Kapitel 7

Berechnung von Stichprobengrößen für die Unterrichtsforschung

Andreas Zendler

7.1 Einleitung

Der Stichprobenumfang spielt die zentrale Rolle bei der Planung experimenteller Untersuchungen, deren Ziel die Prüfung von Ursache-Wirkungszusammenhängen ist. Vom Stichprobenumfang sind abhängig: (1) Die Genauigkeit und die Richtigkeit der Ergebnisse, (2) die Kosten für die Identifizierung und die Messung von Individuen in einer Stichprobe, sowie (3) die verfügbaren Ressourcen wie Zeit und Geld (vgl. Lindley, 2006).

Der Stichprobenumfang in einer experimentellen Untersuchung ist abhängig vom (1) eingesetzten statistischen Hypothesentest und (2) besonders vom verwendeten Versuchsplan und den formulierten Hypothesen mit Angabe erwarteter Unterschiede zwischen Gruppen (Effektgrößen, engl. *effect sizes* – vgl. Bausell & Li, 2002).

In der Literatur wird der Stichprobenumfang oft im Zusammenhang mit der Anwendbarkeit statistischer Hypothesentests diskutiert, z. B. wann parametrische Verfahren (*t*-Tests, Varianzanalysen) gegenüber nicht-parametrischen Verfahren (*U*-Test, Wilcoxon-Test, Rangvarianzanalysen) anzuwenden seien (Siegel, 1956; Wasserman, 2006; Corder & Foreman, 2009; Hollander, Wolfe, & Chicken, 2014), ob nicht-parametrische Verfahren bei kleinem Stichprobenumfang eher zu verwenden seien als parametrische Verfahren (Hettmansperger, 1984; Cooper, 1988; Huck, 2009; Sheskin, 2011), oder ob nicht-parametrische Verfahren eine höhere Power hätten als parametrische Verfahren, wenn deren Voraussetzungen nicht gegeben sind (Van Hecke, 2012).

Entscheidender als der Zusammenhang zwischen Stichprobenumfang und Wahl eines statistischen Hypothesentests ist allerdings der Zusammenhang zwischen Stichprobenumfang und verwendetem Versuchsplan. Aberson (2010) sowie Bausell und Li (2002) behandeln dies ausführlich, und die Empfehlungen von APA (2010) zielen darauf ab:

„Along with the description of subjects, give the intended size of the sample and number of individuals meant to be in each condition." (APA, 2010, S. 30)

Leider sind diese Empfehlungen in der Praxis experimenteller Unterrichtsforschung noch nicht angenommen worden, wie Peng, Long und Adaci (2012) zeigen: Nur in 1.77% der von 2005 bis 2010 veröffentlichten 1357 Artikel in 12 (*refereed*) Journals[1] zur Unterrichtsforschung wurde bei der Planung der Untersuchungen der notwendige Stichprobenumfang berechnet.

Die Befunde von Peng, Long und Adaci (2012) können darin begründet sein, dass in experimentellen Untersuchungen zur Unterrichtsforschung die statistische Signifikanz, die *nach* Datenerhebung bestimmt wird, gegenüber der statistischen Power, die *vor* Datenerhebung bei der Planung einer experimentellen Untersuchung zu berechnen ist, (immer noch) überbewertet wird: „Said another way, statistical significance is used to ascertain whether or not a given effect size can be interpreted as being reliable enough to allow the scientific community to accept a hypothesis once a study is *completed*. Statistical power, in contrast, is used to ascertain how likely a study's data are to result in statistical significance *before t*he study is begun. It is, in effect, a hypothetical or projected test of statistical significance conducted before an investigator has access to data" (Bausell, & Li, 2002, S. 1, Hervorhebungen im Original).

Ein anderer Grund für die Befunde von Peng, Long und Adie (2012) mag darin liegen, dass Stichprobenumfänge für konkrete Versuchspläne in der Literatur nicht verfügbar oder schwer zugänglich waren.

Mit den heute zur Verfügung stehenden Softwarepaketen ist es indes leicht möglich, den Stichprobenumfang in Bezug zu unterschiedlichsten Versuchsplänen und statistischen Verfahren in der Planungsphase einer experimentellen Untersuchung zu berechnen, um zu gewährleisten, dass wissenschaftliche Resultate erzielt werden können, die sich durch richtige Entscheidungen auszeichnen, was die Annahme oder Nichtannahme von Forschungshypothesen betrifft. Dieses Kapitel liefert einen Beitrag zur rich-

[1] *American Educational Research Journal*, *Educational Researcher*, *Journal of Counseling Psychology*, *Journal of Educational Psychology*, *Journal for Research in Mathematics Education*, *Journal of Research in Science Teaching*, *Journal of Research on Technology in Education*, *Journal of Special Education, Journal of School Psychology*, *The Modern Language Journal*, *Research in Higher Education*, *Theory and Research in Social Education.*

tigen Planung von experimentellen Untersuchungen in der Unterrichtsforschung, indem für wichtige Versuchspläne (vgl. Zendler, 2016) entsprechende Stichprobenumfänge auf Grundlage der Poweranalyse berechnet und empfohlen werden.

Der nächste Abschnitt befasst sich mit den Grundlagen poweranalytischer Versuchsplanung. Der dann folgende Abschnitt zeigt die Berechnung von Poweranalysen mit dem Softwarepaket PASS (Version 16). Der vierte Abschnitt enthält die Berechnungen zum Stichprobenumfang wichtiger Versuchspläne für die experimentelle Unterrichtsforschung mit konkreten Empfehlungen. Der fünfte Abschnitt enthält Schlussfolgerungen mit strategischen Hinweisen zur Erhöhung der statistischen Power in der experimentellen Versuchsplanung.

7.2 Poweranalytische Versuchsplanung

Statistische Signifikanz (auch bekannt als *p*-value) informiert, wie groß die Wahrscheinlichkeit ist, dass die Nullhypothese (H_0) falsch ist. Die statistische Signifikanz wird *nach Datenerhebung* berechnet. Falls der *p*-value kleiner ist als ein vorher festgelegtes Signifikanzniveau (z. B. α = .05), wird die H_0 verworfen. Führt eine Untersuchung zu einer so abgesicherten Entscheidung, dann bedeutet dies, dass unter H_0 die Wahrscheinlichkeit für eine korrekte Entscheidung .95 ist.

Statistische Power wird *vor Datenerhebung* bei der Planung einer experimentellen Untersuchung berechnet. Sie stellt den Zusammenhang her zwischen notwendiger Stichprobengröße auf der einen Seite und statistischer Power, Effektgröße Δ sowie Signifikanzniveau α auf der anderen Seite. Die statistische Power gibt die Wahrscheinlichkeit für eine korrekte Entscheidung unter der statistischen Alternativhypothese (H_1) an.

Die Poweranalyse geht auf Cohen (1977, 1988) zurück. Monographien liegen etwa vor von: Kraemer und Thiemann (1987), Lipsey (1990), Bausell und Li (2002), Aberson (2010), Murphy, Myors und Wolach (2009, 2014), Yuan und Zhang (2018).

Bradley, Russel und Reeve (1996) sowie Ryan (2013) informieren über die Poweranalyse bezüglich komplexer Versuchspläne. Murphy, Myors und Wolach (2009, 2014) berichten über die Anwendung der Poweranalyse, wenn in die Versuchsplanung mehrere Variablen (multivariate Versuchsplanung) einbezogen werden sollen.

7.2.1 Ausgangspunkte der Poweranalyse

Das Konstruktionsprinzip eines statistischen Hypothesentests stellt sicher, dass eine richtige Nullhypothese (H_0) bei einem signifikanten Ergebnis höchstens mit einer Wahrscheinlichkeit von α verworfen wird. Aufgrund der Stichprobenergebnisse können ne-

ben zwei korrekten zwei fehlerhafte Entscheidungen getroffen werden (siehe Abbildung 7.1): Eine an sich richtige H_0 wird aufgrund der Stichprobenergebnisse zugunsten der Alternativhypothese (H_1) verworfen (α-Fehler); die H_0 wird akzeptiert, obwohl die H_1 richtig ist (*ß*-Fehler).

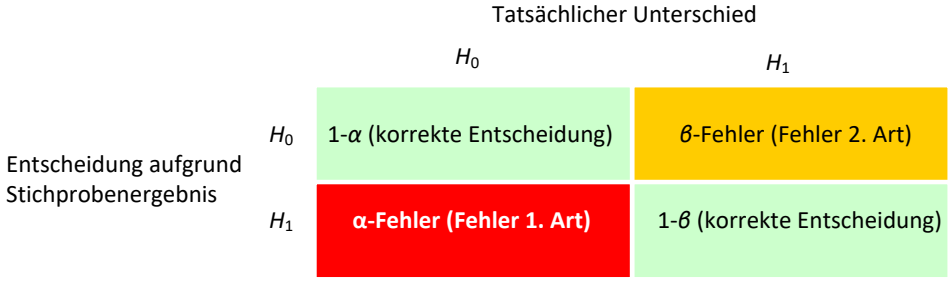

Abbildung 7.1 α-*Fehler* und *ß-Fehler* bei statistischen Entscheidungen

Welche Konsequenzen mit einem α-Fehler und einem *ß*-Fehler verbunden sind, sei am Beispiel der folgenden Forschungshypothese beleuchtet. Angenommen eine Unterrichtsmethode soll evaluiert und gegebenenfalls eingesetzt werden. Dazu ist die Forschungshypothese formuliert: „Mit der Unterrichtsmethode a_2 (z. B. Fallstudie) lassen sich im Vergleich mit der Unterrichtsmethode a_1 (z. B. direkte Instruktion) größere Lernerfolge erzielen“.

Es wird in der Forschungshypothese behauptet, dass sich mit der Unterrichtsmethode a_2 größere Lernerfolge erzielen lassen als mit der etablierten Unterrichtsmethode a_1. Dann lassen sich die folgenden statistischen Hypothesen formulieren:

H_0: Die Unterrichtsmethoden a_2 und a_1 unterscheiden sich nicht bezüglich des Lernerfolgs, oder mit a_2 werden kleinere Lernerfolge erzielt als mit der Unterrichtsmethode a_1.

H_1: Mit der Unterrichtsmethode a_2 werden größere Lernerfolge erzielt als mit der Unterrichtsmethode a_1.

α-*Fehler (Fehler 1. Art)*: Die H_0 wird verworfen, obwohl sie richtig ist, das heißt, es wird fälschlicherweise angenommen, die Unterrichtsmethode a_2 sei besser als die etablierte Unterrichtsmethode a_1. Dies kann die Neuanschaffung von Unterrichtsmaterialien, die Schulung von Lehrern, den Kauf neuer Geräte, usw. zur Folge haben – Maßnahmen, die angesichts der falschen Entscheidung nicht zu rechtfertigen sind.

ß-Fehler (Fehler 2. Art): Die H_1 wird verworfen, obwohl sie richtig ist, das heißt, es wird fälschlicherweise angenommen, dass sich die Unterrichtsmethode a_2 nicht von der Unterrichtsmethode a_1 unterscheidet. Die Folge hiervon wird sein, dass Unterricht weiterhin nach der etablierten Unterrichtsmethode a_1 durchgeführt wird. Es werden

zwar keine Fehlinvestitionen riskiert, aber die Chance verpasst, größere Lernerfolge zu erzielen.

Das Beispiel soll genügen, um zu zeigen, dass je nach Art der Fragestellung entweder der α-*Fehler* oder der *ß-Fehler* zu ungünstigen Konsequenzen führen kann.

7.2.2 Parameter der Poweranalyse

α-*Fehler* und *ß*-Fehler verändern sich in Abhängigkeit des Stichprobenergebnisses $\bar{x}$ (siehe Abbildung 7.2): Mit größer werdendem $\bar{x}$ sinkt die Wahrscheinlichkeit α, bei einer Entscheidung zugunsten der H_1 einen α-*Fehler* zu begehen. Gleichzeitig steigt die Wahrscheinlichkeit *ß* für einen *ß*-Fehler, das heißt, Entscheidungen zugunsten der H_0 werden mit größer werdendem $\bar{x}$ zunehmend unsicherer. Umgekehrt sinkt bei kleiner werdendem $\bar{x}$ die Wahrscheinlichkeit *ß* für einen *ß*-Fehler, während die Wahrscheinlichkeit α einer fälschlichen Annahme der H_1 (α-*Fehler*) steigt. α-*Fehler* und *ß*-Fehler verändern sich gegenläufig. Die Konsequenz dieser gegenläufigen Beziehung ist offensichtlich: Je stärker man sich dagegen absichern will, eine an sich richtige H_0 zu verwerfen (α-*Fehler*), desto eher wird man tatsächlich vorhandene Unterschiede übersehen (*ß*-Fehler).

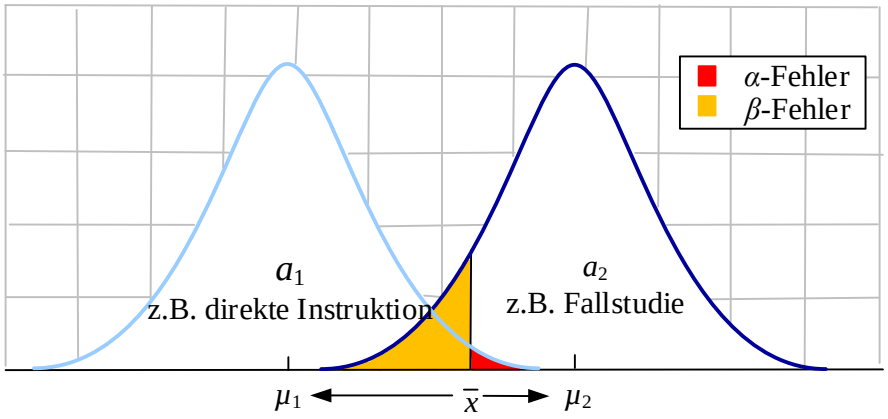

Abbildung 7.2 Schematische Darstellung von α-*Fehler* und *ß-Fehler*

Für den Fehler 2. Art bezeichnet *ß* die Wahrscheinlichkeit, eine an sich richtige Alternativhypothese H_1 fälschlicherweise abzulehnen. Folglich erhält man mit 1–*ß* die Wahrscheinlichkeit, in einer Untersuchung eine richtige H_1 auch als solche zu erkennen. Diese Wahrscheinlichkeit 1–*ß* für die korrekte Entscheidung im Falle der Gültigkeit von H_1 wird statistische Power genannt.

Die statistische Power (1–*ß*) eines Tests hängt von den folgenden Parametern ab:

1. dem Signifikanzniveau α,

2. dem Stichprobenumfang *N*,

3. der Effektgröße Δ.

Signifikanzniveau α

Das Signifikanzniveau α, das für einen statistischen Hypothesentest festgelegt wird, beeinflusst die statistische Power (1–*ß*) (siehe Abbildung 7.3).

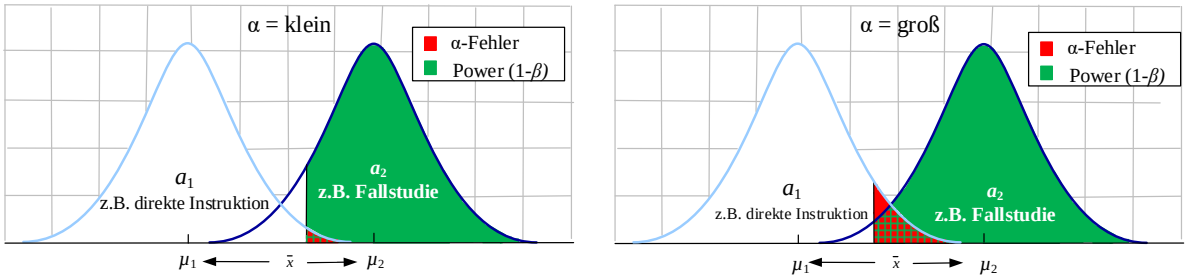

Abbildung 7.3 Einfluss des Signifikanzniveaus α auf die statistische Power (1–*ß*)

Ein großes α (z. B. α = .10) macht es einfacher, ein statistisch signifikantes Ergebnis zu erzielen als ein kleines α (z. B. α = .001). Dementsprechend gilt: Je größer das Signifikanzniveau α gesetzt wird, desto größer ist die statistische Power.

Stichprobenumfang *N*

Der Stichprobenumfang *N* hat einen Einfluss auf die Streuung der Stichprobenverteilung (vgl. Abbildung 7.4). Die Streuung sinkt mit wachsendem Stichprobenumfang, und sie ist fast vernachlässigbar für einen Stichprobenumfang mit sehr großem *N* (z. B. $N > 1000$). Damit wird ein allgemein plausibler Befund untermauert: Je größer der Stichprobenumfang, desto geringer ist die Wahrscheinlichkeit ß, in der statistischen Entscheidung einen Fehler 2. Art zu begehen. Infolgedessen gilt: Je größer *N* für den Stichprobenumfang ist, desto größer ist die statistische Power (1–*ß*).

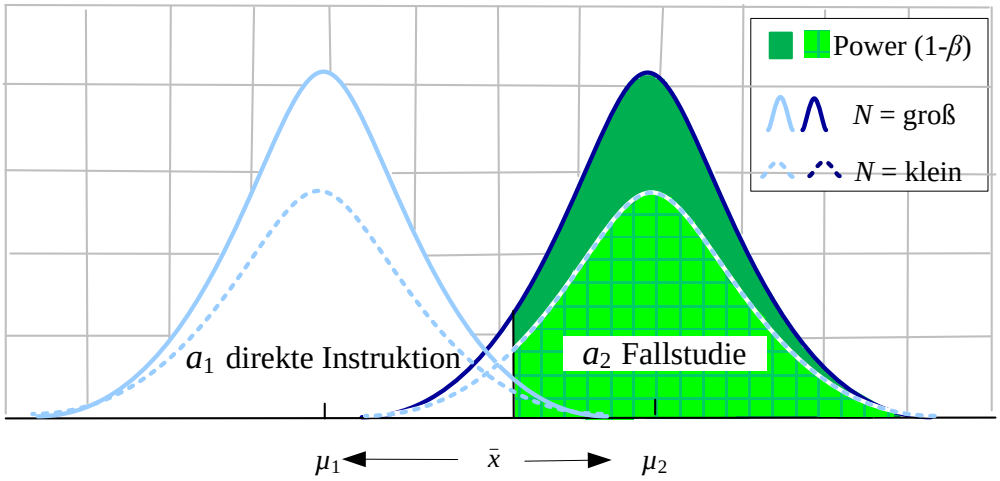

Abbildung 7.4 Einfluss des Stichprobenumfangs *N* auf die statistische Power (1–*ß*)

Effektgröße Δ

Außer vom Signifikanzniveau α und vom Stichprobenumfang *N* hängt die statistische Power vom Grad des tatsächlichen Unterschieds ab, von der Effektgröße (*effect size; ES*) $\Delta = \mu_2 - \mu_1$. Abbildung 7.5 zeigt: Je größer die Effektgröße ist, desto kleiner ist die Wahrscheinlichkeit *ß* für einen *ß*-Fehler und desto größer ist die statistische Power. Für die Effektgröße schlägt Cohen (1977, 1988) die folgenden Richtwerte vor: Δ = .20 bedeutet einen „kleinen" Effekt, Δ = .50 bedeutet einen „mittleren" Effekt und Δ = .80 bedeutet einen „großen" Effekt.

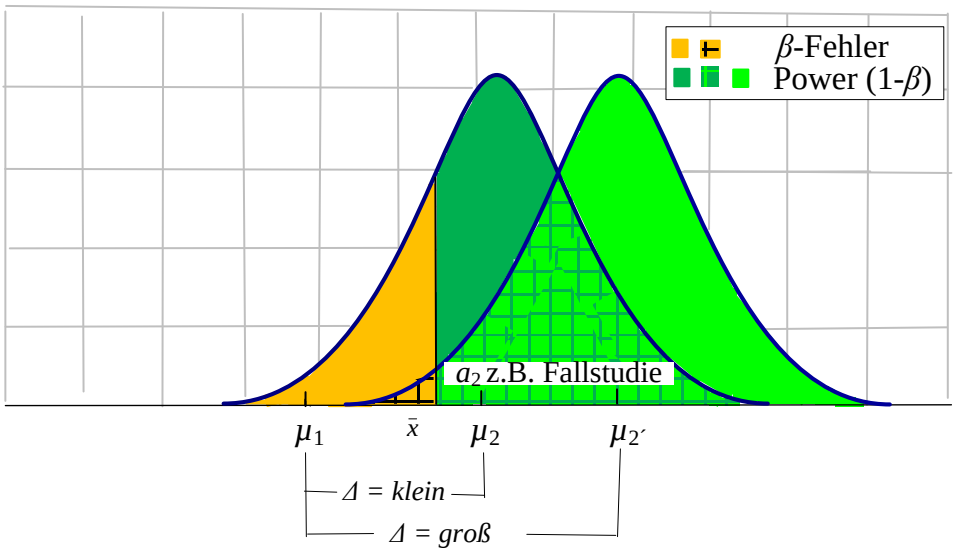

Abbildung 7.5 Einfluss der Effektgröße Δ auf die statistische Power (1–*ß*)

Konventionelle Werte für α, 1–α, *ß*, 1–*ß* und Δ

Zur Überprüfung statistischer Hypothesen wird in den meisten wissenschaftlichen Untersuchungen per Konvention $\alpha = .05$ gesetzt. Für den *ß-Fehler* gibt es keine ähnliche Übereinkunft: „There is no corresponding convention for beta (probability of Type II error), which may be one of the reasons it is so widely neglected in (...) science research“ (Lipsey, 1990, S. 40).

α. Der Wert $\alpha = .05$ bedeutet: Wahrscheinlichkeit, eine an sich richtige H_0 abzulehnen, beträgt 5%.

$1-\alpha$. Der Wert $1-\alpha = .95$ bedeutet: Wahrscheinlichkeit einer korrekten Entscheidung unter H_0 beträgt 95%.

ß. Der Wert *ß* = .20 bedeutet: Wahrscheinlichkeit, eine an sich richtige H_1 abzulehnen, beträgt 20%.

1–*ß*. Der Wert 1–*ß* = .80 bedeutet: Wahrscheinlichkeit einer korrekten Entscheidung unter H_1 beträgt 80%.

Δ. Der Wert $\Delta = .50$ (mittlerer Effekt) bedeutet: Unterschied zwischen $\mu_2 - \mu_1$, relativiert an der Standardabweichung, beträgt .50.

Semantik einer Poweranalyse

Eine Poweranalyse, die für eine experimentelle Untersuchung angibt, dass der Stichprobenumfang $N = 64$ (je Gruppe) betragen muss, hat eine Chance von 80% – das heißt die Power $(1-ß) = .80$ –, um einen Effekt von $\Delta = .50$ (mittlerer Effekt) zwischen Kontroll- und Experimentalgruppe bei gegebenem $\alpha = .05$ aufzudecken.

Anders ausgedrückt: „In non-statistical language, what this is telling us is that*, if everything goes as planned*, our investigator will have an 80% chance of achieving statistical significance. In other words, if the experimental treatment is indeed capable of producing a gain of one-half of the dependent variable's standard deviation as compared to no treatment at all (i.e., if the hypothesized ES is accurate), *then eight out of ten properly performed experiments* using 64 subjects per group will result in statistical significance at the 0.05 level." (Bausell & Li, 2002, S. 10, Hervorhebung im Original)

7.2.3 Typen der Poweranalyse

Cohen (1977, 1988) unterscheidet die folgenden vier Typen von Poweranalysen:

1. Statistische Power $(1-ß)$ als Funktion von α, *N* und Δ;

2. Stichprobengröße *N* als Funktion von $(1-ß)$, Δ und α;

3. Effektgröße Δ als Funktion von α, *N* und $(1-ß)$;

4. α-Fehler als Funktion von *N*, Δ und (1–*ß*).

Sind je drei Parameter bekannt, dann kann der vierte bestimmt werden. Der dritte und vierte Typ der Poweranalyse sind ungewöhnlich und ungebräuchlich; sie werden deshalb im Folgenden auch nicht weiter behandelt. Die ersten beiden Typen werden näher erläutert: Sie sollten fester Bestandteil der Versuchsplanung in der Unterrichtsforschung sein.

7.2.3.1 Allgemeines Vorgehen zur Berechnung von Poweranalysen

Für die Berechnung von Poweranalysen sind sogenannte Powercharts (und Powertabellen) ein nützliches Hilfsmittel (z. B. Cohen, 1997, 1988; Kraemer & Thiemann, 1987). Powercharts sind in Bezug auf die Effektgröße konstruiert. Abbildung 7.6 zeigt die statistische Power 1–*ß* als Funktion von Δ und *N* für gegebenes α = .05

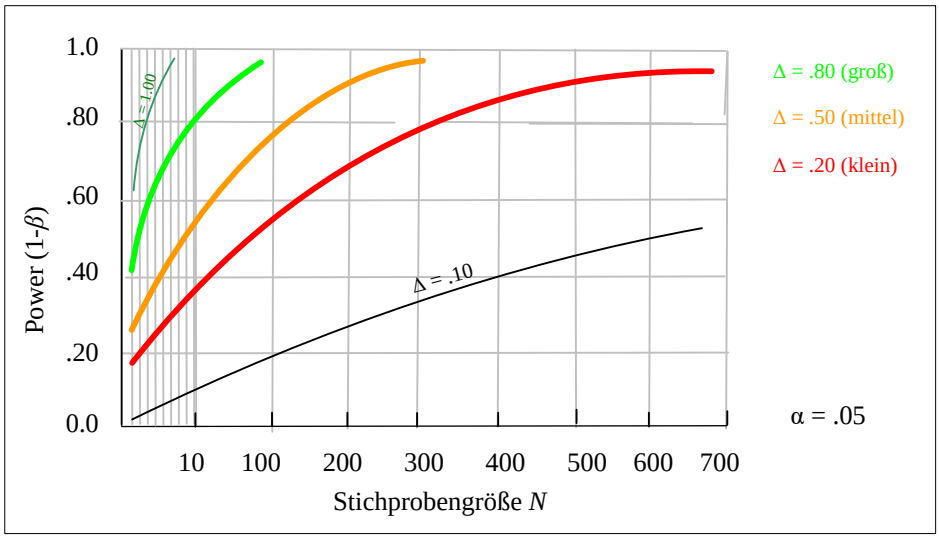

Abbildung 7.6 Power (1–*ß*) als Funktion von Δ und *N* für gegebenes α = .05

Aus Abbildung 7.6 können einige interessante Beziehungen entnommen werden: (1) Zwischen der statistischen Power 1–*ß* und der Stichprobengröße *N* besteht eine asymptotische Beziehung: Für kleines *N* bewirkt eine Erhöhung des Stichprobenumfangs eine starke Erhöhung der statistischen Power, für großes *N* hat die weitere Erhöhung des Stichprobenumfangs nur noch einen geringen Einfluss auf die statistische Power. (2) Bei gegebener Stichprobengröße *N* zwischen 150 und 200 erhöht sich die statistische Power extrem, wenn sich die Effektgröße Δ erhöht. (3) Der gemeinsame Einfluss von Effektgröße Δ und Stichprobengröße *N* auf die statistische Power st der wichtigste Zusammenhang. Falls man Effekte mit einer sehr hohen statistischen Power von mehr als .90 nachweisen möchte, ergeben sich große praktische Schwierigkeiten: Für Stichprobengrößen $N < 1000$ gelingt dieser Nachweis nur für große Effektgrößen ($\Delta > 1.00$); für kleine Effektgrößen benötigt man sehr große Stichproben (z. B. $N > 1000$ für Δ = .10).

7.2.3.2 Poweranalyse Typ I: Statistische Power (1–β) als Funktion von α, *N* und Δ

Bei diesem Typ der Poweranalyse wird folgendermaßen vorgegangen: Zunächst wird das Signifikanzniveau α festgelegt. Dann wird ermittelt, mit welcher Stichprobengröße *N* in einer experimentellen Untersuchung gerechnet werden kann. Für die Effektgröße Δ wird ein für die Untersuchung relevanter Effekt angenommen, der die Alternativhypothese determiniert.

Ist die aus den Powercharts bestimmte statistische Power (1–β) zur Aufdeckung relevanter Effekte zu klein, dann können vor der eigentlichen Versuchsdurchführung die Spezifikationen der Parameter geändert werden.

Insbesondere Änderungen des Versuchsplans können dazu führen, dass – über die Reduzierung der Fehlervarianz – die Effektgröße Δ in ihren Werten höher angesetzt werden kann. Beispiele für Versuchspläne, mit denen ein Teil der Fehlervarianz reduziert werden kann, sind etwa Split-Plot-Versuchspläne (SPF-*p*•*q*, SPF-*p*×*q*•*r*, siehe Abschnitt 7.4) oder allgemein Versuchspläne mit Messwiederholungen.

Abbildung 7.7 veranschaulicht die Schritte 1-4 bei der Durchführung von Poweranalysen des Typs I. Ausgehend von ① einem Signifikanzniveau α = .05, ② einem Stichprobenumfang von *N* = 10 und ③ einer mittleren Effektgröße Δ = .50 erhält man ④ eine statistische Power von ca. 0.52.

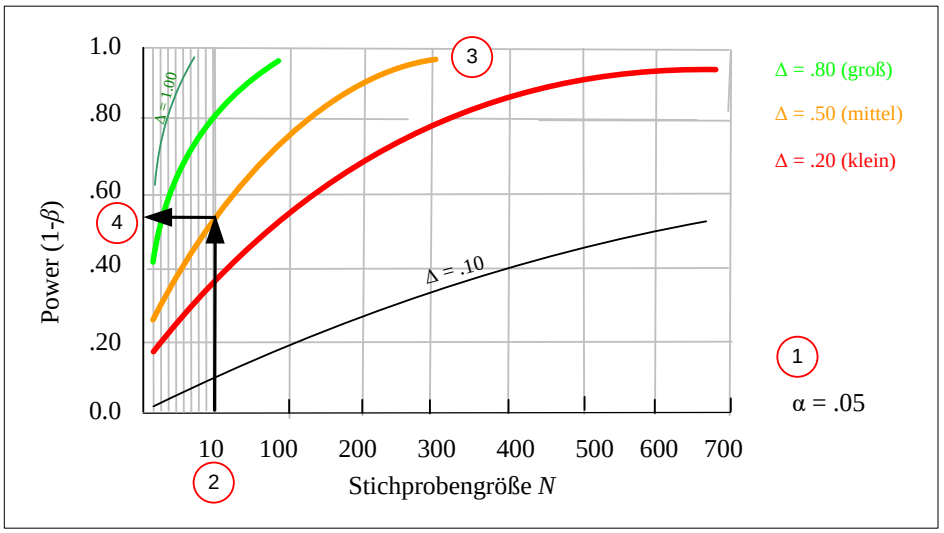

Abbildung 7.7 Poweranalyse vom Typ I

7.2.3.3 Poweranalyse Typ II: Stichprobengröße *N* als Funktion von statistischer Power (1–β), Δ und α

Mit diesem Typ der Poweranalyse lässt sich die Stichprobengröße *N* für einen Versuchsplan bestimmen, die notwendig ist, um mit der Wahrscheinlichkeit (1–*ß*) die erwünschte unterrichtsrelevante Effektgröße Δ bei festgelegtem Signifikanzniveau α nachweisen zu können. Dieser Typ ist die logische Umkehrung der Poweranalyse des Typs I und sollte fester Bestandteil der Versuchsplanung in der experimentellen Unterrichtsforschung sein.

Abbildung 7.8 zeigt die Schritte bei der Durchführung von Poweranalysen des Typs II: Für ① eine gewünschte statistische Power (1–*ß*) von .80, bei ② gegebenem Signifikanzniveau α = .05 und ③ einer mittleren Effektgröße Δ = .50 benötigt man ④ einen Stichproben-umfang von *N* = 120.

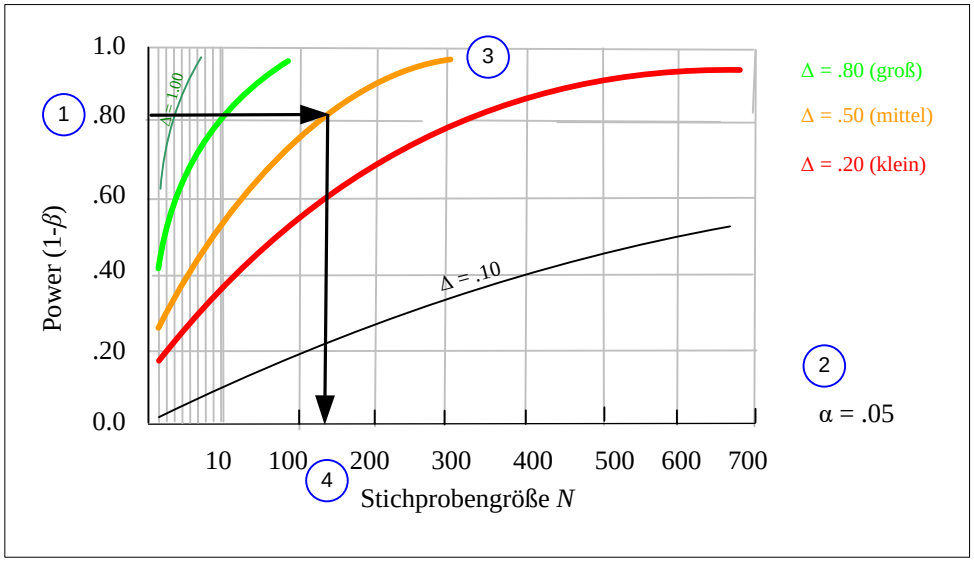

Abbildung 7.8 Poweranalyse vom Typ II

7.3 Poweranalysen mit PASS (Version 16)

Für die Durchführung von Poweranalysen steht eine Vielzahl von Programmpaketen zur Verfügung, z. B. G*Power, SPSS/Sample Power, PASS. In der Bewertung des Funktionsumfangs schneidet PASS (Power Analysis and Sample Size Software) am besten ab (Peng, Long, & Abaci, 2012). Deshalb wird dieses Programmpaket auch im Folgenden für die Berechnung der Stichprobenumfänge im Hinblick auf Versuchspläne des Typs RB-*p,* CR-*p,* CRF-*p*×*q*, SPF-*p*•*q* und SPF-*p*×*r*•*q* in der aktuellen Version 16 (NCSS, 2018) verwendet.

Das Programmpaket zeichnet sich durch ein breites Spektrum von Anwendungsmöglichkeiten aus. PASS erlaubt die Berechnung von Poweranalysen vom Typ I, II, III und IV. Zur Berechnung des Stichprobenumfangs für die ausgewählten Versuchspläne werden im vierten Abschnitt Poweranalysen vom Typ II berechnet.

PASS unterstützt eine große Anzahl statistischer Verfahren, die in 19 Kategorien zusammengefasst sind, z. B. *Correlation*, *Means, Nonparametric*. Abbildung 7.9 zeigt das *PASS Home Window* mit den unterschiedlichen Kategorien statistischer Verfahren, aus denen eine *Procedure* auszuwählen ist.

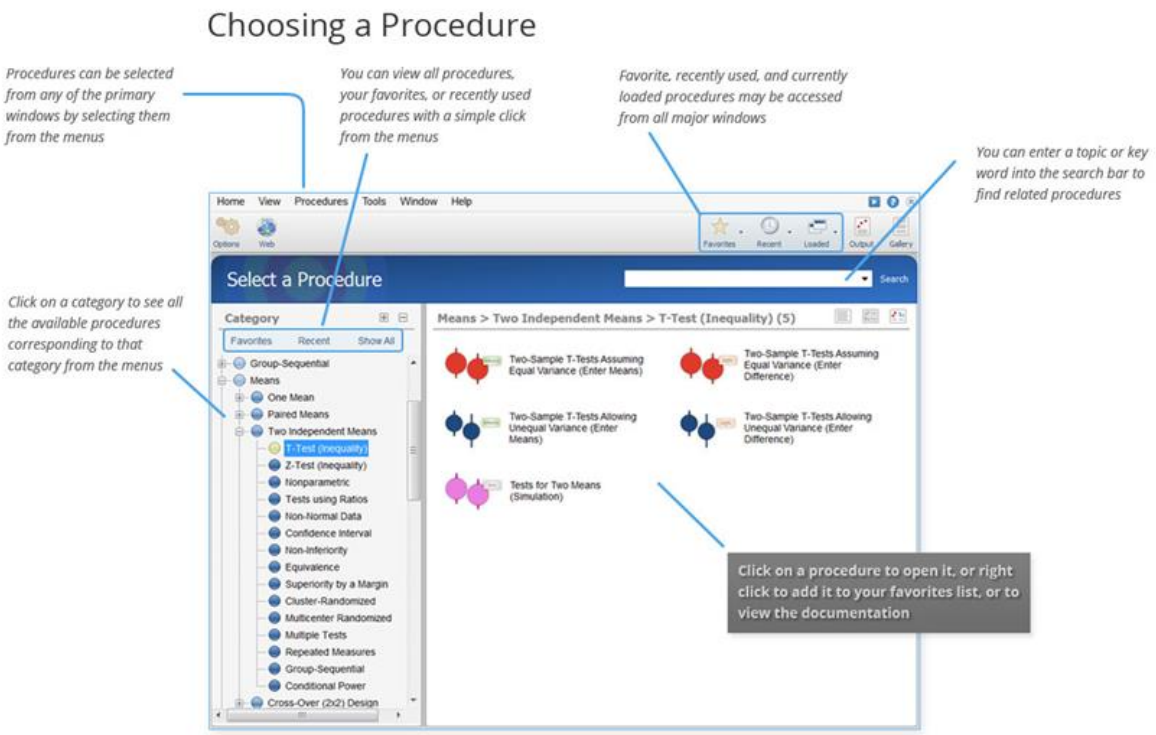

Abbildung 7.9 PASS Home Window (Mit freundlicher Genehmigung von © NCSS LLC, 2018, All Rights Reserved)

Nach Auswahl der *Procedure* werden im *PASS Procedure Window* die Parameter zur Durchführung der Poweranalyse spezifiziert, das heißt, es folgen Eingaben für die Effektgröße Δ, das Signifikanzniveau α und die gewünschte statistische Power $(1-\beta)$.

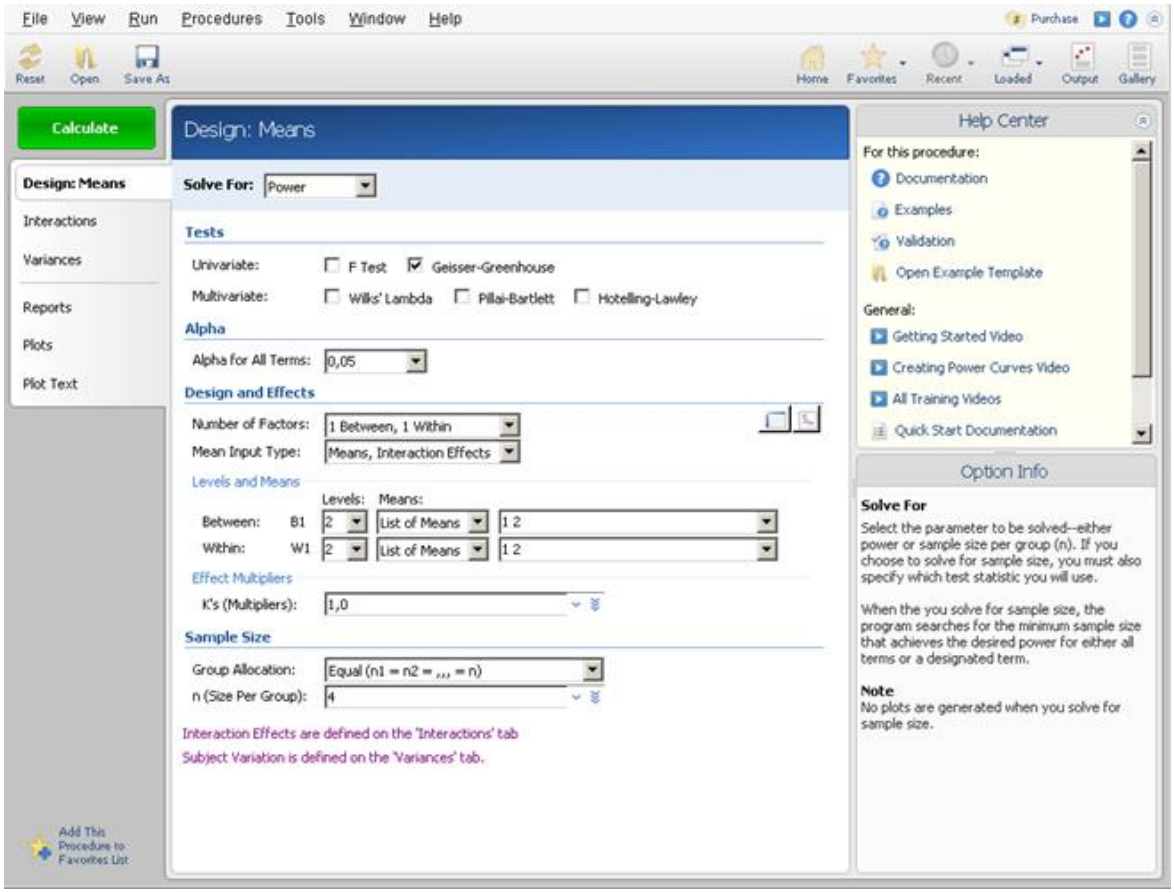

Abbildung 7.10 PASS Procedure Window (Mit freundlicher Genehmigung von © NCSS LLC, 2018, All Rights Reserved)

Die Ergebnisse von Poweranalysen werden im *PASS Output Window* dargestellt, und zwar in Form von Powertabellen (Abbildung 7.11) und Powercharts (Abbildung 7.12). Powertabellen enthalten die numerischen Ergebnisse der Poweranalyse. Abbildung 7.11 zeigt ein typisches Ergebnis im *PASS Output Window* mit Angaben der Stichprobenumfänge für einen *t*-Test bezüglich Effektgröße Δ, statistischer Power $(1-\beta)$ und gegebenem Signifikanzniveau α. Überdies wird auf Referenzen hingewiesen, welche Grundlagen für die Programmierung der entsprechenden *Procedure* waren.

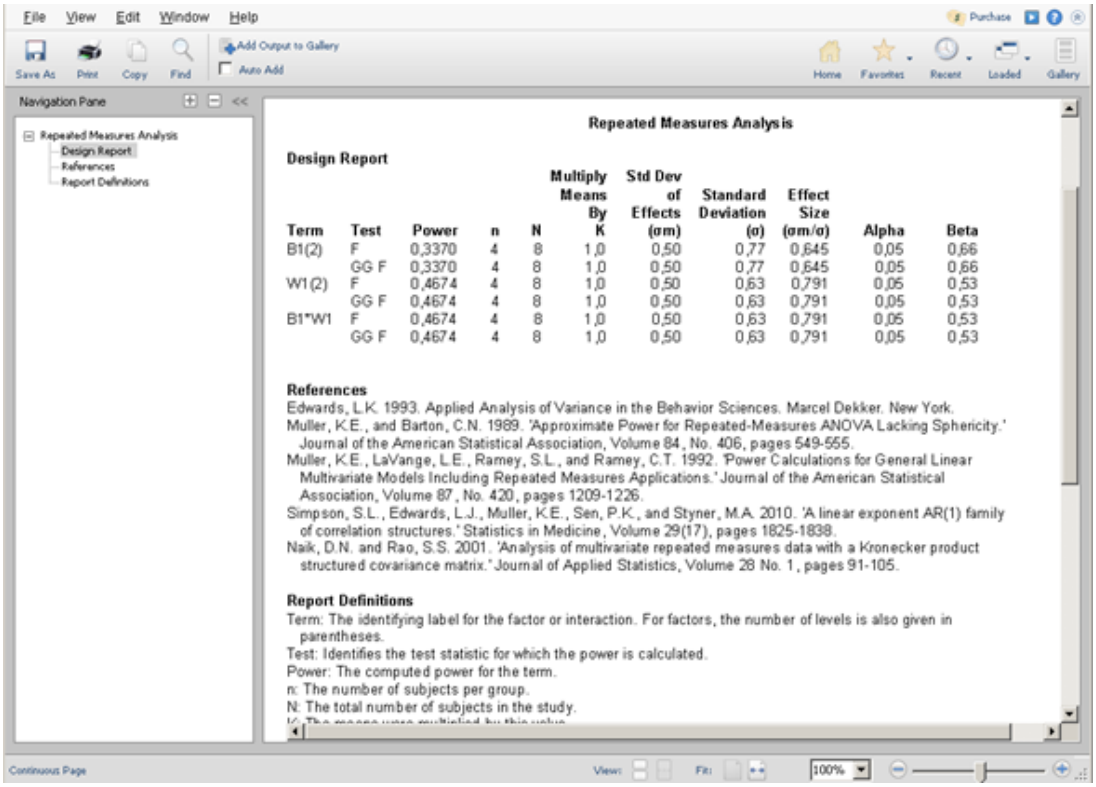

Abbildung 7.11 PASS Output Window mit Tabelle für die Poweranalyse (Mit freundlicher Genehmigung von © NCSS LLC, 2018, All Rights Reserved)

Powercharts visualisieren die Ergebnisse der Poweranalyse. Abbildung 7.12 zeigt für einen *t*-Test die statistische Power in Abhängigkeit der Stichprobengröße bei festgelegter Effektgröße Δ und gegebenem Signifikanzniveau α.

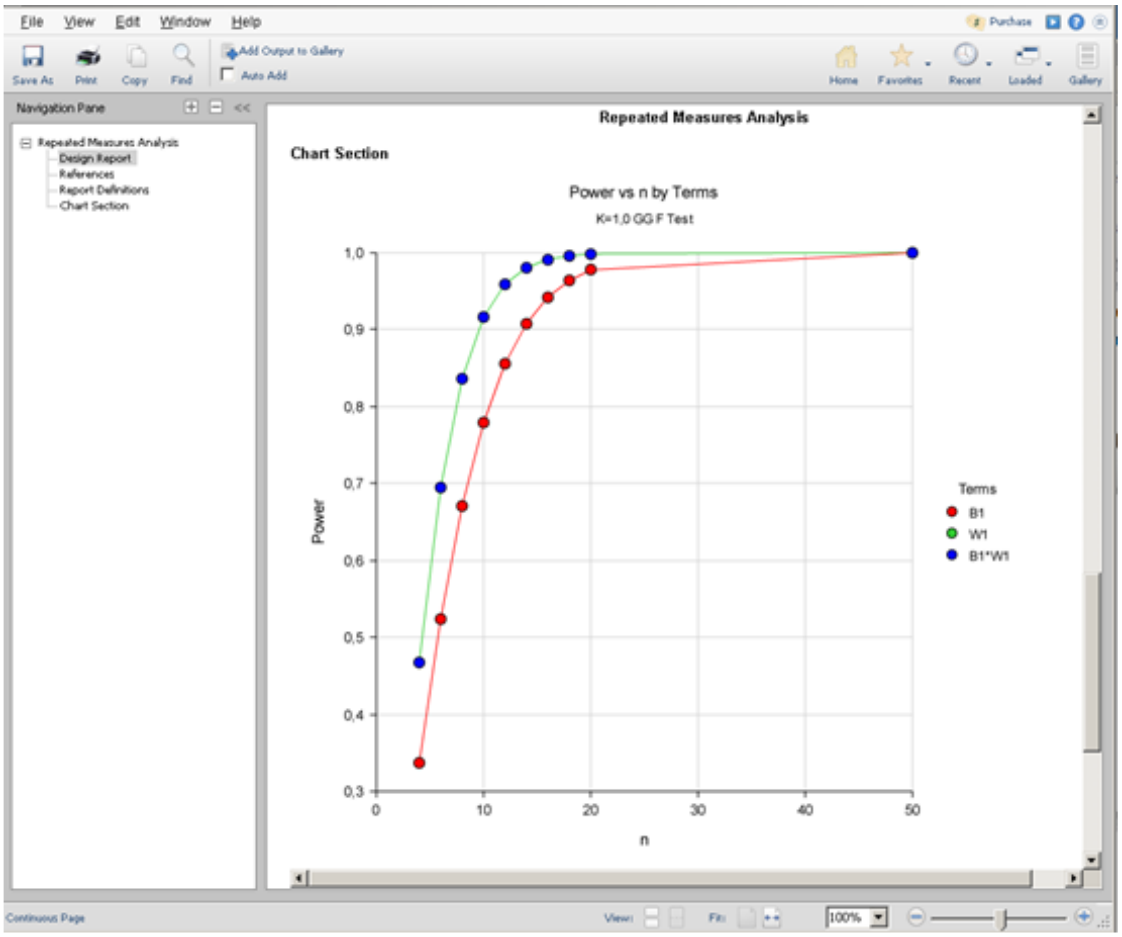

Abbildung 7.12 PASS Output Window mit Powerchart (Mit freundlicher Genehmigung von © NCSS LLC, 2018, All Rights Reserved)

7.4 Stichprobenumfang für ausgewählte Versuchspläne

Dieser Abschnitt enthält die Poweranalysen vom Typ II für fünf Typen von Versuchsplänen. In der Nomenklatur von Kirk (1996, 2012) sind dies: RB-*p* (Randomized Block designs), CR-*p* Completey Randomized designs), CRF-*p*×*q* (Completey Randomized Factorial designs), SPF-*p*•*q* (Split-Plot designs) und SPF-*p*×*r*•*q* (Split-Plot Factorial designs). Für jeden Typ wird der Stichprobenumfang für jeweils drei konkrete Versuchspläne berechnet – bezüglich der *F*-Prüfgröße parametrischer Varianzanalysen (Modell mit festen Effekten).

Die ausgewählten 15 Versuchspläne besitzen für die experimentelle Unterrichtsforschung einen besonderen Stellenwert, besonders zur Untersuchung der Lernwirksamkeit unterschiedlicher Unterrichtsmethoden. Für den Einsatz der Versuchspläne in der experimentellen Unterrichtsforschung werden Hinweise gegeben, zumal für die Untersuchung der Lernwirksamkeit in Schulklassen.

1. RB-*p* Versuchspläne

Tabelle 7.1 enthält den Gesamtstichprobenumfang *N* für die drei RB-*p* Versuchspläne RB-2, RB-3 und RB-4 in Abhängigkeit vom Signifikanzniveau α = .05, den Effektgrößen Δ = .80 (groß), Δ = .50 (mittel), Δ = .20 (klein) und der statistischen Power 1–*ß* = .80, 1–*ß* = .95, 1–*ß* = .99.

Die berechneten Stichprobenumfänge zeigen, dass sich mit nur einer Schulklasse (durchschnittliche Schüleranzahl < 30) experimentelle Unterrichtsforschung (z. B. Behaltensleistungen bei Schülern) insbesondere mit den Versuchsplänen RB-2, RB-3 und RB-4 durchführen lässt, wenn große Effekte (Δ = .80) bei einer statistischen Power von (1–*ß*) = .80 oder (1–*ß*) = .95 erwartet werden. Der Nachweis kleiner Effekte (Δ = .20) mit RB-*p* Versuchsplänen ist im Klassenkontext nicht möglich.

Tabelle 7.1 macht deutlich, dass mit großen Stichprobenumfängen (*N* ≥ 200) geplant werden muss, wenn kleine Effekte (Δ = .20) mit den Versuchsplänen RB-2, RB-3 und RB-4 untersucht werden sollen.

Tabelle 7.1 Stichprobenumfang *N* für drei RB-*p* Versuchspläne

RB-2	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	14	24	30
Δ = .50	34	54	80
Δ = .20	200	330	460
RB-3	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	18	26	38
Δ = .50	42	64	86
Δ = .20	250	400	550
RB-4	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	19	29	40
Δ = .50	45	70	96
Δ = .20	280	440	610

2. CR-*p* Versuchspläne

Tabelle 7.2 enthält den Gesamtstichprobenumfang *N* und den durchschnittlichen Stichprobenumfang (*n*) je Faktorstufe für die drei CR-*p* Versuchspläne CR-2, CR-3 und CR-4 in Abhängigkeit vom Signifikanzniveau α = .05, den Effektgrößen Δ = .80 (groß), Δ = .50 (mittel), Δ = .20 (klein) und der statistischen Power 1–*ß* = .80, 1–*ß* = .95, 1–*ß* = .99.

Die berechneten Stichprobenumfänge zeigen, dass sich nur eine Schulklasse mit durchschnittlicher Schüleranzahl für die experimentelle Unterrichtsforschung unter einem Faktor (z. B. Unterrichtsmethoden) dann eignet, wenn mittlere (Δ = .50) bis große Effekte (Δ = .80) und einer statistischen Power von 1–*ß* = .80, 1–*ß* = .95, 1–*ß* = .99 erwartet werden. Interessant ist, dass die Versuchspläne CR-3 und CR-4 wegen der notwendigen, kleineren Stichprobenumfänge unter den Faktorstufen besser geeignet sind als der Versuchsplan CR-2. Mit Halbklassen kann in der Unterrichtsforschung der CR-2 Versuchsplan eingesetzt werden, wenn große Effekte erwartet werden.

Aus der Tabelle 7.2 lässt sich entnehmen, dass sehr große Stichprobenumfänge (132 ≤ *N* ≤ 600) notwendig sind, wenn kleine Effekte (Δ = .20) mit den Versuchsplänen CR-2, CR-3 und CR-4 nachgewiesen werden sollen.

Tabelle 7.2 Stichprobenumfang *N* und *(n)* für drei CR-*p* Versuchspläne

CR-2	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	14 (7)	22 (11)	30 (15)
Δ = .50	34 (17)	54 (27)	76 (38)
Δ = .20	200 (100)	320 (160)	460 (230)
CR-3	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	18 (6)	27 (9)	36 (12)
Δ = .50	42 (14)	66 (22)	90 (30)
Δ = .20	240 (80)	390 (130)	540 (180)
CR-4	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	24 (6)	32 (8)	40 (10)
Δ = .50	48 (12)	72 (18)	100 (25)
Δ = .20	280 (70)	432 (108)	600 (150)

3. CRF-*p*×*q* Versuchspläne

Tabelle 7.3 enthält den Gesamtstichprobenumfang *N* und den durchschnittlichen Stichprobenumfang (*n*) je Faktorstufe für die drei CRF-*p*×*q* Versuchspläne CRF-2×2, CRF-3×2 und CRF-5×2 in Abhängigkeit vom Signifikanzniveau α = .05, den Effektgrößen Δ = .80 (groß), Δ = .50 (mittel), Δ = .20 (klein) und der statistischen Power 1–*ß* = .80, 1–*ß* = .95, 1–*ß* = .99 von Faktor *A*. Für die Versuchspläne CRF-3×2 und CRF-5×2 ist die Stichprobengröße für Faktor *A* angegeben.

Tabelle 7.3 Stichprobenumfang *N* und *(n)* für drei CRF-*p*×*q* Versuchspläne

CRF-2×2	1–*ß* = .80 für *A*	1–*ß* = .95 für *A*	1–*ß* = .99 für *A*
Δ = .80	16 (4)	24 (6)	32 (8)
Δ = .50	36 (9)	56 (14)	76 (19)
Δ = .20	200 (50)	328 (82)	460 (115)
CRF-3×2	1–*ß* = .80 für *A*	1–*ß* = .95 für *A*	1–*ß* = .99 für A
Δ = .80	24 (4)	30 (5)	42 (7)
Δ = .50	42 (7)	66 (11)	90 (15)
Δ = .20	246 (41)	390 (65)	534 (89)
CRF-5×2	1–*ß* = .80 für *A*	1–*ß* = .95 für *A*	1–*ß* = .99 für *A*
Δ = .80	20 (2)	30 (3)	50 (5)
Δ = .50	40 (4)	60 (6)	80 (8)
Δ = .20	300 (30)	460 (46)	640 (64)

Die berechneten Stichprobenumfänge zeigen, dass man experimentelle Unterrichtsforschung unter zwei Faktoren *A* (z. B. Unterrichtsmethoden) und *B* (z. B. Schülermerkmal) dann mit nur einer Schulklasse durchführen kann, wenn mittlere (Δ = .50) bis große Effekte (Δ = .80) erwartet werden mit einer statistischen Power von 1–*ß* = .80, 1–*ß* = .95. Der Nachweis kleiner Effekte (Δ = .20) mit CRF-*p*×*q* Versuchsplänen ist im Klassenkontext nicht möglich.

Der CRF-2×2 und der CRF-3×2 Versuchsplan eignen sich gut zur Kontrolle des Klassenkontextes. Dann wird unter Faktor *A* mit 2 bzw. 3 Klassen geplant, in denen jeweils mit Halbklassen unter Faktor *B* z. B. die Lernwirksamkeit unter zwei Unterrichtsmethoden untersucht werden. Für beide Versuchspläne lassen sich mittlere Effekte (Δ = .50) unter einer statistischen Power von 1–*ß* = .95 untersuchen.

Tabelle 7.3 verdeutlicht, dass große Stichprobenumfänge verfügbar sein müssen, wenn kleine Effekte (Δ=.20) mit den Versuchsplänen CRF-2×2, CRF-3×2 und CRF-5×2 nachgewiesen werden sollen.

4. SPF-*p*•*q* Versuchspläne

Tabelle 7.4 enthält den Gesamtstichprobenumfang *N* und den durchschnittlichen Stichprobenumfang (*n*) je Faktorstufe für die drei SPF-*p*•*q* Versuchspläne SPF-2•2, SPF-2•3 und SPF-3•2 in Abhängigkeit vom Signifikanzniveau α = .05, den Effektgrößen

Δ = .80 (groß), Δ = .50 (mittel), Δ = .20 (klein) und der statistischen Power 1–*ß* = .80, 1–*ß* = .95, 1–*ß* = .99 von Faktor *A*.

Tabelle 7.4 Stichprobenumfang *N* und *(n)* für drei SPF-*p*•*q* Versuchspläne

SPF-2•2	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	16 (8)	22 (11)	32 (16)
Δ = .50	34 (17)	56 (28)	78 (39)
Δ = .20	190 (95)	320 (160)	450 (225)
SPF-2•3	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	16 (8)	24 (12)	32 (16)
Δ = .50	34 (17)	54 (27)	76 (38)
Δ = .20	196 (98)	320 (160)	452 (226)
SPF-3•2	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	21 (7)	30 (10)	39 (13)
Δ = .50	45 (15)	66 (22)	90 (30)
Δ = .20	258 (86)	408 (136)	564 (188)

Die berechneten Stichprobenumfänge zeigen, dass sich eine Schulklasse mit durchschnittlicher Schüleranzahl für die experimentelle Unterrichtsforschung unter den Faktoren *A* (z. B. Unterrichtsmethoden) und *B* (z. B. Behaltensleistungen) nur dann eignet, wenn große Effekte (Δ = .80) bei einer statistischen Power von 1–*ß* = .80, 1–*ß* = .95, 1–*ß* = .99 erwartet werden. Bemerkenswert ist, dass sich die Versuchspläne SPF-2•2und SPF-2•3 kaum unterscheiden, was die Stichprobenumfänge, die Effektgrößen und die statistische Power betreffen. Der Nachweis kleiner Effekte (Δ = .20) ist mit SPF-*p*•*q* Versuchsplänen im Klassenkontext nicht möglich.

Tabelle 7.4 zeigt, dass große Stichprobenumfänge (*N* ≥ 190) verfügbar sein müssen für die Versuchspläne SPF-2•2, SPF-2•3 und SPF-3•2, wenn kleine Effekte (Δ = .20) erwartet werden.

5. SPF-*p*×*q*•*r* Versuchspläne

Tabelle 7.5 enthält den Gesamtstichprobenumfang *N* und den durchschnittlichen Stichprobenumfang (*n*) je Faktorstufe für die drei SPF-*p*×*r*•*q* Versuchspläne SPF-2×2•2, SPF-2×2•3 und SPF-2×3•2 in Abhängigkeit vom Signifikanzniveau α = .05, den Effektgrößen Δ = .80 (groß), Δ = .50 (mittel), Δ = .20 (klein) und der statistischen Power 1–*ß*

= .80, 1–*ß* = .95, 1–*ß* = .99 von Faktor A. Für die Versuchspläne ist die Stichprobengröße für Faktor *A* angegeben.

Die berechneten Stichprobenumfänge zeigen, dass sich eine Schulklasse mit durchschnittlicher Schüleranzahl für die experimentelle Unterrichtsforschung unter den Faktoren *A* (z. B. Unterrichtsmethoden), *C* (z. B. Schülermerkmal) und *B* (z. B. Behaltensleistungen) nur dann eignet, wenn mittlere (Δ = .50) bis große Effekte (Δ = .80) bei statistischer Power 1–*ß* = .80, 1–*ß* = .95 und 1–*ß* = .99 erwartet werden. Bemerkenswert ist, dass sich die Versuchspläne SPF-2•2und SPF-2•3 kaum unterscheiden – im Hinblick auf Stichprobenumfänge, die Effektgrößen und die statistische Power. Der Nachweis kleiner Effekte (Δ = .20) mit SPF-*p*×*q*•*r* Versuchsplänen ist im Klassenkontext nicht möglich.

Tabelle 7.5 Stichprobenumfang *N* und *(n)* für drei SPF-*p*×*r*•*q* Versuchspläne

SPF-2×2•2	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	16 (4)	24 (6)	32 (8)
Δ = .50	36 (9)	56 (14)	80 (20)
Δ = .20	204 (51)	344 (86)	488 (122)
SPF-2×2•3	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	16 (4)	24 (6)	32 (8)
Δ = .50	36 (9)	56 (14)	80 (20)
Δ = .20	192 (48)	320 (80)	448 (112)
SPF-2×3•2	1–*ß* = .80	1–*ß* = .95	1–*ß* = .99
Δ = .80	18 (3)	24 (4)	36 (6)
Δ = .50	36 (6)	60 (10)	78 (13)
Δ = .20	210 (35)	348 (58)	480 (80)

Die Versuchspläne SPF-2×2•2, SPF-2×2•3 und SPF-2×3•2 eignen sich zur Kontrolle des Klassenkontextes. Dann wird unter Faktor *A* mit 2 bzw. 3 Klassen geplant, in denen jeweils mit Halbklassen unter Faktor *B* z. B. die Behaltensleistung bezüglich des Faktors *C* (z. B. Unterrichtsmethoden) untersucht wird. Für die Versuchspläne lassen sich mittlere Effekte (Δ = .50) unter einer statistischen Power von 1–*ß* = .95 untersuchen.

Tabelle 7.5 macht deutlich, dass mit großen Stichprobenumfängen (*N* ≥ 192) geplant werden muss, wenn kleine Effekte (Δ = .20) mit den Versuchsplänen SPF-2×2•2, SPF-2×2•3und SPF-2×3•2 untersucht werden sollen. Interessant ist die Tatsache, dass der

Versuchsplan SPF-2×2•3 ökonomischer ist als der Versuchsplan SPF-2×2•2 unter der statistischen Power 1–*ß* = .80, 1–*ß* = .95 und 1–*ß* = .99.

6. Kovariate Versuchspläne

Die Berücksichtigung von Kontrollvariablen in der Versuchsplanung hat einen Einfluss auf die Stichprobengröße. Je höher die Korrelation zwischen Kontrollvariable und abhängiger Variable ist, desto mehr sinkt der benötige Stichprobenumfang. Bausell und Li formulieren in ihrer *Strategie 6* zur poweranalytischen Versuchsplanung: „This is due to the fact that the resulting effect upon power is related to the size of the correlation coefficient between the covariate/blocking variable and the dependent variable … When a blocking variable is employed, an additional increment to statistical power may be provided … if this new variable interacts with the treatment and if it can be assumed that the addition of the blocking variable does not increase the overall within-group difference between subjects." (Bausell & Li, 2002, S. 25)

Abbildung 7.13 veranschaulicht den Zusammenhang der Korrelation (*r*) zwischen Kontrollvariable/abhängiger Variable und Stichprobenumfang *N* sowie Power (1–*ß*). Die Abbildung zeigt: (1) Je größer *r* ist, desto kleiner ist *N;* (2) je größer *r* ist, desto größer ist 1–*ß*.

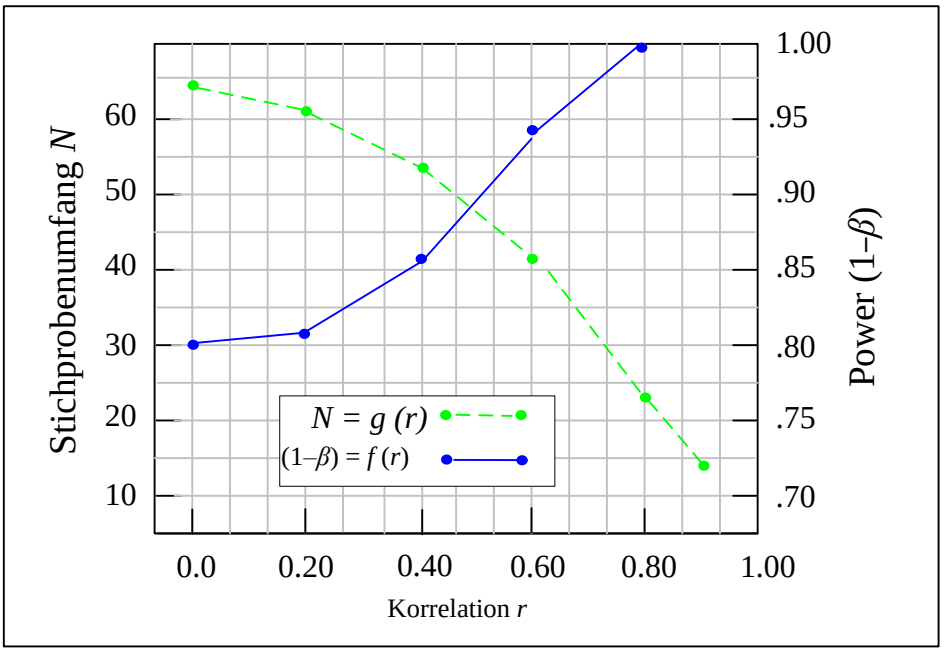

Abbildung 7.13 Korrelation von Kontrollvariable/abhängiger Variable (*r*) in Bezug zu *N* und 1–*ß* (Daten aus Bausell & Li, 2002, S. 26).

7. Multivariate Versuchspläne

Murphy, Myors und Wolach (2009, 2014) machen darauf aufmerksam, dass sich durch Einbeziehung mehrerer abhängiger Variablen in die Versuchsplanung die Power (1–*ß*) erhöhen lässt. (1) Je mehr abhängige Variablen, desto höher ist die Power; (2) je größer die Korrelation zwischen den abhängigen Variablen ist, desto größer ist die Power.

7.5 Zusammenfassung und Ausblick

Dieses Kapitel behandelte die Berechnung von Stichprobenumfängen für 15 konkrete Versuchspläne mit dem Softwarepaket PASS (Version 16). Die 15 konkreten Versuchspläne basieren auf den Versuchsplantypen RB-*p,* CR-*p,* CRF-*p*×*q*, SPF-*p*•*q* und SPF-*p*×*r*•*q*, die für die experimentelle Unterrichtsforschung wichtig sind.

Mit der Poweranalyse wurde das Verfahren vorgestellt, mit dem sich statistische Signifikanz und statistische Power verbinden lassen. Ausgangspunkte waren α-*Fehler* und *ß-Fehler* eines statistischen Hypothesentests. Es wurde veranschaulicht, dass für konkrete Versuchspläne die statistische Power (1–*ß*) vom Signifikanzniveau α, dem Stichprobenumfang *N* und der Effektgröße Δ abhängig ist.

Aus den poweranalytischen Berechnungen zu den Versuchsplänen wurde ersichtlich, dass der Stichprobenumfang abhängig ist von der statistischen Power, der Effektgröße bei gegebenem Signifikanzniveau (was in der Praxis per Konvention durch die Werte α = .05 bzw. α = .01 bestimmt ist) und vom gewählten Versuchsplan.

Folgende wichtige Strategien lassen sich für die Optimierung des Stichprobenumfangs festhalten: (1) Der Stichprobenumfang kann verkleinert werden durch Reduzierung der Power. (2) Der Stichprobenumfang kann verkleinert werden durch Vergrößerung der Effektgröße. (3) Der Stichprobenumfang kann verkleinert werden, indem SPF-Versuchspläne verwendet werden (Versuchspläne mit Messwiederholungen). (4) Der Stichprobenumfang kann verkleinert werden, indem Kontrollvariablen (kovariate Versuchsplanung) einbezogen werden. (5) Der Stichprobenumfang kann verkleinert werden, indem mehrere abhängige Variablen (multivariate Versuchsplanung) berücksichtigt werden.

Die poweranalytischen Berechnungen der Stichprobenumfänge wurden mit dem Softwarepaket PASS (Version 16) in Bezug zur *F*-Prüfgröße parametrischer Varianzanalysen durchgeführt. Die berechneten Stichprobenumfänge dienen als Richtlinie für Stichprobenumfänge, auch wenn nicht-parametrische statistische Verfahren eingesetzt werden müssen. Allerdings ist dann die statistische Power niedriger – beispielweise um 8%, wenn poweranalytische Berechnungen in Bezug zur *H*-Prüfgröße (Rangvarianzanalyse nach Kruskal-Wallis) mit den poweranalytischen Berechnungen in Bezug zur *F*-Prüfgröße verglichen werden (Van Hecke, 2012).

Wie in Kapitel 9 festgestellt, spielen für die Einbeziehung des Klassen- oder Schulkontextes Versuchspläne auf der Grundlage des *Hierarchical Linear Mode*l (HLM) eine besondere Rolle. Auch für diese Versuchspläne (z. B. 2- und 3-Ebenenmodelle) ist es von zentraler Wichtigkeit, Stichprobenumfänge in der Versuchsplanung einzubeziehen. In weiterführenden Arbeiten sollten dazu die notwendigen Stichprobenumfänge dokumentiert sein, unter Verwendung etwa der Software *Optimal Design* (Spybrook, Raudenbush, Congdon, & Martinez, 2011; Optimal Design Software, 2018).

Abschließend sei nochmals auf die enorme Wichtigkeit der Poweranalyse bei der Planung experimenteller Untersuchungen hingewiesen: „It is our firm conviction that no other process possesses more potential for increasing the scientific and societal yields accruing from our experiments." (Bausell & Li, 2002, S. ix)

7.6 Literatur

Aberson, C. L. (2012). *Applied power analysis for the behavioral sciences*. New York: Taylor & Francis.

APA. American Psychological Association (2010). *Publication manual of the American Psychological Association.* Washington. APA.

Bausell, R. B., & Li, Y.-F. (2002). *Power analysis for experimental research*. New York: University Press.

Bradley, D. R., Russell, R. L., & Reeve, C. P. (1996). Statistical power in complex experimental designs. *Behaviour Research Methods, Instruments, and Computers, 28*(2), 319–326.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Auflage). Hillsdale, NJ: Lawrence Erlbaum.

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

Corder, G. W., & Foreman, D. I. (2009). *Nonparametric statistics for non-statisticians: A step-by-step approach*. New York: Wiley.

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

Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric statistical methods*. New York: Wiley.

Huck, S. W. (2009). *Reading statistics*. Boston: Pearson.

Kirk, R. E. (1996). *Experimental design* (3. Auflage). Belmont: Wadsworth.

Kirk, R. E. (2012). *Experimental design* (4. Auflage). Belmont: Wadsworth.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects*? London: Sage Publications.

Lindley, D. V. (2006). Sample size determination. In S. Kotz (Hrsg.), *Encyclopedia of statistical sciences*, Vol. 11. (S. 7405–7406). New York: Wiley.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, California: Sage.

Murphy, K. R., Myors, B., & Wolach, A. (2009). *Statistical power analysis: A simple and general model for traditional and modern hypothesis test*s (3. Auflage). New York: Taylor & Francis.

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (4. Auflage). New York: Taylor & Francis.

NCSS (2018). *PASS 16*. Retrieved January 2, 2018 from http://www.ncss.com/pass

Optimal Design Software (2018). *Produktbeschreibung.* Retrieved January 2, 2018 from http://site-maker.umich.edu/group-based/optimal_design_software

Peng, C.-Y. J., Long, H., & Adaci, S. (2012). Power analysis software for educational researchers. *The Journal of Experimental Education, 80*(2), 113–136.

Ryan, T. P. (2013). *Sample size determination and power*. New York: Wiley.

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

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2011). *Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design Software Version 3.0*. Retrieved from January 2, 2018 from www.wtgrantfoundation.org.

Van Hecke, T. (2012). Power study of ANOVA versus Kruskal-Wallis test. *Journal of Statistics and Management Systems, 15*(2/3), 241–247.

Yuan, K.-H., & Zhang, Z. (2018). *Practical statistical power analysis using WebPower and R*. London: ISDSA Press

Zendler, A. (2016). Versuchspläne und Auswertungsansätze für die experimentelle Unterrichtsforschung. In A. Zendler (Hrsg.), *Empirische Didaktik / Fachmethodik Band 1: Versuchspläne für die Unterrichtsforschung — und Anwendungsbeispiele*. Berlin: epubli.