Tabellen dargestellt, welche nach APA (2010) die folgenden Angaben enthalten sollten: Quadratsummen (*QS*), Freiheitsgrade (*df*), mittlere Quadrate (*MS*), Prüfgröße *F*, *p*-Wert (statistische Signifikanz), (partielle) Effektgröße η^2.

Für die Daten des CRF-3×2 Versuchsplans (vgl. Tabelle 8.1) enthält Tabelle 8.2 die varianzanalytische Auswertung.

Tabelle 8.2 Varianzanalyse für die Daten zum Versuchsplan CRF-3×2

Quelle	*QS*	*df*	*MS*	*F*	*p*	partielles η^2
A	3.46	2	1.73	4.99	< .012	.217
B	.21	1	.21	0.62	< .437	.017
A×*B*	1.11	2	.55	1.59	< .217	.081
Fehler	12.50	36	.35			
Total	295.00	42				

Tabelle 8.2 zeigt einen signifikanten Effekt für Faktor *A*: Die Unterrichtsmethoden (Computersimulation, Lernen durch Lehren, direkte Instruktion) unterscheiden sich hinsichtlich ihrer Lernwirksamkeit signifikant ($p < .012$). Faktor *B* ist nicht signifikant. die Schulklassen (Klasse #1 und Klasse #2) unterscheiden sich in Bezug auf den Lernerfolg nicht ($p < .437$), auch ist die Interaktion $A \times B$ (Unterrichtsmethode x Schulklasse) nicht signifikant ($p < .217$). Das heißt, die Unterrichtsmethoden unterscheiden sich (Computersimulation, Lernen durch Lehren, direkte Instruktion), ohne dass eine Verbindung zum Klassenkontext (Klasse #1, Klasse #2) hergestellt werden müsste.

8.4.2 Parametrischer Ansatz nach Rangtransformation

Der parametrische Ansatz nach Rangtransformation hat dieselben statistischen Verfahren (*t*-Tests, Varianzanalysen, u.a.) zum Gegenstand wie der parametrische Ansatz. Der Unterschied besteht darin, dass vor Verwendung der statistischen Verfahren eine Rangtransformation (Conover & Iman, 1981; Sawilowsky, Blair, & Higgins, 1989) der Daten vorzunehmen ist. Der Ansatz ist dann indiziert, wenn Daten Intervallskalendignität besitzen und den gleichen Verteilungstyp haben.

Die wichtigsten Voraussetzungen für die Anwendbarkeit des parametrischen Ansatzes nach Rangtransformation sind:

1. Skalendignität der Daten: Intervallskala

2. Verteilung der Daten: gleicher Verteilungstyp der Daten für alle Faktorstufen

Die Voraussetzung, ob Daten intervallskaliert sind, ist messtheoretisch zu untersuchen (siehe Abschnitt 8.1). Ob Daten vom gleichen Verteilungstyp sind, lässt sich explorativ mit einem *Stem-and-Leaf*-Diagramm (Stängel-Blatt-Diagramm, siehe Bortz & Schuster, 2010, Kapitel 6) prüfen. Die Prüfung kann mit SPSS durchgeführt werden.

Der parametrische Ansatz nach Rangtransformation verwendet parametrische Verfahren, die in der gängigen Literatur beschrieben sind und in Statistikpaketen (SPSS, R, SAS) sehr leicht zugänglich sind (vgl. Abschnitt 8.1).

Abbildung 8.6 veranschaulicht die Rangtransformation für den Datensatz zum Versuchsplan CRF-3×2 aus Tabelle 8.1.

	a_1 b_1	a_1 b_2	a_2 b_1	a_2 b_2	a_3 b_1	a_3 b_2
	3.0	3.0	2.5	2.0	2.0	1.0
	2.5	3.0	2.5	3.0	1.0	1.0
	3.0	2.5	3.0	2.0	3.0	3.0
	2.5	2.5.	2.5	3.0	3.0	3.0
	3.5	3,5	2.0	2.0	3.0	2.0
	3.0	3.0	3.0	2.5	3.0	2.0
	2.5	4.0	2.0	2.5	3.0	2.0
$\overline{x}$	2.9	3.1	2.5	2.6	2.0	2.0
$s_{\overline{x}}$	.14	.20	.15	.23	.22	.31

Rangtransformation →

a_1 b_1	a_1 b_2	a_2 b_1	a_2 b_2	a_3 b_1	a_3 b_2
31	31	17.5	8	8	2
17.5	31	17.5	31	2	2
31	17.5	31	8	31	31
17.5	17.5	17.5	31	31	31
40.5	40.5	8	8	31	8
31	31	31	17.5	31	8
17.5	42	8	17.5	31	8

Abbildung 8.6 Rangtransformation des Datensatzes zum Versuchsplan CRF-3×2

Tabelle 8.3 enthält die varianzanalytische Auswertung für die rangtransformierten Daten zum CRF-3×2 Versuchsplan.

Tabelle 8.3 Varianzanalyse für die rangtransformierten Daten

Quelle	*QS*	*df*	*MS*	*F*	*p*	*partielles* η^2
A	977.14	2	488.80	4.20	< .023	.189

B	85.71	1	85.71	0.74	< .396	.020
A×B	365.39	2	182.70	1.57	< .222	.080
Fehler	4188.79	36	116.36			
Total	25032.00	42				

Tabelle 8.3 zeigt ein ähnliches Ergebnis wie Tabelle 8.2: Die Unterrichtsmethoden (Computersimulation, Lernen durch Lehren, direkte Instruktion) unterscheiden sich hinsichtlich ihrer Lernwirksamkeit signifikant ($p < .023$). Die Schulklassen (Klasse #1 und Klasse #2) unterscheiden sich in Bezug auf den Lernerfolg nicht signifikant ($p < .396$), und die Interaktion Unterrichtsmethode x Schulklasse ist nicht signifikant ($p < .222$). Das heißt, die Unterrichtsmethoden unterscheiden sich (Computersimulation, Lernen durch Lehren, direkte Instruktion), ohne dass eine Verbindung zum Klassenkontext (Klasse #1, Klasse #2) hergestellt werden müsste.

8.4.3 Nicht-parametrischer Ansatz

Der nicht-parametrische Ansatz bezieht sich auf statistische Verfahren (*U*-Test, Wilcoxon-Test, Rangvarianzanalysen, u.a.), deren Anwendbarkeit wesentlich geringere Annahmen an die Beobachtungsdaten macht als die parametrischen Ansätze. Die Anwendbarkeit des nicht-parametrischen Ansatzes setzt im Allgemeinen nur eine stetig verteilte, abhängige Variable und nur die Homomerität der Populationsverteilung voraus, das heißt, die Verteilungen der untersuchten Populationen müssen vom gleichen Typ sein.

Die Voraussetzungen für die Anwendbarkeit des nicht-parametrischen Ansatzes sind:

1. Skalendignität der Daten: Ordinalskala

2. Verteilung der Daten: Verteilungen der Daten haben die gleiche Form (für alle Faktorstufen)

3. Datensatz: univariat

Die Voraussetzung, ob Daten ordinalskaliert sind, ist unproblematisch, wenn zwischen den Daten größer/kleiner/gleich-Beziehungen bestehen. Die Verteilungsform (z. B. unimodal versus bimodal) der Daten lässt sich leicht explorativ mit einem *Stem-and-Leaf*-Diagramm (Stängel-Blatt-Diagramm, siehe Bortz & Schuster, 2010, Kapitel 6) prüfen. Die Prüfung lässt sich mit SPSS durchführen.

Einfaktorielle Versuchspläne

Statistische Verfahren des nicht-parametrischen Ansatzes sind für die Auswertung der einfaktoriellen Versuchspläne vom Typ RB-*p* und CR-*p* in der gängigen Statistikliteratur

(Bortz, Lienert, & Boehnke, 2008; Siegel & Castellan, 1988; Conover, 1999; Sheskin, 2011) beschrieben (siehe Tabelle 8.4, Zeile 1 und Zeile 2).

Für die einfaktoriellen Versuchspläne vom Typ RB-*p* und CR-*p* sind statistische Auswertungsroutinen in den Statistikpaketen SPSS (Field, 2013) und R (Field & Miles, 2012) leicht zugänglich.

Mehrfaktorielle Versuchspläne

Für den mehrfaktoriellen Fall sind statistische Verfahren des nicht-parametrischen Ansatzes nur in der statistischen Spezialliteratur veröffentlicht (Hettmansperger, 1984; Thompson, 1991; Puri & Sen, 1993; Brunner & Dette, 1992; Brunner, Dette, & Munk, 1997; Brunner & Langer, 1999; Brunner, Domhof, & Langer, 2002; Brunner & Munzel, 2013; Noguchi, Gel, Brunner, & Konietschke, 2012).

Für die mehrfaktoriellen Versuchspläne vom Typ CRF-$p \times q$, CRF-$p \times q \times r$, SPF-$p \bullet q$ und SPF-$p \times q \bullet r$ sind in SAS Routinen für die Auswertung verfügbar (Brunner & Langer, 1999; Brunner, Domhof, & Langer, 2002; Brunner & Munzel, 2013). Tabelle 8.4 (Zeile 3, 4, 5 und 6) gibt die exakten Seitenzahlen für diese Verfahren in der Literatur an.

Noguchi, Gel, Brunner und Konietschke (2012) exemplifizieren in ihrem Softwarepaket *nparLD* unter R die Datenauswertung für Versuchspläne der Typen RB-*p*, SPF-$p \bullet q$ und SPF-$p \times q \bullet r$.

Tabelle 8.4 Versuchspläne und nicht-parametrische statistische Verfahren

Versuchsplan	statistisches Verfahren	Literaturangabe
RB-*p*	für $p = 2$: Wilcoxon-Test	Bortz, Lienert, & Boehnke (2010, S. 259)
	für $p > 2$: Friedman-Test	Bortz, Lienert, & Boehnke (2008, S. 267)
CR-*p*	für $p = 2$: *U*-Test	Bortz, Lienert, & Boehnke (2010, S. 200)
	für $p > 2$: *H*-Test (Kruskal-Wallis)	Bortz, Lienert, & Boehnke (2008, S. 222)
CRF-$p \times q$	für $p \geq 2$, $q \geq 2$: F_N-Test	Brunner & Munzel (2013, S. 137–139)
CRF-$p \times q \times r$	für $p \geq 2$, $q \geq 2$, $r \geq 2$: F_N-Test	Brunner & Munzel (2013, S. 171–185)
SPF-$p \bullet q$	für $p = 2$, $q = 2$: U_n-Test	Brunner & Langer (1999, S. 94–95)
	für $p > 2$, $q > 2$: Q_n-Test	Brunner & Langer (1999, S. 112–122)
SPF-$p \times q \bullet r$	für $p \geq 2$, $q \geq 2$:, $r \geq 2$: Q_n-Test	Noguchi, Gel, Brunner, & Konietschke (2012, S. 6)

Tabelle 8.5 enthält die rangtransformierten Daten von $n_{ij} = 7$ ($i = 1, \ldots, 3$; $j = 1, 2$) Lernern zum CRF-3×2 Versuchsplan aus Tabelle 8.1 mit den Rängen R_{ijk} ($i = 1, \ldots, 3$; $j = 1, 2$; $k = 1, \ldots, 7$) und den Rangmittelwerten $\bar{R}_{ij.}$.

Tabelle 8.5 Rangtransformierte Daten zum Versuchsplan CRF-3×2

a_1 b_1	a_1 b_2	a_2 b_1	a_2 b_2	a_3 b_1	a_3 b_2
31	31	17.5	8	8	2
17.5	31	17.5	31	2	2
31	17.5	31	8	31	31
17.5	17.5	17.5	31	31	31
40.5	40.5	8	8	31	8
31	31	31	17.5	31	8
17.5	42	8	17.5	31	8

Für die rangtransformierten Daten des CRF-3×2 Versuchsplan wird der F_N-Test verwendet, wie er in Brunner und Munzel (2013, S. 137–139) beschrieben ist. Der Test überprüft die Auswirkungen eines p-fach gestuften Faktors A und eines q-fach gestuften Faktors B auf eine abhängige Variable. Mit dem F_N-Test lassen sich prinzipiell dieselben Nullhypothesen wie mit einer zweifaktoriellen Varianzanalyse prüfen.

Unter Verwendung der Hilfsgrößen $\tilde{R}_{i..} = \frac{1}{q} \cdot \sum_{j=1}^{q} \bar{R}_{ij.}$, $\tilde{R}_{...} = \frac{1}{p \cdot q} \cdot \sum_{i=1}^{p} \sum_{j=1}^{q} \bar{R}_{ij.}$ und

$S_0^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{n_{ij}} (R_{ijk} - \bar{R}_{ij.})^2 / \left[n_{ij} \cdot (n_{ij} - 1) \right]$ lauten die Teststatistiken

unter H_0^A: $$F_N(T_A) = \frac{p \cdot q^2}{(p-1) \cdot S_0^2} \cdot \sum_{i=1}^{p} (\tilde{R}_{i..} - \tilde{R}_{...})^2 ,$$

unter H_0^B: $$F_N(T_B) = \frac{q \cdot p^2}{(q-1) \cdot S_0^2} \cdot \sum_{j=1}^{q} (\tilde{R}_{.j.} - \tilde{R}_{...})^2 ,$$

unter $H_0^{A \times B}$: $$F_N(T_{AB}) = \frac{p \cdot q}{(p-1) \cdot (q-1) \cdot S_0^2} \cdot \sum_{i=1}^{p} \sum_{j=1}^{q} (\bar{R}_{ij.} - \tilde{R}_{i..} - \tilde{R}_{.j.} + \tilde{R}_{...})^2 .$$

Die Teststatistik $F_N(T_A)$ ist asymptotisch χ^2-verteilt mit $df = p–1$ Freiheitsgraden, $F_N(T_B)$ asymptotisch χ^2-verteilt mit $df = q–1$ und $F_N(T_{AB})$ asymptotisch χ^2-verteilt mit $df = (p–1)(q–1)$ Freiheitsgraden.

Tabelle 8.6 fasst die Ergebnisse des F_N-Tests für die rangtransformierten Daten aus Tabelle 8.5 zusammen.

Tabelle 8.6 F_N-Test für die rangtransformierten Daten

Quelle der Variation	F_N	*df*	*p*
A	13.33	2	< .013
B	.16	1	< .690
A×B	.64	2	< .726
Total	14.13	5	< .015

Die Auswertung in Tabelle 8.6 mit dem F_N-Test zeigt dieselben signifikanten und nicht-signifikanten Ergebnisse, wie sie auch mit den beiden parametrischen Ansätzen (vgl. Abschnitte 8.1 und 8.2) ermittelt wurden (allerdings mit Unterschieden in den *p*-Werten). Die Unterrichtsmethoden (Computersimulation, Lernen durch Lehren, direkte Instruktion) unterscheiden sich hinsichtlich ihrer Lernwirksamkeit signifikant ($p < .013$). Die Schulklassen (Klasse #1 und Klasse #2) unterscheiden sich in Bezug auf den Lernerfolg nicht signifikant ($p < .690$); auch ist die Interaktion Unterrichtsmethode x Schulklasse nicht signifikant ($p < .726$).

8.4.4 *Ranking after Alignment*

Ranking after Alignment zielt darauf ab, Daten vor Durchführung eines statistischen Tests für die Untersuchung eines speziellen Effekts hinsichtlich anderer möglicher Effekte zu bereinigen (vgl. Bortz, Lienert, & Boehnke, 2008). Das Verfahren geht auf Hildebrand (1980) zurück. *Ranking after Alignment* ist für die Auswertung univariater Daten in der gängigen Literatur zu verteilungsfreien Verfahren beschrieben (Bortz, Lienert, & Boehnke, 2008), für multivariate Daten ist es nur in der statistischen Spezialliteratur zugänglich (Puri & Sen, 1993, S. 286–308). In den Statistikpaketen SPSS und SAS *Ranking after Alignment* nicht verfügbar. Seit kurzem ist unter R das package *ARtool* verfügbar (Kay, 2018), das *Ranking after Alignment* für Datensätze in Bezug zu Versuchspläne der Typen RB-*p*, CRF-*p×q,* SPF-*p•q* und SPF-*p×q•r* erlaubt.

Die wichtigsten Voraussetzungen für die Anwendbarkeit des Verfahrens sind dieselben wie für den nicht-parametrischen Ansatz (siehe Abschnitt 8.1), allerdings kann *Ranking after Alignment* auch für den multivariaten Fall eingesetzt werden:

1. Skalendignität der Daten: Ordinalskala

2. Verteilung der Daten: Verteilungen der Daten haben die gleiche Form (für alle Faktorstufen)

3. Datensatz: univariat oder multivariat

Das Prinzip des *Ranking after Alignment* besteht darin, Daten so zu transformieren, dass interessierende Effekte mit einfaktoriellen, nichtparametrischen Verfahren (*H*-Test, Friedman-Test, u.a.) ausgewertet werden können.

Am Beispiel des Datensatzes zum CRF-3×2 Versuchsplan aus Tabelle 8.2 werden die Schritte beim *Ranking after Alignment* gezeigt.

1. Berechnung der Mediane

Es wird davon ausgegangen, dass die Rohdaten zum Versuchsplan CRF-3×2 nur ordinale Informationen besitzen. Deshalb werden zuerst Mediane (*Md*) unter den Faktoren *A* und *B* sowie unter den Faktorstufenkombinationen *A* x *B* berechnet (siehe Abbildung 8.7).

Rohdaten

a_1 b_1	a_1 b_2	a_2 b_1	a_2 b_2	a_3 b_1	a_3 b_2
3.0	3.0	2.5	2.0	2.0	1.0
2.5	3.0	2.5	3.0	1.0	1.0
3.0	2.5	3.0	2.0	3.0	3.0
2.5	2.5	2.5	3.0	3.0	3.0
3.5	3.5	2.0	2.0	3.0	2.0
3.0	3.0	3.0	2.5	3.0	2.0
2.5	4.0	2.0	2.5	3.0	2.0

Medianberechnung $\longrightarrow$

Mediane (*Md*)

	a_1	a_2	a_3	b_j
b_1	3.0	2.5	2.0	2.5
b_2	3.0	2.5	2.0	2.5
a_i	3.0	2.5	2.0	2.5

Abbildung 8.7 Mediane für den Datensatz zum Versuchsplan CRF-3×2

2. Alignierung der Rohdaten

Für die Alignierung des Rohdatensatzes werden folgende Berechnungsvorschriften verwendet:

unter *A*: $x_{ijm} = x_{ijm} - Md\,(ab_{ij}) + Md\,(a_{i.}),$

unter *B*: $x_{ijm} = x_{ijm} - Md\,(ab_{ij}) + Md\,(b_{.i}),$

unter *A* x *B*: $x_{ijm} = x_{ijm} - Md\,(a_{j.}) + Md\,(b_{.j}) - 2\,Md(ab_{..})$.

Abbildung 8.8 veranschaulicht die berechneten alignierten Datensätze für die Faktoren *A* und *B* sowie für die Faktorkombination *A* x *B*

Alignierter Datensatz für *A*

a_1 b_1	a_1 b_2	a_2 b_1	a_2 b_2	a_3 b_1	a_3 b_2
3.0	3.0	2.5	2.0	2.0	1.0
2.5	3.0	2.5	3.0	1.0	1.0
3.0	2.5	3.0	2.0	3.0	3.0
2.5	2.5	2.5	3.0	2.0	3.0
3.5	3.5	2.0	2.0	2.0	2.0
3.0	3.0	3.0	2.5	2.0	2.0
2.5	4.0	2.0	3.5	2.0	2.0

Alignierter Datensatz für *B*

a_1 b_1	a_1 b_2	a_2 b_1	a_2 b_2	a_3 b_1	a_3 b_2
2.5	2.5	2.5	2.0	2.5	1.5
2.0	2.0	2.5	3.0	1.5	1.5
2.5	2.0	3.0	2.0	3.5	3.5
2.0	2.0	2.5	3.0	2.5	3.5
3.0	3.0	2.0	2.0	2.5	2.5
2.5	2.5	3.0	2.5	2.5	2.5
2.0	3.5	2.0	3.5	2.5	2.5

Alignierter Datensatz für $A \times B$

a_1 b_1	a_1 b_2	a_2 b_1	a_2 b_2	a_3 b_1	a_3 b_2
2.5	2.5	2.5	2.0	2.5	1.5
2.0	2.5	2.5	3.0	1.5	1.5
2.5	2.0	3.0	2.0	3.5	3.5
2.0	2.0	2.5	3.0	2.5	3.5
3.0	3.0	2.0	2.0	2.5	2.5
2.5	2.5	3.0	2.5	2.5	2.5
2.0	3.5	2.0	3.5	2.5	2.5

Abbildung 8.8 Alignierte Datensätze zum Versuchsplan CRF-3×2

3. Rangtransformation der alignierten Daten

Im dritten Schritt werden die alignierten Datensätze einer Rangtransformation unterzogen. Ergebnis dieses Schrittes sind drei rangtransformierte, alignierte Datensätze (vgl. Abbildung 8.9).

Rangtransformierter, alignierter Datensatz für *A*

a_1 b_1	a_1 b_2	a_2 b_1	a_2 b_2	a_3 b_1	a_3 b_2
32.0	32.0	21.0	10.0	10.0	2.0
21.0	32.0	21.0	32.0	2.0	2.0
32.0	21.0	32.0	10.0	32.0	32.0
21.0	21.0	21.0	32.0	10.0	32.0
40.0	40.0	10.0	10.0	10.0	10.0
32.0	32.0	32.0	21.0	10.0	10.0
21.0	42.0	10.0	40.0	10.0	10.0

Rangtransformierter, alignierter Datensatz für *B*

a_1 b_1	a_1 b_2	a_2 b_1	a_2 b_2	a_3 b_1	a_3 b_2
22.5	22.5	22.5	8.5	22.5	2.0
8.5	22.5	22.5	34.5	2.0	2.0
22.5	8.5	34.5	8.5	40.0	40.0
8.5	8.5	22.5	34.5	22.5	40.0
34.5	34.5	8.5	8.5	22.5	22.5
22.5	22.5	34.5	22.5	22.5	22.5
8.5	40.0	8.5	40.0	22.5	22.5

Rangtransformierter, alignierter Datensatz für $A \times B$

a_1 b_1	a_1 b_2	a_2 b_1	a_2 b_2	a_3 b_1	a_3 b_2
22.5	22.5	22.5	8.5	22.5	2.0
8.5	22.5	22.5	34.5	2.0	2.0
22.5	8.5	34.5	8.5	40.0	40.0
8.5	8.5	22.5	34.5	22.5	40.0
34.5	34.5	8.5	8.5	22.5	22.5
22.5	22.5	34.5	22.5	22.5	22.5
8.5	40.0	8.5	40.0	22.5	22.5

Abbildung 8.9 Rangtransformatierte, alignierte Datensätze

4. Auswertung der rangtransformierten, alignierten Daten

Im vierten Schritt werden die rangtransformierten, alignierten Daten mit dem entsprechenden einfaktoriellen, nichtparametrischen Verfahren (*H*-Test) ausgewertet.

Die Teststatistik lautet: $H = \frac{12}{N \cdot (N+1)} \cdot \sum_{i=1} \frac{T_i^2}{n_i} - 3 \cdot (N+1)$.

Sie ist für den Faktor *A* χ^2-verteilt mit $df = p–1$ Freiheitsgraden, für den Faktor *B* χ^2-verteilt mit $df = q–1$ Freiheitsgraden, für die Interaktion $A \times B$ χ^2-verteilt mit $df = (p–1)(q–1)$ Freiheitsgraden.

Tabelle 8.7 enthält die Auswertung mit dem einfaktoriellen *H*-Test für die rangtransformierten alignierten Daten zum Versuchsplan CRF-3×2.

Tabelle 8.7 Einfaktorieller *H*-Test für die rangtransformierten, alignierten Daten

Quelle der Variation	χ^2	*df*	*p*
A	14.32	2	< .002
B	.12	1	< .730
A × *B*	.93	2	< .978

Die Auswertung in Tabelle 8.7 mit dem *ein*faktoriellen *H*-Test zeigt dieselben signifikanten und nicht-signifikanten Ergebnisse, die mit den F_N-Tests (vgl. Abschnitt 8.3) erhalten wurden. Die Unterrichtsmethoden (Computersimulation, Lernen durch Lehren, direkte Instruktion) unterscheiden sich in Bezug zu ihrer Lernwirksamkeit signifikant ($p < .002$). Die Schulklassen (Klasse #1 und Klasse #2) unterscheiden sich in Bezug auf den Lernerfolg nicht signifikant ($p < .730$), auch ist die Interaktion Unterrichtsmethode x Schulklasse nicht signifikant ($p < .978$). Deshalb kann festgestellt werden, dass sich Unterrichtsmethoden (Computersimulation, Lernen durch Lehren, direkte Instruktion) unterscheiden, ohne dass eine Verbindung zum Klassenkontext (Klasse #1, Klasse #2) herzustellen ist.

Mehrfaktorielle Versuchspläne

Ranking after Aligment ist universell einsetzbar. Wobbrock, Findlater, Gergle und Higgins (2011) liefern für ein-, zwei-, drei-, vier- und *n*-faktorielle Versuchspläne entsprechende Berechnungsvorschriften.

Multivariate Versuchspläne

Auch für multivariate, mehrfaktorielle Versuchspläne ist *Rank after Alignment* einsetzbar (Puri & Sen, 1993, S. 286–308), wenn Daten so transformiert sind, dass sie mit einfaktoriellen, multivariaten, nicht-parametrischen Verfahren (multivariater *H*-Test, multivariater Friedman-Test) ausgewertet werden können (Katz & McSweeney, 1980; Gerig, 1969).

8.5 Zusammenfassung und Ausblick

Dieses Kapitel skizzierte vier statistische Ansätze (parametrischer Ansatz, parametrisches Ansatz nach Rangtransformation, nicht-parametrischer Ansatz, *Ranking after Aligment*), die es erlauben, Daten bezüglich wichtiger Versuchspläne der Unterrichtsforschung (RB-*p*, CR-*p*, CRF-*p*×*q*, SPF-*p*•*q*, SPF-*p*×*r*•*q*) auswerten zu können, je nachdem welche Datenvoraussetzungen zutreffen.

Besonderes Augenmerk wurde auf neuere Ansätze des nicht-parametrischen Ansatzes und auf *Ranking after Aligment* gelegt, so dass experimentelle Untersuchungen mit kleinen Stichprobengrößen (*N = small*) durchgeführt werden können, wie sie von neueren Ansätzen zum *small-N*-Paradigma durch die Betonung der Selektion von Einzelfällen gefordert werden. Abhängige Variablen zur Messungen des Lernzuwachses brauchen zudem nur ordinalskaliert sein.

Für die recherchierten Verfahren wurden neben exakten Quellenangaben für das Auffinden der Teststatistiken in der neueren, statistischen Literatur die entsprechenden Hinweise zur Verfügung gestellt.

Abschließend sei darauf hingewiesen, dass es über die vorgestellten Versuchspläne hinaus weitere Möglichkeiten gibt, einzelne Schulklassen ungeteilt einzubeziehen, etwa mit der wichtigen Gruppe der hierarchischen faktoriellen Versuchspläne (vgl. Kirk, 2012). Sie eignen sich besonders gut, komplexe unterrichtsrelevante Fragestellungen ökonomisch behandeln zu können, wenn differenzierte Forschungshypothesen vorliegen. Nachfolgende Arbeiten sollten solche Versuchspläne wie auch Lateinische Quadrate und Griechisch-Lateinische Quadrate berücksichtigen.

8.6 Literatur

APA. American Psychological Association. (2010). *Presenting your findings*. Washington. APA.

Blatter, J., & Haverland, M. (2012). *Designing case studies: Explanatory approaches in small-N research*. New York: Palgrave Macmillan.

Bortz, J., & Schuster, C. (2010). *Statistik*. Berlin: Springer.

Bortz, J., Lienert, G. A., & Boehnke K. (2008). *Verteilungsfreie Methoden in der Biostatistik.* Berlin: Springer.

Brunner , F., Domhof, S., & Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial experiments.* New York: Springer.

Brunner, E., & Dette, H. (1992). Rank procedures for the two factor mixed model. *Journal of the American Statistical Association, 87*, 884–888.

Brunner, E., & Langer F. (1999). *Nichtparametrische Analyse longitudinaler Daten*. München: Oldenbourg.

Brunner, E., & Munzel, U. (2013). *Nichtparametrische Datenanalyse* (2. Auflage). Berlin: Springer.

Brunner, E., Dette, H., & Munk, A. (1997). Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association, 92*, 1494-1502.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Collier, D., Brady, H. E., & Seawright, J. (2004). Sources of leverage in causal inference: Toward an alternative view of methodology. In H. E. Brady, & D. Collier (Hrsg.), *Rethinking social inquiry* (S. 229–266). Lanham: Rowman & Littlefield.

Conover, W. J. (1999). *Practical nonparametric statistics*. New York: Wiley.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistics, 35*(3), 124–129.

Cooper, M. (1988). Nonparametric statistics. In J. P. Keeves (Hrsg.), *Educational research, methodology, and measurement: An international handbook* (S. 705–710). Oxford: Pergamon Press.

Field, A. (2013*). Discovering statistics using SPSS*. London: Sage.

Field, A., & Miles, J. (2010). *Discovering statistics using S.A.S.* London: Sage.

Field, A., & Miles, J. (2012. *Discovering statistics using R.* London: Sage.

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. Cambridge: MIT Press.

Gerig, T. (1969). A multivariate extension of Friedman´s χ_r^2-Test. *Journal of the American Statistical Association, 64*, 1595–1608.

Gerring, J. (2007). *Case study research*. Cambridge: University Press.

Gibbons, J. D., & Chakraborti, S. (2003). *Nonparametric statistical inference*. New York: Dekker.

Glaser, R. E. (2006). Levene´s robust test of homogeneity of variance. In S. Kotz (Hrsg.), *Encyclopedia of statistical sciences,* Vol. 6 (S. 4132–4134). New York: Wiley.

Goodwin, L. D., & Goodwin, W. L. (1985). Statistical techniques in AERJ articles, 1979–1983: The preparation of graduate students to read the educational research literature. *Educational Researcher, 14*(2), 5–11.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. Malabar, Fl: Krieger.

Hildebrand, H. (1980). Asymptotisch verteilungsfreie Rangtests in multivariaten linearen Modellen. In W. Köpcke & K. Überla (Hrsg.), *Biometrie – heute und morgen* (S. 344–349). Berlin: Springer.

Hox, J. J. (2000). Multilevel analysis of grouped and longitudinal data. In T. D. Little, K. U. Schnabel, & J. Baumert (Hrsg.), *Modeling longitudinal and multilevel data* (S. 15–32). London: Lawrence Erlbaum Associates.

Kabacoff, R. L. (2015). *R in action*. Shelter Island, NY: Manning.

Katz, B. M., & McSweeney, M. (1980). A multivariate Kruskal-Wallis test with post hoc procedures. *Multivariate Behavioral Research, 15*, 281–297.

Kay, M. (2018). Package *'ARTool'.* Retrieved January 2, 2018, from https://cran.r-project.org/web/packages/ARTool/ARTool.pdf

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350–386.

Kirk, R. E. (2012). *Experimental design* (4. Auflage). Pacific Grove, CA: Brooks/Cole Publishing Company.

Krantz, D. H., Luce, R.D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1). New York: Academic Press.

Kvam, P. H., & Vidakovic, B. (2007). *Nonparametric statistics with applications to science and engineering*. New York: Wiley.

Leeuw, de J., & Meijer, E. (2008). *Handbook of multilevel analysis* (S. 207–236). Berlin. Springer.

Lewis, D. G. (1968). *Experimental design in education*. London: University of London Press.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York. Wiley.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*, 579–620.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. Probleme und Lösungen. *Psychologische Rundschau, 58*(2), 103–117.

Narens, L. (2014). *Introduction to the theories of measurement and meaningfulness and the use of symmetry in science*. New York: Chapman & Hall.

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software, 50*(12), 1–23.

Puri, M. L., & Sen, P. K. (1993). *Nonparametric methods in multivariate analysis* (2. Auflage). New York: Wiley.

Ragin, C. C. (2004). Turning the tables: How case-oriented research challenges variable-oriented research. In H. E. Brady and D. Collier (Hrsg.), *Rethinking social inquiry* (S. 123–138). Lanham: Rowman & Littlefield.

Raudenbush, S. W. (2008). Many small groups. In J. de Leeuw & E. Meijer (Hrsg.), *Handbook of multilevel analysis* (S. 207–236). Berlin. Springer.

Roberts, F. S. (1979). *Measurement theory with applications to decision making, utility, and the social sciences.* London: Addison Wesley.

Royston, J. P. (2006). Shapiro-Wilk W statistics. In S. Kotz (Hrsg.), *Encyclopedia of statistical sciences*, Vol. 12 (S. 7679–7680). New York: Wiley.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–590.

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics, 14*(3), 255–267.

Sheskin, D. (2011). *Handbook of parametric and nonparametric statistical procedures* (5. Auflage). Boca Raton: CRC Press.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics*. New York: McGraw-Hill.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics*. Harlow: Pearson.

Thompson, G. L. (1991). A unified approach to rank tests for multivariate and repeated measure designs. *Journal of the American Statistical Association, 86*, 410–419.

Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, Fl: CRC Press.

Van Buuren, S., Groothuis-Oudshoorn, K. (2011). `mice`: Multivariate **i**mputation by **c**hained equations in `R`. *Journal of Statistical Software*, *45*(3), 1–67.

West, C. K., Carmody, C., & Stallings, W. M. (1983). The quality of research articles. *Journal of Educational Researcher, 25*, 26–30.

Winer, B. J. (1971). *Statistical principles in experimental design*. New York: Wiley.

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). *The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures*. CHI 2011, Session: Research.

Zendler, A., & Pfeiffer, T. (2009). Methodische Befunde zu durchgeführten Unterrichtsexperimenten. *Empirische Pädagogik, 23*(2), 208–221.