Planungen eines Experiments

2

Vor einem Experiment müssen mehrere Überlegungen getätigt und Entscheidungen getroffen werden. Welchen Effekt möchte ich untersuchen? Welche Einflussvariablen muss ich dafür manipulieren? Wie genau soll mein Experiment aussehen? Wie viele Messwiederholungen benötige ich?

2.1 Anatomie eines Experiments

Bevor wir über die konkrete Planung eines Experiments sprechen, sollten wir uns einen Überblick über die Grundbegriffe und die Bestandteile eines herkömmlichen computergestützten Experiments der Experimentalpsychologie verschaffen. Natürlich können sich diese Bestandteile je nach ihrer inhaltlichen Orientierung auch ändern oder weitere Phasen hinzukommen. Hier wird der Fokus jedoch primär auf Experimente gelegt, wie sie in der visuellen Aufmerksamkeitsforschung verwendet werden. Nichtsdestotrotz sind die Begrifflichkeiten und Konzepte auch über diesen spezifischen Forschungsbereich hinaus gebräuchlich. Es ist daher sehr wichtig, sich über die Bedeutung dieser Begriffe im Klaren zu sein, da dies nicht nur das Verständnis von Experimenten, welche vorgestellt werden, erhöht, sondern auch den Austausch mit Kollegen und Kolleginnen bei der Planung von Experimenten wesentlich erleichtert. Also, was sind Bildschirme, Durchgänge, gemischte Blöcke und reine Blöcke?

Bildschirme

Wenn Sie einen Blick auf Abb. 2.1 werfen, sehen Sie, dass ein Bildschirm die kleinste „Einheit“ eines Experiments darstellt. Auch wenn schnell klar ist, was

C. Büsel, *Psychologische Experimente*,
https://doi.org/10.1007/978-3-662-65524-5_2

„Bildschirm“ im gegenwärtigen Kontext bedeutet, mag der Begriff zunächst etwas verwirrend sein. Logischerweise hieven wir während eines Experiments nicht eine Reihe von Computermonitoren vor die Versuchspersonen. Bildschirm im Kontext eines Experiments bezieht sich auf jenen Reiz bzw. jene Konfiguration von Reizen, die einer Versuchsperson zu einem gegebenen Zeitpunkt präsentiert werden.

Durchgänge

Ein oder (meist) mehrere Bildschirme konstituieren einen Durchgang. Generell gesagt, sind Durchgänge jene Aufgaben, die Versuchspersonen im Verlauf eines Experiments erledigen sollen. Ein Durchgang ist beendet, wenn Versuchspersonen die von ihnen geforderte Aufgabe erledigt haben, woraufhin ein neuer Durchgang folgt. Welche Aufgabe die Versuchspersonen erledigen sollen, hängt natürlich vom jeweiligen Experiment ab. Oft sollen Versuchspersonen eine manuelle Antwort geben, eine schnelle Augenbewegung auf einen Reiz hin oder von einem Reiz weg ausführen. Das bedeutet, dass es in einem Verhaltensexperiment pro Durchgang meist einen einzelnen Datenpunkt gibt. Verwendet man andere Methoden, wie etwa Elektroenzephalografie (EEG), werden Daten über die Hirnaktivität über einen gesamten Durchgang hinweg gesammelt. Das Prinzip bleibt aber auch bei der Verwendung eines EEGs gleich: Man vergleicht Datenpunkte zu einem bestimmten Zeitpunkt innerhalb eines Durchgangs unter verschiedenen Experimentalbedingungen.

Blöcke

Gleich wie Durchgänge eine Sammlung an Bildschirmen darstellen, sind Blöcke eine Sammlung von sukzessiven aufeinanderfolgenden Durchgängen. Werfen Sie wieder einen Blick in Abb. 2.1: In dieser Abbildung sind zwei Blöcke mit den Bezeichnungen „Block 1“ und „Block 2“ dargestellt. Natürlich wäre es auch möglich, die Übungsphase als „Übungsblock“ zu bezeichnen, da Versuchspersonen auch in dieser Phase meist mehrere aufeinanderfolgende Durchgänge absolvieren und nicht lediglich einen. Es gibt jedoch verschiedene Möglichkeiten, experimentelle Bedingungen innerhalb eines Blockes zu präsentieren: gemischt oder rein (oft auch „geblockt“ genannt) (siehe Abb. 2.2). Kurz gesagt: in *gemischten Blöcken* kommen alle Bedingungen, die Sie in Ihrem Experiment testen, randomisiert vor. Das bedeutet, innerhalb eines Blocks kann ein Durchgang aus der Bedingung

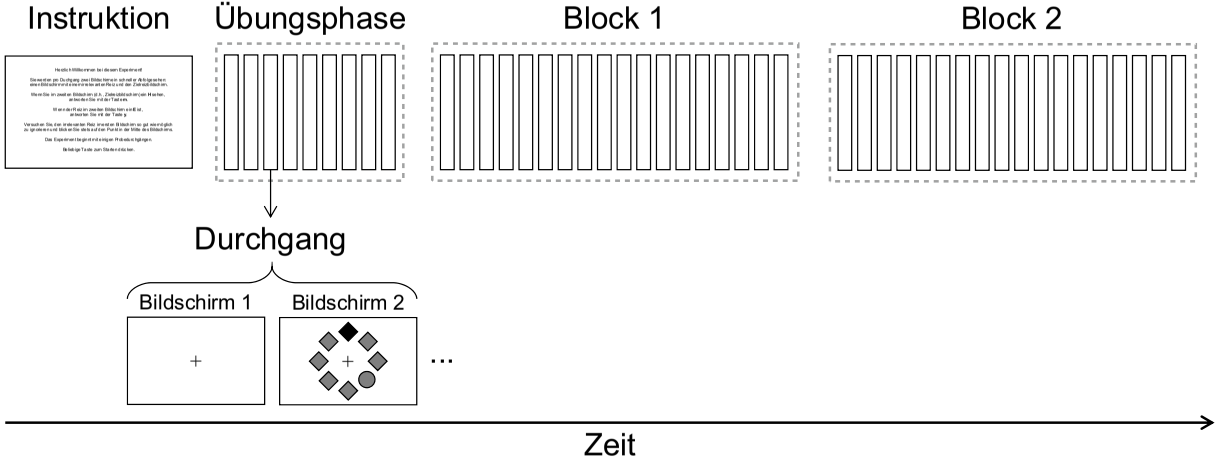


Abb. 2.1 Ein typischer Aufbau eines herkömmlichen Experiments: Nach der Instruktion für die Aufgabe erledigen die Versuchspersonen einige Durchgänge zur Übung (Übungsphase), woraufhin das Hauptexperiment mit der Datenerhebung beginnt. Dieses Hauptexperiment ist wiederum meist in mehrere Abschnitte (Block 1 & Block 2) unterteilt

A stammen und der nächste Durchgang zufällig aus Bedingung A oder Bedingung B. In *reinen Blöcken*, bzw. *geblockten Designs*, erledigen die Versuchspersonen über eine längere Zeit hinweg Durchgänge aus stets der gleichen experimentellen Bedingung. Für beide Arten von Blöcken kann es gute Gründe geben. Wollen Sie etwa Lerneffekte oder den Einfluss von Erwartungen und Vorbereitung über mehrere Durchgänge hinweg untersuchen, kann es sinnvoll sein, stets die gleiche Bedingung zu präsentieren. Wollen Sie aber Lerneffekte oder damit verwandte Alternativerklärungen möglichst ausschließen, dann sind gemischte Blöcke das Mittel der Wahl. Warum diese beiden Arten von Blöcken einen relevanten Einfluss auf die Studienergebnisse und deren Interpretation haben können, wird in Abschn. 2.6 noch zusätzlich näher beleuchtet.

Reines und gemischtes Design am Beispiel von Treisman und Gelade (1980)

Die Rolle der Aufmerksamkeit in der visuellen Suche ist eine traditionelle und seit Langem erforschte Forschungsfrage in der Kognitionspsychologie. Unter welchen Umständen können wir effizient relevante Zielreize von irrelevanten Distraktoren unterscheiden? Welche Suche ist schwieriger und welche ist leichter? Wann benötigen wir mehr oder weniger Aufmerksamkeit? Anne

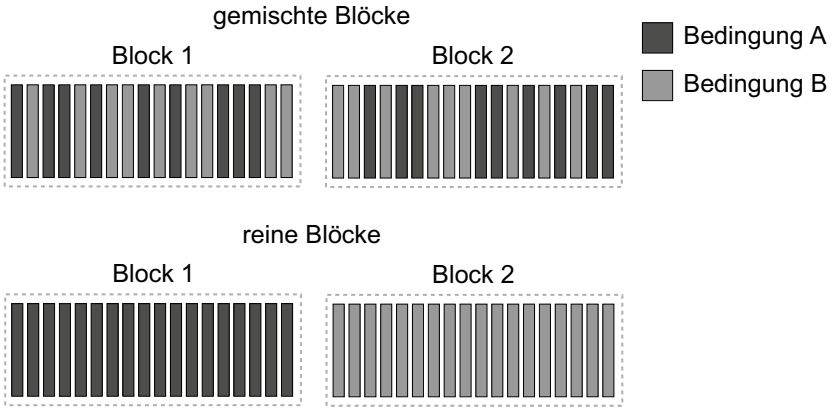
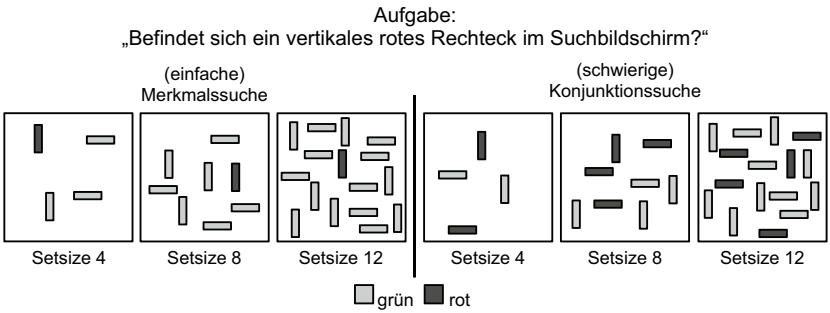


Abb. 2.2 In einem Experiment mit Messwiederholungen wird zumeist nicht nur eine Bedingung untersucht, sondern mehrere, die anschließend verglichen werden. Diese Bedingungen können entweder (pseudo-)randomisiert (gemischte Blöcke) oder in zwei verschiedenen Blöcken (reine Blöcke) dargeboten werden

Treisman und ihre Kolleginnen und Kollegen leisteten einen so großen Beitrag zur Beantwortung dieser Fragen, dass Treisman 2011 sogar von Präsident Obama mit der National Medal of Science ausgezeichnet wurde.

In ihrer *Feature-Integration Theory* stellten Treisman und Gelade (1980) ein Modell auf, das eine Erklärung für den Einfluss der Aufmerksamkeit unter schwierigen und einfachen Suchbedingungen erklären soll. Dazu ließen Treisman und Gelade ihre Versuchspersonen einfache und schwierige Suchaufgaben erledigen. Während des gesamten Experiments sollten die Versuchspersonen mittels Tastendruck angeben, ob sich ein Zielreiz im Suchbildschirm befindet oder nicht. In der *einfachen Suchbedingung* unterschied sich der Zielreiz stark von den Distraktoren (für eine sinngemäße Darstellung siehe Abbildung unten links). Genauer gesagt unterschied sich der Zielreiz in der einfachen Suchbedingung in einem Merkmal (z. B. Farbe) von den umgebenden Distraktoren.

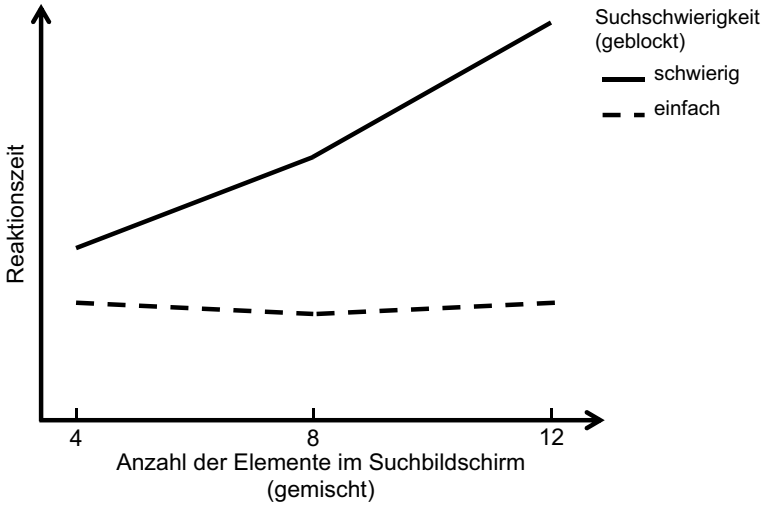
In der schwierigen Suchbedingung unterschied sich der Zielreiz von den Distraktoren jedoch nicht nur in einer Farbe oder einer Form von den Distraktoren, sondern in einer Kombination dieser beiden Merkmale. Das bedeutet, dass eine Merkmalssuche nicht mehr ausreichte, um den Zielreiz aufzuspüren, sondern eine *Merkmalskombinationssuche* (kürzer: Konjunktionssuche) nötig war, um den Zielreiz zu identifizieren. Versuchspersonen erledigten abwechselnd mehrere schwierige und mehrere einfache Suchblöcke.



Nun gut, wir sehen, wie Treisman und Gelade (1980) den Faktor der Suchschwierigkeit geblockt manipulierten. Wie wiesen sie aber nach, ob die Versuchspersonen effizient nach dem Zielreiz suchten oder nicht? Hier kommt ein zweiter Faktor ins Spiel: die Anzahl der Elemente im Suchbildschirm (bzw. kürzer im Englischen: *setsize*). In der Abbildung sehen Sie, dass 4, 8 oder 12 Reize im Suchbildschirm sein konnten. Warum?

Nehmen wir an, wir können die Anwesenheit eines Zielreizes ohne Umschweife sofort erkennen, weil sich der Zielreiz so markant von seiner Umgebung abhebt. Wie wirkt sich Ihrer Meinung nach die Anzahl der Distraktoren im Suchbildschirm aus? Genau, so gut wie gar nicht. Wir können den Suchbildschirm schnell scannen, die visuellen Reize praktisch *parallel* verarbeiten und entsprechende Abweichungen effizient und schnell erkennen. Wenn die Zielreize und Distraktoren einander allerdings hinreichend ähnlich sind, dann genügt diese parallele Suche nicht mehr. Stattdessen müssen wir die einzelnen Reize *seriell* absuchen und nach jeder Selektion eines Reizes entscheiden, ob es sich dabei um den Zielreiz handelt oder nicht. Entsprechend länger dauert die Suche.

Diese Annahme klingt vernünftig, doch wollen wir sie mit Daten stützen. Genau hier kommt die Setsize-Manipulation ins Spiel: Wenn die Suche effizient und parallel vonstatten geht, sollte sich die Zeit, die man zur Identifikation des Zielreizes benötigt, zwischen den verschiedenen Setsize-Bedingungen nicht wesentlich unterscheiden. Anders verhält es sich bei einer ineffizienten und seriellen Suche. Unter diesen Suchbedingungen hat man mit mehr ähnlichen Distraktoren noch mehr Reize, die man absuchen und klassifizieren muss. Daraus folgt, dass die Suchzeit als eine Funktion der Anzahl der Elemente im Suchbildschirm ansteigen sollte. Um diese Hypothese untersuchen zu können, variierten Treisman und Gelade (1980) die Anzahl der Elemente im Suchbildschirm zufällig von Durchgang zu Durchgang. Die Setsize-Bedingung wurde also innerhalb der Blöcke (= Suchschwierigkeit) gemischt dargeboten. Stellt man die Reaktionszeiten der Versuchspersonen im Verhältnis zur Setsize als eine Suchfunktion dar, konnten Treisman und Gelade ganz genau das finden:



In ihrer *Feature-Integration Theory* argumentieren Treisman und Gelade (1980), dass man Unterschiede in einzelnen Merkmalsdimensionen bereits ohne Aufmerksamkeitszuwendung („prä-attentiv“) identifizieren kann (ein hinreichend großer Merkmalskontrast vorausgesetzt, vgl. Duncan & Humphreys, 1989). Muss jedoch nach einer Kombination mehrerer Merkmale gesucht werden, benötigt es Aufmerksamkeit, um diese unterschiedlichen Merkmale zu kombinieren.

Wir kennen das auch aus dem realen Leben: Stellen Sie sich vor, sie fahren nachts mit dem Auto auf einer wenig befahrenen Landstraße. Sie passieren Bäume, Sträucher, Felder und plötzlich sehen sie etwas: einen gesperrten Bahnübergang. Für Sie ist es in einer solchen Situation vermutlich belanglos, wie viele Sträucher Sie neben der Straße sehen. Sofern die Sträucher oder Bäume die Ampel am Bahnübergang nicht verdecken, nehmen Sie dieses Warnsignal sofort wahr.

Fahren Sie mit Ihrem Auto jedoch in einer größeren Stadt (why? … why???) und wollen herausfinden, ob sie Vorfahrt geben müssen, dann ist ein Vorfahrt-achten-Schild nicht auffällig genug, dass Sie sich locker zurücklehnen können und das Verkehrsschild schon Ihre Aufmerksamkeit einfängt. Stattdessen müssen Sie den Schilderwald gezielt und aufmerksam nach der Information absuchen, die für Sie relevant ist. Dreieckig sind nämlich mehrere Warnzeichen, ebenso wie rot. Sie müssen also gezielt nach einem auf dem Kopf stehenden, rot umrahmten Dreieck suchen. ◄

2.2 Max-Kon-Min Prinzip

Wir versuchen für gewöhnlich, in unseren Experimenten bestimmte Effekte nachzuweisen. Stellen wir uns ein Experiment daher kurz anhand eines sehr zeitgemäßen Beispiels vor: Wir möchten manuell zwischen Radiosendern wechseln (OK, Boomer...). Dabei ist die gewünschte Radiostation der Effekt, den wir erreichen wollen. Das Rauschen zwischen den einzelnen Stationen ist etwas, das wir minimieren wollen. Nur durch ein sensibles Herumdrehen des Knopfes können wir dabei unser gewünschtes Ergebnis erreichen: ein optimales Verhältnis des Signals (der Musik des Radiosenders) zum Hintergrundrauschen (in der englischsprachigen Literatur werden Sie hierzu oft „signal-to-noise ratio“ lesen können). Um unsere grauen Zellen noch zusätzlich zu fordern, stellen wir uns vor, dass der gewünschte Radiosender in Wien eine Frequenz von 92 MHz hat und in Innsbruck 87,6 MHz. Wir müssen also (1) das Signal **max**imieren, (2) die für die Örtlichkeit korrekte Frequenz wählen, also für den Ort **kon**trollieren und (3) das Rauschen **min**imieren.

Was hat das jetzt mit der Experimentalpsychologie zu tun? Um einen spezifischen Effekt zu finden, sollten wir uns zunächst bewusst sein, was dieser Effekt denn an und für sich ist: das Variieren der abhängigen Variable in Abhängigkeit der jeweiligen Bedingung (der Experimental- oder Kontrollbedingung). Wir untersuchen also, ob die Varianz unserer Daten durch die von uns gewählten Bedingungen erklärt werden kann. Das klingt nun vielleicht weniger trivial, als es eigentlich ist. Die Varianz, die wir in den Daten beobachten (Gesamtvarianz) kann nämlich durch drei Quellen zustande kommen (siehe Abb. 2.3): der Primärvarianz, der Sekundärvarianz und des Zufallsfehlers (Kerlinger, 1973).

Primärvarianz

Unter der Primärvarianz versteht man den Anteil der systematischen Varianz der durch die systematische Variation der Experimentalbedingungen (UV) zustande kommt. In einem guten Experiment gilt es, diese Primärvarianz zu maximieren. Dies wird durch die Wahl von optimalen Faktoren und Faktorenstufen erreicht, die miteinander verglichen werden sollen. Ideal ist es hier, Extremstufen von Faktoren zu wählen, welche die Unterschiede zwischen den Bedingungen maximiert. Wenn Sie zum Beispiel demonstrieren wollen, dass kongruente Bedingungen in der Stroop-Aufgabe besonders hilfreich sind (beispielsweise das Wort „Rot“ in roter Farbe), dann vergleichen Sie diese Bedingung für gewöhnlich

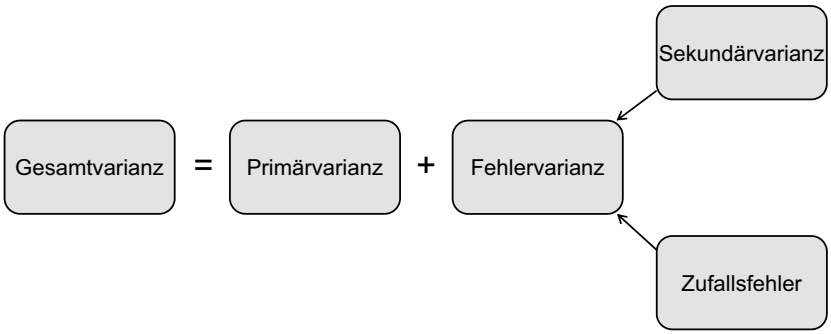


Abb. 2.3 Die Gesamtvarianz der abhängigen Variable in einem Experiment setzt sich aus der Primär- und Fehlervarianz zusammen. Die Fehlervarianz setzt sich ihrerseits wiederum aus der Sekundärvarianz und dem Zufallsfehler zusammen

nicht mit einer neutralen Bedingung (beispielsweise das Wort „Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz“ in roter Farbe), sondern mit einer inkongruenten Bedingung (bspw. das Wort „Blau“ in roter Farbe). Zu Faktoren werden Sie im gleichnamigen Abschnitt noch Näheres lernen.

Sekundärvarianz

Unter der Sekundärvarianz versteht man den Anteil systematischer Varianz, der durch nicht berücksichtigte und unkontrollierte Faktoren zustande kommt. Die Sekundärvarianz kann die Interpretation der gefundenen Ergebnisse erschweren bzw. im schlimmsten Falle sogar verunmöglichen. Stellen Sie sich folgendes „Experiment“ vor: Wir wollen Geschlechtsunterschiede in den arithmetischen Fähigkeiten in einem Computerexperiment überprüfen, in dem die Versuchspersonen randomisiert Additionen, Subtraktionen, Multiplikationen und Divisionen durchführen sollen. Dafür rekrutieren Sie junge Frauen aus einem humanistisch ausgerichteten Gymnasium und junge Männer aus einer Höheren Technischen Lehranstalt (HTL). Ihre Ergebnisse suggerieren große Unterschiede in den mathematischen Kompetenzen zwischen den Geschlechtern: Männer erreichten signifikant mehr Punkte als Frauen. Kann man die Ergebnisse jedoch dahingehend interpretieren, dass Männer generell besser in Mathematik sind als Frauen? Mitnichten!

Eine Vielzahl an nicht berücksichtigten Variablen könnten diese Unterschiede erklären. Hier nur zwei wahrscheinlich höchst relevante Konfundierungen:

1. *Selbstselektion:* Es ist anzunehmen, dass sich technisch und mathematisch interessierte und begabte Personen eher für eine technische Schule entscheiden als Personen, die sich eher für Sprachen begeistern.
2. *Unterrichtsfächer:* Schüler:innen an einer HTL haben wesentlich mehr Unterrichtseinheiten, die sich mit Mathematik und verwandten Fächern beschäftigen, als Schüler:innen an einem humanistischen Gymnasium. Das bedeutet klarerweise, dass Schüler:innen an der HTL wesentlich mehr Übung in Arithmetik haben als jene, die ein humanistisches Gymnasium besuchen.

Sie sehen, die Interpretation des eben beschriebenen (und erfundenen!) Ergebnisses ist, streng genommen, gar nicht möglich. Die gefundenen Unterschiede könnten genauso gut durch die Konfundierungen erklärt werden.

Kollinearität
Kollinearität beschreibt in der Statistik das Ausmaß eines Zusammenhanges zwischen zwei Variablen (spezifischer: UVs). Kollinearität sollte in Verfahren wie etwa einer Regressionsanalyse tunlichst vermieden werden. Die Problematik sollte durch das oben genannte Beispiel ersichtlich sein: Korrelieren zwei Prädiktoren (etwa das Geschlecht und die Anzahl an Mathematikstunden) zu hoch miteinander, dann ist eine getrennte Interpretation der einzelnen Prädiktorvariablen in einem Regressionsmodell nicht möglich.

Zufallsfehler

Den Zufallsfehler kann man mit dem Messfehler aus der klassischen Testtheorie vergleichen, und tatsächlich gibt es zwei hervorstechende Gemeinsamkeiten mit dem Messfehler der klassischen Testtheorie:

1. Der Erwartungswert, also jener Wert, den eine Variable im Mittel nach unendlich vielen Messwiederholungen annimmt, ist beim Messfehler 0. Ähnlich verhält es sich mit der Varianz. Wir werden die Sekundärvarianz realistisch gesehen wohl nie auf 0 bekommen, werfen Sie aber einen Blick auf die Berechnung der Varianz:

$$var = \frac{\sum_{i=0}^{n} (x_i - \overline{x})}{n}$$

Anhand der Formel ist gut ersichtlich, dass die Größe der Varianz als eine Funktion der Stichprobengröße (n) abnimmt – je größer der Nenner wird, desto kleiner wird das Resultat (siehe Abb. 2.4).

2. Fast wichtiger als der erste Punkt: Der Zufallsfehler korreliert nicht mit dem gemessenen Effekt. Das bedeutet, dass der Zufallsfehler die Primärvarianz nicht systematisch beeinflusst, sondern lediglich Rauschen in den Daten ist, das es zu minimieren gilt, da eine Kontrolle des Zufalls ein gleichermaßen ermüdendes wie hoffnungsloses Unterfangen ist.

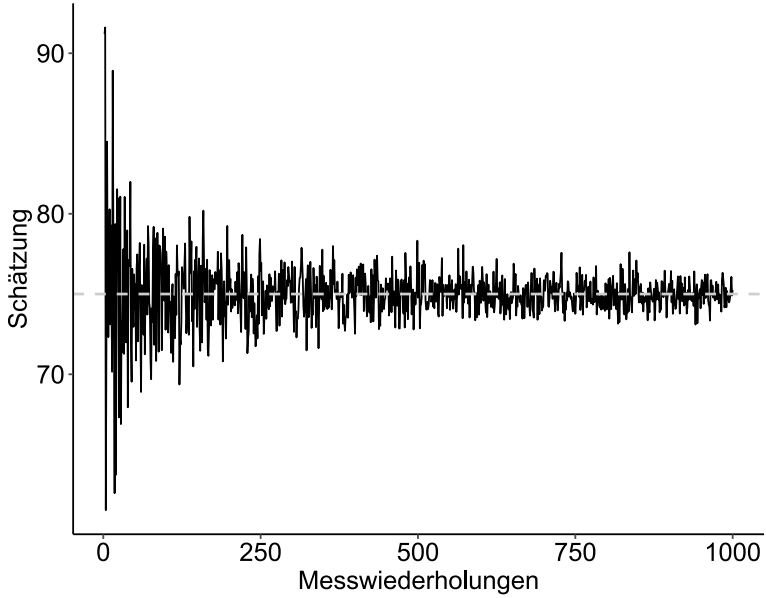


Abb. 2.4 Simulation zur Zunahme der Messgenauigkeit. Zwei bis 1000 Werte werden zufällig aus einer Normalverteilung (N[75; 25]) gezogen und gemittelt. Je mehr Werte gezogen werden, desto präziser wird die Schätzung des tatsächlichen Mittelwerts einer Variable (hier: die gestrichelte graue Linie)

2.3 Faktoren

Selbstverständlich hängt ein experimentelles Design von der exakten Fragestellung ab. Mithilfe eines Experiments lassen sich Wenn-dann-Vorhersagen, sprich: Hypothesen, testen. Stellen wir uns kurz folgende Forschungsfrage vor: Führen unbewusste Merkmalswiederholungen zu verbesserten Leistungen (z. B. zu einem schnelleren Erkennen/Klassifizieren desselben Merkmals zu einem späteren Zeitpunkt)? Lassen wir vorerst außer Acht, dass diese Forschungsfrage bereits seit Langem intensiv be- und erforscht wird (im supraliminalen Bereich z. B. Maljkovic & Nakayama, 1994), und überlegen wir uns ein angemessenes Design für diese Forschungsfrage: Wir wollen von unseren Versuchspersonen Urteile über einfache Reize erfragen. Beispielsweise könnten wir unseren Versuchspersonen Hunde- und Katzenbilder präsentieren und die Versuchsperson soll angeben, ob es sich bei dem gezeigten Vierbeiner um ein Exemplar der Gattung Canis oder Felis handelt (den zoologischen Hintergrund der Versuchspersonen sollte man in den Instruktionen selbstverständlich berücksichtigen). So weit, so einfach.

Wie aber kann man unbewusst einen mit den beiden Kategorien verwandten Reiz präsentieren? Eine Methode wäre es, kurz vor den Tierbildern für sehr kurze Zeit ausgeschriebene Tierlaute zu präsentieren und nachfolgend unmittelbar mit einem anderen Reiz zu überdecken (d. h. zu maskieren), damit sie von den Versuchspersonen nicht bewusst wahrgenommen werden könnten. Diese kurz davor präsentierten Reize nennen wir in weiterer Folge Primes (zu Deutsch manchmal Bahnungsreize genannt). Geeignete Primes wären zum Beispiel „Wuff“ und „Miau“.

Wenn wir nun die Primes und die Tierbilder zufällig zusammenwürfeln, sollten sich folgende Bedingungen gleich oft ergeben:

1. Prime: Wuff – Zielreiz: Hund
2. Prime: Miau – Zielreiz: Katze
3. Prime: Miau – Zielreiz: Hund
4. Prime: Wuff – Zielreiz: Katze

Wir haben nun zum einen kongruente Bedingungen (1. und 2.) und zum anderen inkongruente Bedingungen (3. und 4.). Wie Sie bereits gemerkt haben, stimmt die Spezies beider Reize (des Primes und des Zielreizes) in kongruenten Bedingungen überein, während sie in inkongruenten Bedingungen nicht übereinstimmt.

Unsere Hypothese ist nun wie folgt: Wenn Versuchspersonen die subliminal präsentierten Primes verarbeiten, dann sollten kongruente Durchgänge zu signifikant besseren Leistungen (d. h. schnelleren Antworten und weniger Fehlern) führen, als inkongruente.

Behaupten wir nun auch noch, dass es einen Unterschied zwischen Katzen und Hunden insofern gibt, als dass mögliche Kongruenzeffekte lediglich für eine Spezies vermutet werden. Wir haben also ein 2×2-faktorielles Experiment:

1. Faktor: Spezies des Zielreizes (Hund oder Katze)
2. Faktor: Prime (kongruent oder inkongruent)

In der eben verwendeten und konventionellen Schreibweise beschreibt das $\times$-Symbol, dass mehrere Faktoren miteinander kombiniert werden. Die exakte Zahl beschreibt die Anzahl der Faktorstufen. In unserem Beispiel haben wir zwei zweistufige Faktoren (Hund und Katze/kongruent und inkongruent). Wir könnten unser Experiment aber auch einfach in ein 3×2-Experiment verwandeln, in dem wir aus dem Faktor Spezies des Zielreizes einen dreistufigen Faktor machen (z. B. Hund, Katze oder Huhn). Die Anzahl der Faktorstufenkombinationen wird dabei stets gleich berechnet $\rightarrow$ wie es geschrieben steht.

Mit unseren zwei zweistufigen Faktoren hat unser Experiment 2 mal 2, also 4 mögliche Faktorstufen (Hund-kongruent, Hund-inkongruent, Katze-kongruent und Katze-inkongruent). Hätten wir ein $2 \times 2 \times 2 \times 3$-Experiment, hätte unser Experiment 24 Faktorstufenkombinationen.

2.4 Zwischen- & Innersubjektfaktoren

Zwischensubjektfaktoren

Wie Sie im letzten Abschnitt bemerkt haben, steigt die Anzahl der benötigten Durchgänge exponentiell mit den variierten Faktoren an. Selbst die freundlichste Versuchsperson wird allerdings nach etwa 3000 Versuchsdurchgängen ihre Kooperationswilligkeit verlieren. Eine Variante, dieses Problem zu umgehen, ist die Verwendung eines Faktors als Zwischensubjektfaktors (engl. between-subjects): Eine Gruppe von Versuchspersonen ist in der Bedingung A, während eine andere Gruppe in Bedingung B ist. Manchmal haben wir aber auch gar keine andere Wahl, als ein Between-Subjects-Design zu verwenden, da die Versuchspersonen bereits vor dem Experiment einer bestimmten Gruppe zugehören. Wollen wir zum Beispiel Leistungsunterschiede zwischen verschiedenen Berufsgruppen

untersuchen, können wir Versuchspersonen nicht zufällig einem Beruf zuordnen. Ein Nachteil eines Between-Subjects-Designs ist aber, dass die Sekundärvarianz (systematische, aber unbedachte Varianz) nur durch eine wesentlich größere Stichprobe kontrolliert werden kann. Warum? Nun, wenn wir z. B. nur fünf Personen aus der Gruppe der Pflegeberufe mit wieder nur fünf Personen aus der Gruppe der Bürokräfte vergleichen, können sich die Personen zwischen den beiden Gruppen aus Gründen unterscheiden, die nicht zwangsläufig mit der eigentlichen Forschungsfrage zu tun haben. Je mehr Versuchspersonen sich jedoch in beiden Vergleichsgruppen befinden, desto höher ist die Wahrscheinlichkeit, dass sich diese Unterschiede zwischen den beiden Gruppen egalisieren.

Innersubjektfaktoren

Das Problem der unkontrollierten Sekundärvarianz ist bei Innersubjektfaktoren (engl. within-subjects) weniger gegeben. In einem reinen Within-Subject-Design durchläuft jede Versuchsperson alle in einem Experiment möglichen Faktorstufenkombinationen und dient daher beim Vergleich zwischen den Bedingungen gleichsam als eigene Vergleichsstichprobe. Um das zu verdeutlichen: Neigt eine Person generell zu schnelleren Reaktionen, dann können Differenzen zwischen zwei Experimentalbedingungen trotz der schnelleren Antworttendenz miteinander verglichen werden, da die Versuchsperson zwar im Schnitt schneller sein mag als andere Versuchspersonen, die Differenz zwischen den Bedingungen aber in etwa von der gleichen Größe sein kann.

Gemischte Designs

Die dritte Alternative, Faktoren in einem Experiment zu variieren, ist das sogenannte gemischte Design. Wie der Name schon anklingen lässt, ist das gemischte Design eine Kombination aus Inner- und Zwischensubjektfaktoren: Zum einen unterscheiden sich zwei oder mehrere Gruppen an Versuchspersonen in einer relevanten Variable, durchlaufen aber in einem Experiment dieselben experimentellen Manipulationen. Es gibt verschiedene Gründe, auf ein gemischtes Design zurückzugreifen. Ein Grund kann sein, dass es die Fragestellung kaum anders zulässt, als zumindest einen Faktor zwischen Versuchspersonen zu variieren. Stellen wir uns folgende Fragestellung vor: Sind Raucher:innen, die lange keine Zigarette mehr geraucht haben, anfälliger dafür, irrelevante Informationen zu verarbeiten als jene Raucher:innen, die ihre Sucht erst kürzlich befriedigen konnten? Sie

wollen Ihre Forschungsfrage mit der klassischen Stroop-Aufgabe beantworten, die wir in Abschn. 2.2.1 bereits kurz erwähnt haben. Zur Wiederholung: In der Stroop-Aufgabe sollen Versuchspersonen die Farbe, in der Farbwörter gedruckt sind, benennen. Die Bedeutung der Wörter ist an sich für die Aufgabe vollkommen irrelevant, da nur die Druckfarbe des Wortes benannt werden soll und die Wörter nur zufällig die gleiche Farbe angeben, in der sie gedruckt sind. Der herkömmliche Fund in der Stroop-Aufgabe ist, dass Versuchspersonen die Farbe, in der ein Wort gedruckt ist, schneller benennen können, wenn das Wort mit der Druckfarbe übereinstimmt (Stroop, 1935).

Nun haben Sie zwei Varianten, die Untersuchung anzustellen. Zum einen können Sie eine Hälfte Ihrer rauchenden Versuchspersonen vor dem ersten Block eine Zigarette rauchen lassen, nach diesem Block eine längere Pause ansetzen und die mittlerweile schon zitternden Versuchspersonen zu einem späteren Zeitpunkt den zweiten Block mitten im Nikotinentzug erledigen lassen. Würden Sie nur diese Versuchspersonen auswerten, könnten Sie zu dem Befund kommen, dass Entzugserscheinungen zu kleineren Stroop-Effekten führen. Alternativ könnten Sie aber einfach einen Übungseffekt gefunden haben, da die Versuchspersonen im zweiten Block schlicht schon vertrauter mit der Aufgabe waren. Darum wäre es möglich, eine gleiche Anzahl an Versuchspersonen zu bitten, vor dem Experiment mindestens drei Stunden keine Zigarette zu rauchen, um sie dann nach dem ersten Block von ihrem gesundheitsförderndem Leid zu erlösen, ihnen eine Zigarette erlauben und danach den zweiten Block erledigen lassen. Dabei handelt es sich bereits um ein gemischtes Design: Wenn Sie kontrollieren wollen, ob die Reihenfolge der Bedingungen (Nikotin vor dem ersten Block vs. Nikotin vor dem zweiten Block) die Ergebnisse signifikant beeinflusst, dann ist die *Blockreihenfolge* ein Zwischensubjektfaktor und die *Kongruenz zwischen Druckfarbe und Farbwort* (kongruent vs. inkongruent) ein Innersubjektfaktor.

Alternativ könnten Sie sich auch aus zeitlichen Gründen für ein gemischtes Design entscheiden. Sollte ein Experiment zu lange werden, ist es auch möglich, zumindest einen Faktor als Zwischensubjektfaktor zu realisieren. Um bei unserem vorherigen Beispiel zu bleiben: Wenn Ihre Versuchspersonen drei Stunden zwischen den Blöcken warten müssten, damit sich die Entzugserscheinungen in ihrer vollen Pracht entfalten können, wäre eine zeitlich schonendere Alternative, eine Gruppe vor der Stroop-Aufgabe rauchen zu lassen und die zweite Gruppe schlicht zu bitten, mindestens drei Stunden vor dem Experiment keine Zigarette mehr zu rauchen. Der Zwischensubjektfaktor wäre somit *Rauchen vor dem Experiment* (ja vs. nein), während der zweite Faktor (*Kongruenz zwischen Druckfarbe und Farbwort:* kongruent vs. inkongruent) weiterhin einen Innersubjektfaktor darstellt.

2.5 Messwiederholungen

Sollten Sie schon einmal das Vergnügen gehabt haben, an einem (verhaltens-) psychologischen Experiment teilzunehmen, werden Sie festgestellt haben, dass Sie in einem 2×2-faktoriellen Design nicht einfach 4 oder in einem $2 \times 2 \times 2 \times 3$-faktoriellen Design nur 24 Durchgänge absolvieren. Im Gegenteil: Viele Experimente scheinen eher die Relativitätstheorie überprüfen zu wollen und eruieren, wie viel subjektiv empfundene Zeit in eine halbe oder volle Stunde passt. Sind wir Experimentalpsychologinnen und -psychologen Sadisten, die aus reinem Glück die Gefängnismauern von außen betrachten? Die Antwort auf diese Frage ist ein klares „ja“, aber das hat mit der Länge von Experimenten nur bedingt etwas zu tun.

Wir kennen den Grund für die Notwendigkeit von Messwiederholungen schon aus dem Abschnitt zum Zufallsfehler. Nehmen wir dazu noch ein Beispiel: Werfen Sie einen Blick auf das linke Bild in Abb. 2.5 und stellen Sie sich folgende Fragestellung vor, die sie beantworten sollen: Befindet sich in diesem Bild ein Tier und ist dieses Tier ein Hund oder eine Katze? Übersetzt in die Experimentalpsychologie fragen wir uns also: Gibt es ein Signal bzw. einen Effekt? Und wenn es einen Effekt gibt: Wie sieht dieser Effekt genau aus? Wenn nur wenige Pixel/Messungen einen großen Bereich eines Bildes repräsentieren, fällt es uns sehr schwer, uns auf Basis des sehr unscharfen Bildes ein zuverlässiges Urteil zu bilden. Nach rechts wandernd stellen wir aber fest, dass mehr Messungen der Farbwerte in immer kleineren Regionen die Zuverlässigkeit unseres Urteils erhöhen. (Das ist übrigens auch eine nette Gelegenheit, kurz die Messgenauigkeit des amerikanischen Wahlmännersystems zu reflektieren.)

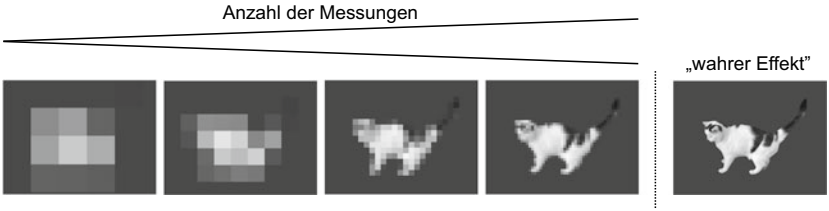


Abb. 2.5 Mehr Pixel ergeben ein klareres Bild. Vergleichbar verhält es sich auch mit der Anzahl an Messwiederholungen in (psychologischen) Experimenten: Je mehr Datenpunkte wir sammeln, desto näher kommen wir der wahren Gestalt eines Effektes. Ich danke meiner Katze für die (sehr kurze und widerwillige) Kooperation für diese Abbildung.

Derselben Logik folgen wir in psychologischen Experimenten: Würden wir alle Menschen auf der Welt unzählige Male testen können, könnten wir mit beinahe absoluter Sicherheit sagen, ob es diesen oder jenen Effekt gibt und wie dieser Effekt exakt aussieht. De facto können wir für jedes unserer Experimente aber nur eine begrenzte Anzahl an Versuchspersonen testen. Auch wenn die Anzahl der Versuchspersonen nicht das einzige Kriterium für zuverlässige Messungen ist (eine eigentlich unkontroversielle Aussage, die bei manchen dennoch zu Schnappatmung führt), sollte man über eine ausreichend große Stichprobe verfügen. Für Experimente mit Zwischensubjektfaktoren gibt es online bereits eine Vielzahl an Rechnern, die uns Auskunft über die optimale Stichprobengröße geben („Power"). In Experimenten, die rein aus Innersubjektfaktoren bestehen, ist die Ermittlung der optimalen Stichprobengröße allerdings alles andere als trivial – und je nach Perspektive auch nicht sonderlich sinnvoll (vgl. Smith & Little, 2018; siehe auch Exkurs zu *Versuchspersonenanzahl und Messwiederholungen*).

Fast wesentlicher für die zuverlässige Messung eines Effekts in einem Experiment mit Innersubjektfaktoren ist die Anzahl an Messwiederholungen. Jeder einzelne Tastendruck als Reaktion auf eine experimentelle Bedingung ist schlussendlich das Ergebnis vieler nicht kontrollierbarer Prozesse. Vielleicht war die Versuchsperson gerade abgelenkt, müde, mit den Gedanken woanders und, und, und. Alle diese Umstände können die Messung ungenauer machen. Dieser Einfluss unkontrollierbarer Prozesse wird oft als Rauschen bezeichnet, während die Variation in den Daten, die durch einen tatsächlichen Effekt zustande kommt, als Signal bezeichnet wird. Der Effekt nimmt systematisch Einfluss auf die Daten, während andere Prozesse, wie etwa Blinzeln, Gähnen oder Sonstiges, unsystematisch auf die Daten einwirken. Das Ziel der Messwiederholungen ist daher, durch mehrere Messungen das Rauschen in den Daten herauszumitteln und das Signal besser vom Rauschen abzugrenzen.

Anders als in Abb. 2.4 messen wir jedoch denselben Effekt nicht bis zu tausendmal innerhalb eines Experiments. Als eine Faustregel werden für gewöhnlich **mindestens** 25 bis 50 Messwiederholungen für Verhaltensexperimente geplant. Das bedeutet, dass jede einzelne Faktorstufenkombination mindestens 25-mal gemessen werden soll. Unser Hund-Katz-Experiment müsste daher aus mindestens 100 Durchgängen bestehen (4*25), während ein hypothetisches $2 \times 2 \times 2 \times$ 3-faktorielles Experiment aus mindestens 600 (24*25) Durchgängen bestehen müsste.

Zu beachten ist, dass wir hier stets von komplett ausbalancierten Designs sprechen, in denen jede Bedingung gleich oft vorkommt. Möchten wir in unserem Hund-Katz-Experiment noch eine Wahrscheinlichkeitsmanipulation, sodass etwa die Wahrscheinlichkeit für einen Hund als Zielreiz doppelt so hoch ist wie für eine

Katze als Zielreiz, dann müsste die Katze 50-mal (25*kongruent + 25*inkongruent) gemessen werden, während der Hund 100-mal getestet werden müsste (50*kongruent + 50*inkongruent).

Messwiederholungen mit anderen experimentellen Methoden

Die Faustregel von 25–50Messwiederholungen ist vorrangig in Verhaltensexperimenten gültig. Bei anderen Methoden, wie beispielsweise der Elektroenzephalographie (EEG), sind andere Tatsachen zu berücksichtigen: die neuronale Aktivität einiger weniger Neuronen von Interesse kann durch die Aktivität von umliegenden Neuronen überlagert werden. Die Effekte, die man beispielsweise mittels ereigniskorrelierten Potenzialen (ERPs) messen möchte, sind oft im μV (Mikrovolt) Bereich, weshalb unsystematische Einflüsse besonders gründlich eliminiert werden müssen. Anders ausgedrückt: Ein relativ schwaches Signal soll unter einer großen Menge an Rauschen gefunden werden. Daher werden in EEG/ERP-Experimenten oft um die 100 Messwiederholungen pro Bedingung angestrebt.

Versuchspersonenanzahl und Messwiederholungen

Kaum ein Effekt ist in der visuellen Aufmerksamkeitsforschung empirisch durch Replikationen so stark abgesichert, wie *Intertrial Priming*. Intertrial Priming bezeichnet (perzeptuelle) Bahnungseffekte, die von einem Durchgang in den nächsten stattfinden. Erstmals wurde dieser Effekt von Maljkovic und Nakayama (1994) beschrieben. Sie baten ihre Versuchspersonen, nach einem Zielreiz zu suchen, der sich in seiner Farbe von den Distraktoren unterschied. Ein Suchbildschirm konnte bei Maljkovic und Nakayama aus einem roten Zielreiz und zwei grünen Distraktoren bestehen oder aber auch aus einem grünen Zielreiz und zwei roten Distraktoren. Welche Farbe der Zielreiz und die Distraktoren in einem jeweiligen Durchgang genau hatten, konnte von Durchgang zu Durchgang wechseln und war für die Aufgabe völlig irrelevant. Der einzige relevante Umstand war, dass sich der Zielreiz in seiner Farbe von allen anderen Reizen unterschied.

Dadurch, dass sich die Zielreiz- und Distraktorfarben zufällig von Durchgang zu Durchgang ändern konnten, gab es geprimte Durchgänge (d. h. Durchgänge, in denen die Zielreiz- und Distraktorfarben zwischen zwei Durchgängen identisch blieben) und ungeprimte Durchgänge (d. h. Durchgänge, in denen die vorherige Zielreizfarbe nun die Distraktorfarbe und die frühere Distraktorfarbe nun die Zielreizfarbe war). Vielleicht nehmen Sie in weiser Voraussicht bereits das Ergebnis vorweg: Selbst wenn die exakte Zielreizfarbe per se nicht relevant für die Aufgabe war, waren die Versuchspersonen in jenen Durchgängen signifikant schneller, in denen sich die

Zielreizfarbe wiederholte, und langsamer, wenn sich die Zielreizfarbe zwischen zwei Durchgängen änderte. Das ist der sogenannte *Intertrial Priming Effekt.*

Das ist jetzt natürlich erstmal eine schöne Anekdote und der absolute Burner beim ersten Date – glaube ich, ich habe immer noch keine Feedbackkarten zurückgeschickt bekommen. Aber wozu dieser Exkurs? So spannend und robust der Effekt des Intertrial Primings auch ist, so unvorstellbar mag für manche die Stichprobe sein, die Maljkovic und Nakayama (1994) verwendeten: In den neun berichteten Experimenten nahmen jeweils zwei bis drei Versuchspersonen teil. Damit aber noch nicht genug: Vera Maljkovic und Ken Nakayama waren auch selbst in mehr als der Hälfte der Experimente Versuchspersonen. Bedeutet das, wir können den Daten aus den Experimenten nun nicht mehr glauben? Waren die ganzen Replikationen reine Glückssache oder sind Maljkovic und Nakayama schlicht so repräsentative Versuchspersonen?

Die Antwort auf diese Fragen ist vielfältig. Zunächst einmal werden massiv große Effekte auch schon bei weniger Versuchspersonen die statistische Signifikanz erreichen – auch wenn Maljkovic und Nakayama (1994) ihre Daten deskriptivstatistisch und nicht inferenzstatistisch präsentierten. Zusätzlich kommt die Güte von Daten nicht nur daher, wie viele Versuchspersonen getestet werden, sondern wie viele Messungen die einzelnen Versuchspersonen durchlaufen. Wie sah das bei Maljkovic und Nakayama aus? Ihre Versuchspersonen/sie durchliefen bis zu 2500 Messungen pro Experiment und lieferten somit sehr genaue Schätzungen. Zudem waren die Ergebnisse der Versuchspersonen sehr (!) ähnlich zueinander. Smith und Little (2018) argumentieren deshalb, dass man in gewissen Fragestellungen der Experimentalpsychologie eher präzise Messungen anstreben sollte, anstatt zig Versuchspersonen zu erheben. Werden die Versuchspersonen präzise gemessen, könnte man sie nämlich weniger als einzelne Datenpunkte in einer Varianzanalyse betrachten, sondern als „Replikationseinheiten“: Haben mehrere, penibelst genau gemessene Versuchspersonen dieselben Ergebnisse, dann haben wir einen Effekt in mehreren Versuchspersonen repliziert.

Ob Sie diese Sichtweise für gut befinden oder nicht, kann ich Ihnen natürlich nicht vorschreiben. Es ist jedoch wichtig, sich in Zeiten der Replikationskrise, die seit etwa 2015 in der Psychologie bekannt ist (Open Science Collaboration, 2015), darüber Gedanken zu machen, was Merkmale guter Forschung sind. Heute wird oft der Fokus auf ausreichend große Stichproben gelegt, die bei tatsächlich existierenden Effekten zu statistisch signifikanten Ergebnissen führen sollen (~ Power). Greift diese Sichtweise aber nicht etwas zu kurz? Drücken wir es etwas drastisch aus: Sind die Daten von

500 Versuchspersonen, die jeweils nur zweimal gemessen wurden, wirklich zuverlässiger als die Daten von 2 Versuchspersonen, von die jeweils 500-mal gemessen wurden?◀

2.6 Randomisierung und Balancierung

Die Randomisierung und Balancierung sind zwei sehr mächtige Kontrolltechniken die uns die Kontrolle der Sekundärvarianz erlauben. Zur Erinnerung: Die Sekundärvarianz ist jene Varianz in den Daten, die durch systematische Einflüsse verursacht werden, die jedoch nichts mit der Fragestellung zu tun haben. Stellen wir uns eine fiktive Forschungsfrage vor: Wir wollen herausfinden, wie schnell Versuchspersonen auf emotionale Wörter („fröhlich“ vs. „traurig“) reagieren, wenn sie in einem Gesicht eingebettet sind, die ebenfalls entweder fröhlich oder traurig sind (siehe etwa Kar et al., 2017). Wir haben also ein 2×2-Design mit den Faktoren Zielwort („fröhlich“ vs. „traurig“) und Gesicht (fröhlich oder traurig). Wir könnten diese Faktorstufenkombinationen jetzt auf viele unterschiedliche Arten präsentieren, zum Beispiel:

1. Im ersten Abschnittes des Experiments werden die randomisierten Zielwörter (also zufällig „fröhlich“ oder „traurig“) stets in einem traurigen Gesicht präsentiert, während sie im zweiten Abschnitt des Experiments stets in einem fröhlichen Gesicht gezeigt werden.
2. Die Emotion des Gesichtes, in dem die Zielwörter präsentiert werden, wird randomisiert, also zufällig variiert. In der ersten Hälfte der Durchgänge wird stets das Zielwort „fröhlich“, während in der zweiten Hälfte der Durchgänge durchgehend das Zielwort „traurig“ präsentiert wird.
3. Sowohl das Zielwort als auch die Emotion des Gesichtes werden in jedem Durchgang zufällig gezogen.

Für alle oben genannten Möglichkeiten könnte es valide Gründe geben. Haben Sie etwa den Verdacht, dass sich die Wirkung der Emotion des Hintergrundgesichtes erst langsam und über mehrere Durchgänge aufbaut, dann macht es Sinn, über mehrere Durchgänge hinweg die gleiche Emotion zu zeigen. Gleiches gilt für die zweite oben genannte Variante: Wollen Sie demonstrieren, dass Motorpriming unabhängig von Distraktoreigenschaften (Emotion des Gesichtes) ist, dann würden Sie eine Zeit lang stets nur eines der beiden Zielwörter präsentieren.

Stellen wir uns nun folgenden experimentellen Ablauf vor: Sie begrüßen Ihre Versuchspersonen und erklären ihnen die Aufgabe. Danach erledigen die Versuchspersonen für 10 min nur Durchgänge, in denen die Zielwörter in einem fröhlichen Gesicht eingebettet sind, und danach für 10 min nur Durchgänge, in denen die Zielwörter in einem traurigen Gesicht eingebettet sind. Sie testen nach diesem Schema 20 Versuchspersonen und werfen – panisch und voller Erwartungen – einen Blick in die Daten. Und da schau her, Versuchspersonen antworten generell signifikant schneller, wenn das Zielwort im traurigen Gesicht eingebettet ist. Fantastisch, yolo, dab …

Ihre Ergebnisse sind tatsächlich spannend. Ignorieren wir einmal, dass Kar et al. (2017) eine Erklärung dafür vorgeschlagen haben, und lassen Sie einmal Ihre Ergebnisse in *Nature* einreichen. Um meine Geschichte weiterspinnen zu können, nehmen wir zudem einmal an, dass die Arbeit den Schreibtisch des Editors bzw. der Editorin überlebt und an Reviewer:innen zur Begutachtung geschickt wird. Nach zwei Monaten erhalten Sie Rückmeldung von den Gutachterinnen und Gutachtern und nach anfänglicher Trunkenheit vor lauter Erfolg und Hoffnung bleibt Ihnen Ihr Lachen im Halse stecken. Was ist passiert?

Reviewer:in 1 findet alles schön und super und turbotoll. Reviewer:in 2 bekrittelt einiges, hält die Problemchen aber für lösbar. Reviewer:in 3 (es ist gefühlt immer die Nummer 3) stellt Sie jedoch vor ein unlösbares Problem: Wenn alle Versuchspersonen im ersten Block ein fröhliches Gesicht im Hintergrund sahen und im zweiten Block ein trauriges Gesicht, woher wissen Sie dann, dass die schnelleren Reaktionszeiten im zweiten Block auf den Einfluss der Emotion des Gesichtes zurückgehen, und nicht bloß darauf, dass die Versuchspersonen zuvor einen ganzen Block Zeit hatten, die Aufgabe zu üben und deshalb naturgemäß schneller wurden? So ungerne wir es zugeben, Reviewer:in 3 spricht damit einen wichtigen Punkt an. Übungseffekte *müssen* nicht immer auftreten (vgl. Experiment 1B in Theeuwes, 1992), aber *können*! Durch unser Design, die Emotionen des Gesichtes im Hintergrund geblockt darzubieten und noch dazu die Blockreihenfolge nicht zu variieren, sind der Grad an Übung und die Emotion des Gesichtes konfundiert, d. h., es ist nicht mehr möglich, die Quelle des Effektes klar zu bestimmen, weil zwei Faktoren stets zusammen auftreten (in unserem Beispiel: *Grad an Übung* und *Emotion des Gesichtes*).

Reviewer:in 3 ist im Gegensatz zu uns aber noch nicht fertig. Er fragt, wie wir denn sicher sein könnten, dass das fröhliche Gesicht im ersten Block die Versuchspersonen nicht so angesteckt hat, dass die miesepetrigen Gesichter im zweiten Block umso auffallender waren. Reviewer:in 3 spricht hier einen wesentlichen Punkt an, der durchaus oft vorkommt und teils schwerwiegende Folgen

haben kann: den **Carry-Over-Effekt.** Der Carry-Over-Effekt bezeichnet den Einfluss, den eine vorherige experimentelle Manipulation auf das Verhalten der Versuchsperson im darauffolgenden Block haben kann.

Der Einfluss der Blockreihenfolge auf Suchstrategien

Ein besonders eindrückliches Beispiel für einen solchen Carry-Over-Effekt in der visuellen Aufmerksamkeitsforschung konnten Leber und Egeth (2006) demonstrieren. Sie teilten Ihre Versuchspersonen in zwei Trainingsbedingungen ein: Eine Gruppe sollte in einem Suchbildschirm lediglich nach einer einzigartigen Form suchen (etwa ein grünes Quadrat unter grünen Kreisen), während eine andere Gruppe nach einer ganz spezifischen Form suchen musste (etwa ein grünes Quadrat unter grünen Kreisen, Dreiecken und anderen Polygonen). In dieser zweiten Bedingung reichte es also nicht mehr, nur nach der besonderen, einzigartigen Form zu suchen, wie es etwa in der ersten Bedingung der Fall war (engl. *singleton detection mode*), sondern die Versuchspersonen mussten nach der ganz spezifischen Form suchen, da die Distraktoren nicht mehr homogen geformt waren *(feature search mode).* In beiden Bedingungen tauchte in der Hälfte der Durchgänge ein roter Distraktor auf, der die Aufmerksamkeit ablenken kann (vgl. Additional Singleton Paradigma in Kap. 7). Es ist bereits seit Bacon und Egeth (1994) bekannt, dass dieser rote Distraktor die Aufmerksamkeit nur anzieht, wenn Versuchspersonen im Singleton Detection Mode suchen, nicht aber, wenn sie im Feature Search Mode suchen.

In einem zweiten Abschnitt ließen Leber und Egeth (2006) ihre Versuchspersonen dann nochmals nach einer einzigartigen Form suchen, die unter homogenen Distraktoren eingebettet war, unabhängig davon, ob die Versuchspersonen davor in der Singleton Detection Mode oder Feature Search Mode ihr Training absolvierten. Wie immer konnte in der Hälfte der Durchgänge ein irrelevanter Farbdistraktor auftauchen, der für die Aufgabe völlig irrelevant war. Interessanterweise zeigte sich jetzt, dass dieser Farbdistraktor nur bei jenen Versuchspersonen die Aufmerksamkeit anzog, die in der Trainingsphase die Singleton Detection geübt haben, nicht aber bei jenen, die zuvor die Feature Search Bedingung absolvierten. Dieser Carry-Over-Effekt zeigt eindrücklich, dass Vorerfahrung und Übung selbst solche Effekte beeinflussen können, die gemeinhin als automatisch gelten.◀

Wie könnten wir Reviewer:in 3 in einem Nachfolgeexperiment von unseren Ergebnissen überzeugen? Prinzipiell hätten wir hier zwei Möglichkeiten zur Auswahl:

1. Wir balancieren die Blockreihenfolge über die Versuchspersonen hinweg.
2. Wir randomisieren die Durchgänge vollständig, sodass jede Faktorstufenkombination in jedem Durchgang gleich wahrscheinlich ist.

In der ersten Variante führen wir eine neue Variable in unser Experiment ein, die für uns zwar inhaltlich nicht relevant ist, uns aber erlaubt, mögliche Einflüsse der Blockreihenfolge aufzudecken. Wir variieren zufällig, welche Versuchsperson welche Blockreihenfolge erledigt (entweder fröhlich/traurig oder traurig/fröhlich; siehe Infobox für Vorschläge). Nach der Datenerhebung können wir also die Blockreihenfolge als einen Zwischensubjektfaktor in unsere Analyse einfließen lassen, um so etwaige Effekte dieser Variable zu finden. Dieses Vorgehen verlangt natürlich auch nach ausreichend großen Stichproben in beiden Gruppen, um Unterschiede zwischen den Blockreihenfolgen zuverlässig aufspüren zu können – sofern es welche gibt.

Die wohl einfachste Variante für unser Problem wäre aber wohl, die Faktorstufenkombinationen komplett randomisiert zu präsentieren. Jede Versuchsperson hat so eine andere Abfolge an Durchgängen und mögliche Einflüsse von Reihenfolge-Effekten mitteln sich auf diese Art und Weise innerhalb einer Versuchsperson und zwischen den Versuchspersonen über eine ausreichend große Anzahl an Durchgängen heraus (z. B. 25 Durchgänge pro Faktorstufenkombination).

Randomisierung vs. Balancierung

Der Zufall ist – per Definition – unserer Kontrolle entzogen. Auch wenn die Wahrscheinlichkeit sehr gering ist, könnten wir beispielsweise bei einem Münzwurf fünf- oder zehnmal hintereinander „Zahl“ werfen. Nur eine ausreichend große Ziehung zufälliger Werte nähert sich daher dem Erwartungswert (beim Münzwurf: 50 %) an (vgl. Gesetz der großen Zahlen). Sollen Versuchspersonen also zwei Bedingungen durchlaufen, kann es sein, dass bei einer rein zufälligen Wahl der Reihenfolge der Bedingung 8 von 10 Versuchspersonen die Bedingung A vor der Bedingung B erledigen müssen.

Eine Alternative zur Randomisierung stellt daher die Balancierung dar: Versuchspersonen werden anhand eines unwillkürlichen Merkmals der einen oder anderen Blockreihenfolge zugewiesen. Wichtig ist dabei, dass das entscheidende unwillkürliche Merkmal in keiner Weise mit dem untersuchten Effekt korreliert. Ein mögliches Merkmal, das oft zur Wahl der Blockreihenfolge verwendet wird, ist die Versuchspersonenzahl (gerade oder ungerade).

Lernziele

Denken Sie an die zweite anfangs geschilderte Möglichkeit, unser fiktives Experiment durchzuführen: „Die Emotion des Gesichtes, in dem die Zielwörter präsentiert wird, wird randomisiert, also zufällig variiert. In der ersten Hälfte der Durchgänge wird stets das Zielwort ‚fröhlich', während in der zweiten Hälfte der Durchgänge durchgehend das Zielwort ‚traurig' präsentiert wird."

Zu welchen Problemen könnte es hier kommen und wie könnte man diese Probleme umgehen?