DLR-IB-FT-BS-2017-43

Der Mensch als herausfordernde Variable im Experiment

Interner Bericht

Autor: Helene Walter

Deutsches Zentrum
DLR für Luft- und Raumfahrt

Institutsbericht

DLR-IB-FT-BS-2017-43

Der Mensch als herausfordernde Variable im Experiment

H. Walter

Institut für Flugsystemtechnik (DLR FT)
Braunschweig

33 Seiten
7 Abbildungen
4 Tabellen
37 Literaturstellen

Stufe der Zugänglichkeit: II, intern und extern beschränkt zugänglich

Deutsches Zentrum für Luft- und Raumfahrt e.V.
Institut für Flugsystemtechnik
Abt. Flugdynamik und Simulation
Lilienthalplatz 7, D-38108 Braunschweig

Braunschweig, 10.03.2017

Unterschriften:

Institutsleiter:	Prof. Dr.-Ing. S. Levedag	
Abteilungsleiter:	Dr.-Ing. H. Duda	
Betreuer:	Dipl.-Ing. C. Seehof	
Verfasser:	Helene Walter	

Dieser Bericht soll einen Überblick darüber geben, was man beachten muss, wenn man Experimente und Messungen mit Menschen als Teilnehmer durchführt. Menschliche Wahrnehmung, Kognition und Verhaltensweisen sind verantwortlich dafür, dass Effekte auftreten können, die eine Messung verfälschen oder für ihren Zweck unbrauchbar machen. Die Psychologie bietet in Form ihrer Theorien Erklärungen für Phänomene, die bei Messungen mit Menschen auftreten können. Außerdem hat sie einen großen Erfahrungsschatz und eine eigene Disziplin in Form der psychologischen Methodenlehre für den Umgang mit diesen Phänomenen.

Um die Vorgehensweisen in der Psychologie besser zu verstehen, wird sich zunächst mit der Vorstellung von Wissenschaft allgemein beschäftigt, um dann zu erläutern, welche Aufgaben die Psychologie als Wissenschaft an sich stellt und auf welche Probleme sie dabei trifft. Es wird das Experiment als Methode erklärt und kurz veranschaulicht, welche Möglichkeiten es bietet und weshalb es für die Psychologie unerlässlich ist. Das Konzept der Gütekriterien soll dem Leser einen Eindruck davon geben, wie in psychologischen Messungen Qualität sichergestellt wird. Daraufhin beginnt der Hauptteil dieses Berichts, die Beschäftigung mit Einflüssen auf die Wahrnehmung von Menschen in Messungen und Experimenten. Diese werden anhand von Beispielen zunächst allgemein dargestellt, woraufhin im folgenden Abschnitt die kognitiven Mechanismen erläutert werden, die diese Einflüsse erlauben. Das Ende des Berichts bildet eine Liste von Effekten, die psychologische Messungen verzerren können, zusammen mit Maßnahmen, die diese Verzerrungen verhindern.

Inhaltsverzeichnis

1. Wissenschaft

Bevor sich der Psychologie zugewendet werden soll, ist es hilfreich, einen Schritt zurückzutreten, um sich mit der Wissenschaft im Allgemeinen zu beschäftigen.
In der Wissenschaftsphilosophie von Thomas Kuhn sind neue Erkenntnisse untrennbar von bestehenden wissenschaftlichen Theorien. Sie entwickeln sich aus dem herrschenden System und werden von dessen Denkweise geleitet (Kuhn, 1970). Außerdem beschreibt Thomas Kuhn in seinem Werk „Die Struktur wissenschaftlicher Revolutionen“, dass in der Wissenschaft Wissen nicht gesammelt wird, sondern dass eine bestehende Theorie ersetzt wird, sobald ein Phänomen nicht mehr damit erklärt werden kann. Dieser Prozess nennt sich wissenschaftliche Revolution.
Sonst sind sich die Forscher in der „normalen Wissenschaft“ über die wichtigen Themengebiete und Fragestellungen, die erklärenden Theorien, sinnvolle Methoden und die zu benutzenden Begriffe einig. In dieser Einigkeit besteht laut Kuhn der Kern der Wissenschaft. Er nennt dies das herrschende *Paradigma* (Sedlmeier & Renkewitz, 2013). Sie haben ein Konstrukt erstellt, das die Fragen ihres Themenbereichs in einem etablierten Vokabular und mit den gewählten Methoden beantwortet. Tritt eine Anomalie auf, also etwas, das die aktuelle Sichtweise nicht erklären kann, wird versucht, das Gerüst noch durch Abänderungen aufrechtzuerhalten, sodass die Anomalie darin eingepasst werden kann. Gelingt dies trotz aller Bemühungen nicht, werden verschiedene neue Theorien gebildet, die nun um den freiwerdenden Platz des herrschenden Paradigmas konkurrieren. Letztendlich wird die Theorie gewählt, die die beste Erklärung liefert. Allerdings kann man nicht sagen, dass das neue Paradigma, das alte vollständig in sich enthält. Kuhn nannte das *Inkommensurabilität*: Die zwei Paradigmen können sich so weit unterscheiden, dass sogar die gleichen Worte nicht mehr das Gleiche meinen. Dies war zum Beispiel beim Wechseln vom geozentrischen zum heliozentrischen Weltbild der Fall. Der Begriff Planet wurde vor der kopernikanischen Wende auch für die Sonne genutzt, was in unserem heutigen Paradigma falsch wäre.

Wissenschaft ist laut dieser Vorstellung also ein ständiges Bemühen um eine bessere Vorstellung von der Welt. Wer von endgültigen Wahrheiten spricht, überschätzt laut Kuhns Vorstellung die Bestrebungen der Wissenschaft. Dass es, zumindest zu dieser Zeit, nicht möglich ist, absolute Aussagen zu machen, wird sichtbar, wenn man sich mit einem typischen Problem der Psychologe beschäftigt. Dieses Problem soll im nächsten Abschnitt genauer betrachtet werden, in dem zunächst die Wissenschaft Psychologie definiert wird.

1.1 Die Wissenschaft Psychologie

Psychologie wird als die Wissenschaft vom Verhalten und Erleben des Menschen bezeichnet (Hussy, Schreier & Echterhoff, 2010). Im Gegensatz zum Verhalten ist das Erleben etwas, das unsichtbar bleibt. Ein bekannter Ausspruch zum Geheimnis des Bewusstseins ist der Titel eines Aufsatzes des Philosophen Thomas Nagel: „What is it like to be a bat?“ (1974). Jedes Lebewesen erlebt in sich und allein. Das Erleben zu messen gestaltet sich deswegen ziemlich schwer. Variablen, die im Verborgenen bleiben und nicht so einfach gemessenen werden können, wie zum Beispiel die Länge eines Holzstücks, nennt man *latente Variablen*. Letztlich führen psychologische Fragestellungen immer dann zu latenten Variablen, wenn man die zugrunde liegenden inneren Prozesse eines Verhaltens herausfinden möchte. Es ist also wichtig, sich mit diesen latenten Variablen zu beschäftigen, bevor psychologisch geforscht wird.

1.1.1 Latente Variablen

Dass hinter dem Beobachtbaren etwas unsichtbares Lenkendes steht, ist eine sehr alte Vorstellung (Bollen, 2002). In der Psychologie sind latente Variablen interessant, da man bei der Beobachtung von menschlichem Verhalten auch nach Ursachen für dieses Verhalten sucht. Der Mensch reagiert nicht passiv auf äußere Reize, sondern verarbeitet diese. Reaktionen entstehen somit aus einem Zusammenspiel zwischen Reiz und Verarbeitung. Ein Beweis dafür ist, dass zwei verschiedene Menschen auf den gleichen Reiz auf verschiedene Weise reagieren können. In einer Notfallsituation reagiert ein Pilot schneller und angemessener als ein anderer. Eine Hypothese könnte lauten, dass der Pilot, der besser reagiert hat, ein höheres Situationsbewusstsein hatte. In diesem Beispiel wären die Reaktionen der Piloten das von außen sichtbare Verhalten. Das Situationsbewusstsein ist eine Variable, die das Erleben der Piloten beschreibt, also eine latente Variable.
Auf latente Variablen gibt es verschiedene Sichtweisen. Eine der häufigsten Fragen ist, ob es latente Variablen, wie Situationsbewusstsein, wirklich gibt.
Laut Loevinger (1957) gibt es Konstrukte (also latente Variablen) und Eigenschaften von Menschen. Dies sind aber zwei verschiedene Dinge. Die menschlichen Eigenschaften bestehen in der „wirklichen" Welt, Konstrukte werden von Wissenschaftlern im Bestreben, diese Eigenschaften greifbarer zu machen, erfunden. Das würde bedeuten, dass es die Einheit „Situationsbewusstsein" in der wirkliche Welt nicht gibt. Dazu passt auch eine andere Definition von latenten Variablen: Sie werden von Wissenschaftlern geschaffen, um beobachtete Phänomene zusammenzufassen und zu erklären (Bollen, 2002). Hierzu wird das statistische Instrument der *Faktorenanalyse* genutzt.

Grundlagen einer Faktorenanalyse sind *manifeste Variablen*. Deshalb werden zunächst Daten in Messungen gesammelt. Manifeste Variablen sind also im Gegensatz zu latenten Variablen durch Empirie zugänglich. Ziel einer Faktorenanalyse ist das Finden von den latenten Variablen (hier werde sie *Faktoren* genannt; Sedlmeier & Renkewitz, 2013). Auch bei der Faktorenanalyse hängt es von der gewählten Sichtweise ab, ob man sagt, dass die latenten Variablen nur abstrakte, von Wissenschaftlern ausgedachte Begriffe sind, die die manifesten Variablen zusammenfassen, oder ob man davon ausgeht, dass die wirklich existierenden latenten Variablen die gemessenen manifesten Variablen verursachen. In Faktorenanalysen untersucht man durch Korrelationen vermutete Zusammenhänge der gemessenen Daten untereinander und schließt auf dieser Grundlage auf die latenten Variablen.
Ohne sich im Detail mit den Rechenprozeduren der Faktorenanalyse zu beschäftigen, wird jedoch klar, dass das Konzept der latenten Variable von den manifesten Variablen abhängt, die zuvor gemessen wurden. Ohne ermittelte Daten, die erklärt werden sollen, ist die Idee von „Situationsbewusstsein" unklar und nutzlos. Hier wird eine der wichtigsten Aufgaben in der psychologischen Forschung deutlich: Operationale Definitionen. Aufgrund der Eigenschaft latenter Variablen, nicht direkt beobachtbare Merkmale zu beschreiben, herrscht keine Einigkeit darüber, ob sie tatsächlich existieren oder nur von Wissenschaftlern konstruiert sind. Jedoch wollen wir weiterhin Begriffe wie Situationsbewusstsein oder auch Intelligenz benutzen, da sie nützlich sind. Das können sie allerdings nur bleiben, wenn sie gut definiert sind. Ein bekannter Ausspruch von Edwin Boring ist seine „Definition" von Intelligenz: „Intelligenz ist das, was der Intelligenztest misst." (Boring, 1923) Diese Formulierung erscheint zunächst ironisch und inhaltslos. Allerdings ist genau dies – Messungen mithilfe von Intelligenztest – die einzige Möglichkeit und Art, Intelligenz zu erfassen. Wenn man genau beschreibt, mithilfe welcher manifesten Variablen eine latente Variable gemessen werden kann, stellt man sicher, dass alle Forscher, die sich an diese Anweisungen halten, vom gleichen Konstrukt sprechen. Dies macht Forschungsergebnisse vergleichbar. Außerdem: Da latente Variablen die manifeste Welt besser begreifbar machen sollen, ist eine Annäherung über ebendiese manifeste Welt sinnvoll. Aus diesem Grund sind operationale Definitionen, also die genaue Anweisung, wie eine latente Variable messbar gemacht wird, Grundlage in der Psychologie und Teil jeder Beschreibung einer Studie. Operationale Definitionen umgehen die Frage, ob z. B. Intelligenz tatsächlich existiert,

indem von vornherein gesagt wird, dass ein bestimmtes Messmodell ein Merkmal misst, das in diesem Modell den Namen „Intelligenz“ trägt.
In Kuhns Vorstellung von Wissenschaft sind absolute Definitionen von latenten Variablen nicht nötig, da Forschung sowieso ein Prozess ist, der seine Ansichten ständig erneuert und widerruft. Latente Variablen sind wie Paradigmen der Wissenschaft Stützen, die für einige Zeit die Grundlage für weiteren Erkenntnisgewinn bilden.

2. Zusammenhänge zwischen Variablen

Bleiben wir weiterhin beim Beispiel vom Reaktionsverhalten von Piloten in einem Notfall. Natürlich möchte man wissen, wie diese Reaktion mit bestimmten anderen Bedingungen zusammenhängt. Eine weitere Hypothese könnte lauten: Piloten, die im Notfall von Nebenaufgaben abgelenkt werden, zeigen ein größeres Fehlverhalten. Der vermutete Zusammenhang besteht in diesem Fall zwischen Ablenkung und dem situationsgerechten Verhalten. Führt man einfach nur eine Messung durch und beobachtet, dass mehr Ablenkung und schlechteres Verhalten meist zusammenauftreten, kann man nur aufgrund dieser Korrelation nicht sagen, dass das beobachtete Verhalten durch die Ablenkung verursacht wird. Für Kausalaussagen benötigt man Experimente (Sedlmeier & Renkewitz, 2013).

2.1 Die Logik des Experiments

Der wichtigste Vorteil von Experimenten ist, dass nur sie Kausalaussagen ermöglichen. Kausalität liegt vor, wenn ein Merkmal A (und zwar *nur* ein Merkmal A) verändert wird und sich daraufhin *regelhaft* das Merkmal B verändert. In diesem Fall sagt man, dass das Merkmal A das Merkmal B kausal beeinflusst (Eggert, 2014).
Überträgt man diese Logik auf das Experiment, so heißt das: Eine Variable wird *manipuliert* (die unabhängige Variable UV) und alle anderen Variablen, die Einfluss auf die Experimentalsituation nehmen können (Störvariablen) werden *kontrolliert*. Verändert sich in der Folge die Ausprägung der abhängigen Variablen (AV), so sagt man, dass die UV die AV kausal beeinflusst.
Diese Definition beinhaltet die drei Bedingungen für Kausalschlüsse (Sedlmeier & Renkewitz, 2013):

- Kovariation – Es besteht ein Zusammenhang zwischen A und B,
- Zeitliche Präzedenz - A tritt zeitlich vor B auf,
- Ausschluss von Alternativerklärungen (andere Ursache als A).

Für unser Beispiel bedeutet dies, dass (1), wenn Piloten abgelenkt werden, immer auch ihr Reaktionsvermögen schlechter sein muss, dass außerdem (2) zunächst die Ablenkung durch andere Aufgaben und dann die Verschlechterung in der Aufgabenausführung stattfinden muss und dass (3) keine weiteren Einflüsse als Erklärung für die Leistungsverschlechterung möglich sein dürfen.

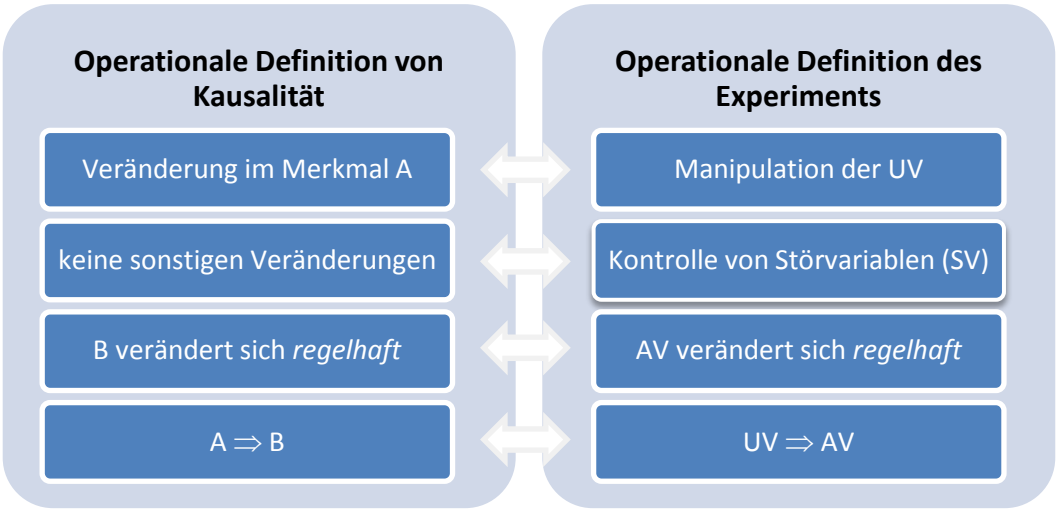

Abbildung 1. Aus den Merkmalen der Kausalität resultierende Merkmale eines Experiments.

2.1.1 Experimentelles Design

Weil man prüfen möchte, ob eine Kovariation zwischen der UV und der AV besteht, ändert man die Ausprägung der UV und beobachtet, ob sich daraufhin auch die Ausprägung der AV ändert. Die Veränderung der Ausprägung der UV nennt man *Manipulation*. Die verschiedenen Ausprägungen, die eine UV im Experiment einnimmt, heißen *Stufen* (Sedlmeier & Renkewitz, 2013). Führen wir für unser Beispiel ein Experiment durch, können wir als unabhängige Variable „Ablenkung durch eine andere Aufgabe“ und als abhängige Variable „Reaktion auf Problemsituation“ festlegen. Oft hat die UV in Experimenten nur zwei Stufen: Die Experimental- und Kontrollbedingung. In der Kontrollbedingung ist die UV nicht anwesend. Wenn die Hypothese lautet, dass andere Aufgaben die Leistung in einer Problemsituation mindern, dann ist in unserem Experiment die Experimentalbedingung ein Szenario mit Ablenkungsaufgabe und die Kontrollbedingung eines ohne.

Möchte man in einem Experiment die Einflüsse von mehr als einer Variable prüfen, dann kann man die Stufen der verschiedenen UVs auch kombinieren. Die resultierenden Kombinationen nennt man *Bedingungen* (Sedlmeier & Renkewitz, 2013).

Tabelle 1. *Kombination der Stufen von zwei unabhängigen Variablen*

	Stufen der UV1	
Stufen der UV2	Stufe A	Stufe B
Stufe A	Bedingung 1_A2_A	Bedingung 1_B2_A
Stufe B	Bedingung 1_A2_B	Bedingung 1_B2_B

In unserem Beispiel könnte man als zweite UV noch die Variable „Müdigkeit“ prüfen, wie in der Tabelle 2 dargestellt.

Tabelle 2. *Reaktionszeiten von Piloten auf Gefahrensituationen unter den Bedingungen „Ablenkung“ und „Müdigkeit“*

	UV1: Ablenkung	
UV2: Müdigkeit	mit Ablenkung	ohne Ablenkung
müde	4	3
ausgeruht	2	2

Anmerkungen. In diesem fiktiven Beispiel wurde die Reaktionszeit (in Sekunden) von Probanden bei einem bei der Landung in die Bahn rollenden Hindernis gemessen.

Experimente mit mehreren UVs erlauben nicht nur die Prüfung von Haupteffekten, sondern auch von Interaktionen. Ein *Haupteffekt* bedeutet, dass die Variation der Ausprägung einer Variable einen Einfluss auf die Ausprägung der abhängigen Variable hat. Eine *Interaktion* hingegen bedeutet, dass die Ausprägung einer Variable bestimmt, ob eine andere Variable einen Einfluss auf die abhängige Variable hat (Sedlmeier & Renkewitz, 2013).
Zum Beispiel kann es sein, dass nur Piloten, die müde sind, im Falle einer Ablenkung eine längere Reaktionszeit haben.

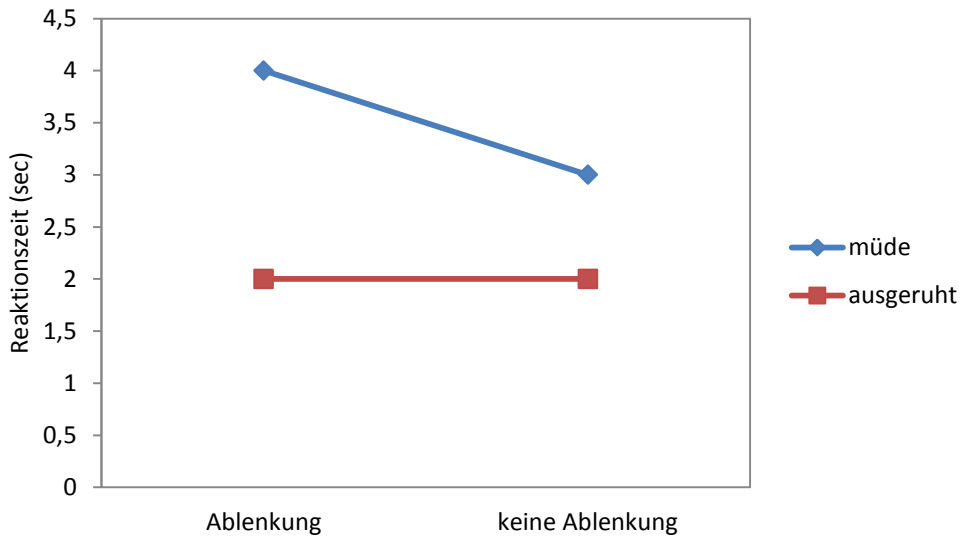

Abbildung 2. Reaktionszeiten von Piloten auf Gefahrensituationen unter den Bedingungen „Ablenkung“ und „Müdigkeit“.

In dem Graphen erkennt man neben der Interaktion auch den Haupteffekt der UV2 „Müdigkeit“: Müde Piloten benötigen immer mehr Zeit für eine Reaktion, als ausgeruhte Piloten.

2.1.2 Störvariablen

Wie oben aufgeführt, gehört zu den Bedingungen für einen Kausalschluss, dass andere Ursachen neben der vermuteten – also des Reizes, der in Form der UV untersucht wird – ausgeschlossen werden müssen. Variablen, die Alternativerklärungen für beobachtete Effekte sein können, nennt man *Störvariablen (SV)* (Sedlmeier & Renkewitz, 2013).

In einem Experiment soll die Leistungsfähigkeit von Piloten in Abhängigkeit vom Einsatz eines neuen Displays geprüft werden.

UV: Einsatz des Displays

AV: Leistung

Eine Störvariable könnte die Uhrzeit des Experiments sein.

Problematisch wäre es, wenn alle Experimente *ohne* Display kurz vor dem Mittagessen stattfinden und alle Experimente *mit* dem Display danach. Ergibt das Experiment tatsächlich, dass die Piloten in der Versuchsbedingung mit Display höhere Leistungen erbringen, kann man sich nicht sicher sein, ob dies an der Hilfe des Displays liegt oder ob die Probanden der anderen Bedingung einfach durch ihren Hunger eingeschränkt waren. Allerdings gibt es Kontrollmethoden für Störvariablen, die man nutzen kann, um diese Unsicherheit zu vermeiden (siehe Abschnitt 2.1.2.1).

Oft wird zwischen personalen und situativen Störvariablen unterschieden. Bei *situativen Störvariablen* befinden sich in der Experimentalumgebung Reize, die verhindern, dass man eindeutig sagen kann, dass ein Kausalzusammenhang zwischen UV und AV herrscht (Sedlmeier & Renkewitz, 2013).

So können in einem Experiment die Lichtverhältnisse eine situative Störvariable darstellen. Wenn alle Versuchsdurchgänge der Experimentalbedingung „mit Ablenkung“ in einer dunkleren Umgebung stattfinden, als die der Kontrollbedingung „ohne Ablenkung“, kann ein besseres Verhalten der Piloten in der Kontrollbedingung nicht mehr ausschließlich auf die Bedingung „ohne Ablenkung“ der UV zurückgeführt werden. Vielleicht war es in der Experimentalbedingung zu dunkel, um die Checklisten für Notfallsituationen zu lesen.

Bei *personalen Störvariablen* können verschiedene Eigenschaften von Probanden eine Veränderung der Ausprägung der AV verursachen (Sedlmeier & Renkewitz, 2013).

Zum Beispiel nehmen an einem Experiment sowohl Piloten teil, die bereits viele Erfahrungen haben, aber auch junge Piloten. Die personale Störvariable „Erfahrung“ kann mit der AV „situationsgerechtes Verhalten“ zusammenhängen. So weiß man nicht mehr, ob Veränderungen im Verhalten der Piloten nur auf die experimentelle Manipulation der Variable „Ablenkung“ zurückzuführen sind, oder ob möglicherweise in einer Experimentalbedingung mehr erfahrene Piloten sind, als in der anderen, weshalb in einer Gruppe die Leistung höher ist.

2.1.2.1 Störvariablenkontrolle

Am Beispiel der personalen Störvariablen sieht man, dass es nicht immer möglich ist, alle Störvariablen auszuschließen. Selbst wenn man alle situativen Störvariablen eliminieren könnte (im Fall von Störgeräuschen kann man durch eine Schallisolierung den Experimentalkontext vollkommen still halten), bringen Probanden immer unterschiedliche Eigenschaften und Vorerfahrungen mit.

Dennoch gibt es Möglichkeiten, Störvariablen entgegenzutreten. Für verschiedene Formen von Störvariablen gibt es unterschiedlich gut passende Kontrollmöglichkeiten. Diese werden nun aufgelistet.

Randomisieren

Personale Störvariablen

Die Versuchspersonen werden zufällig auf die verschiedenen Bedingungen des Experiments zugeteilt. Die Störvariable soll so über die verschiedenen Bedingungen hinweg durchschnittlich den gleichen Wert aufweisen (Sedlmeier & Renkewitz, 2013).

Situative Störvariablen

Zur Kontrolle von Störvariablen in der Versuchssituation können natürlich auch alle anderen Faktoren des Experiments zufällig auf die Bedingungen aufgeteilt werden. Stehen für ein Experiment zwei verschiedene Räume zur Verfügung, könnte man durch Randomisieren sicherstellen, dass sich die Ausprägungen von Störvariablen wie Lautstärke, Helligkeit oder Luftqualität, die sich in den Räumen unterscheiden, gleichmäßig auf die Bedingungen verteilen.

Parallelisieren
Die Störvariable, die kontrolliert werden soll, wird so über die Bedingungen des Experiments aufgeteilt, dass ihre verschiedenen Ausprägungen in jeder Bedingung gleich häufig vorkommen. Bei personalen Störvariablen muss natürlich die Ausprägung der Störvariablen vor dem Experiment bei den Versuchspersonen gemessen werden. (Sedlmeier & Renkewitz, 2013).

Konstanthalten
Alle Merkmale der Untersuchungssituation und der Probanden haben in allen Bedingungen die gleiche Ausprägung (Sedlmeier & Renkewitz, 2013). Beispielsweise nutzt man in den meisten Tierversuchen nur männliche Versuchstiere, mit der Begründung, dass weibliche Tiere aufgrund des Hormonzyklus schwankende Ergebnisse liefern (Zucker & Beery, 2010). Eine spezielle Form des Konstanthaltens ist das *Eliminieren* der Störvariablen. Bei der Variable „Lärm" wäre das das Ausschließen von allen Geräuschen (Sedlmeier & Renkewitz, 2013).

Zum Parallelisieren und Konstanthalten muss die Störvariable bekannt sein, die kontrolliert werden soll. Hierin liegt der große Vorteil des Randomisierens: Durch eine zufällige Zuteilung der Probanden zu den Bedingungen werden immer alle personengebundenen Störvariablen gleichzeitig kontrolliert und sie müssen nicht einmal bekannt sein (Sedlmeier & Renkewitz, 2013). Deswegen sollte man seine Probanden *immer* zufällig auf die Bedingungen verteilen, sofern dies möglich ist.

Die Tabelle 3 listet Vor- und Nachteile der drei Kontrolltechniken auf. Darauf basierend kann man die passende Kontrolltechnik für den Gegenstand seiner Messung wählen.

Tabelle 3. *Vor- und Nachteile verschiedener Kontrolltechniken für Störvariablen*

Kontrollmethode	Vorteile	Nachteile
Randomisieren	– Alle Störvariablen werden kontrolliert – Störvariablen müssen nicht bekannt sein – Sollte immer angewandt werden	– Erhöhung der Fehlervarianz[1] – Ausreichend große Stichprobe nötig (je heterogener die Population, aus der gezogen wird, desto größer muss die Stichprobe sein)
Parallelisieren	– Bei kleiner Stichprobe Alternative zum Randomisieren	– Erhöhung der Fehlervarianz – Störvariable muss bekannt sein (allerdings kennt man nie alle!) – Störvariable muss messbar sein – Bei situationaler Störvariable: – SV muss manipulierbar sein – Nur eine (oder kleine Zahl von) Störvariablen wird kontrolliert
Konstanthalten	– Keine Erhöhung der Fehlervarianz[1]	– Schränkt Übertragbarkeit ein – SV muss bekannt, messbar und manipulierbar sein

[1] Die Fehlervarianz ist die Varianz, in den erhobenen Daten, die nicht durch den untersuchten Effekt erklärt werden kann.
Anders formuliert: Fehlervarianz + systematische Varianz (ergibt sich durch die unterschiedlichen experimentellen Bedingungen) = Gesamtvarianz

3. Gütekriterien des Messens

Latente Variablen bilden den Dreh- und Angelpunkt psychologischer Forschung. Operationalisierungen ermöglichen uns einen Zugang zu ihnen. Es ist also wichtig, dass man sich auf diese Messanweisungen verlassen kann. Aus diesem Grund gibt es Bewertungsmaßstäbe für Messinstrumente.

Der „DLR-Test“ soll die psychologische Eignung von Piloten untersuchen und dient als Instrument für das Auswahlverfahren vieler Fluggesellschaften (DLR - Institut für Luft- und Raumfahrtmedizin - Luft- und Raumfahrtpsychologie, 2017). Dieser Test entscheidet also, in wessen Hände viele Menschenleben und teure Flugzeuge gegeben werden. Es wäre also gut, wenn dieses Prüfverfahren auch vertrauenswürdig ist. Anhand des DLR-Tests sollen die Gütekriterien nun erklärt werden.

3.1 Objektivität

Eine Messung sollte unabhängig davon sein, wer diese durchführt (Sedlmeier & Renkewitz, 2013). Wenn also zehn verschiedene Versuchsleiter mit der gleichen Person beim DLR-Test zum gleichen Ergebnis kommen, dann ist der DLR-Test objektiv.
Man unterscheidet drei verschiedene Arten von Objektivität:

Durchführungsobjektivität
Die Durchführungsobjektivität beschreibt, ob das Verhalten des Versuchsleiters während der Testung einen Einfluss auf die Ergebnisse hat (Sedlmeier & Renkewitz, 2013).
Um sie zu gewährleisten, wird die Durchführung von Experimenten und Tests möglichst maximal standardisiert. Dazu gehört, dass die Instruktionen – ob auf Papier oder mündlich – immer die gleichen sind. Auch sollten sich die Versuchsleiter den Versuchspersonen gegenüber immer gleich verhalten und jede weitere Interaktion zwischen Teilnehmer und Versuchsleiter sollte verhindert werden, weshalb Instruktionen für jeden verständlich gestaltet werden sollten (Sedlmeier & Renkewitz, 2013).
In unserem Beispiel, dem DLR-Test, kann man die Durchführungsobjektivität also durch einfache und standardisierte Instruktionen sichern.

Auswertungsobjektivität
Die Auswertungsobjektivität ist dann gegeben, wenn gleiches Verhalten des Probanden immer nach exakt den gleichen Regeln in ein „empirisches Relativ“ übersetzt wird.
Der Proband zeigt ein Verhalten oder macht in einem Fragebogen bestimmte Angaben, dies muss allerdings noch vergleichbar gemacht werden. Die Form, die die Daten einer Person annehmen, damit sie mit Daten anderer Personen verglichen werden können, nennt man Rohwerte, sie bilden das empirische Relativ. (Schmidt-Atzert & Amelang, 2012)
Bekommen die Kandidaten beim Auswahlerfahren für Piloten die Aufgabe, eine Gruppendiskussion zu führen, so muss ein geeignetes Instrument gefunden werden, das das Verhalten der Probanden in Daten übersetzbar macht. Perfekte Auswertungsobjektivität ist gegeben, wenn gleiches Verhalten immer gleich übersetzt wird.

Interpretationsobjektivität
Hat man gezeigten Verhalten oder Angaben in einem Fragebogen in Rohwerte übersetzt, muss diesen noch eine Bedeutung gegeben werden. Dies nennt man Interpretation. Die Interpretation ist objektiv, wenn gleiche Rohwerte immer in die gleichen Aussagen über den Probanden umformuliert werden. Dies kann sichergestellt werden, wenn es Angaben gibt, welches Merkmal durch welche Aufgaben

gemessen wird (*Benennung des Merkmals*) und welche Rohwerte welche Ausprägung dieses Merkmals darstellen (*Benennung der Ausprägung*; Schmidt-Atzert & Amelang, 2012)
Wird beim DLR-Test mithilfe eines Fragebogens die Persönlichkeit gemessen, so muss klar sein, welche Fragen, welche Eigenschaft messen. Kreuzt jemand bei der Frage „Gehen Sie gerne von sich auf andere Menschen zu?“ die Antwortmöglichkeit „eher nicht“ an, so muss im sogenannten Testmanual nachgeschaut werden, dass die Frage zu der Gruppe von Fragen gehört, die „Offenheit messen“, in welchen Rohwert „eher nicht“ übersetzt werden soll und was der Rohwert aussagt. In den meisten Fragebögen gibt es mehrere Test-Items (Fragen oder Aufgaben) für ein Merkmal, um die Reliabilität (siehe unten) zu erhöhen. Deshalb müssen die Rohwerte einzelner Items meist addiert werden und können erst dann in eine Aussage umgewandelt, also interpretiert werden.

3.2 Reliabilität

Die Reliabilität beschreibt, wie genau ein Test ein Merkmal erfasst. Je genauer die Messung ist, desto weniger wird sie von unsystematischen Fehlern (Messfehlern) gestört. Es gibt verschiedene Schätzgrößen für die Reliabilität, die allerdings nicht alle das gleiche angeben. Die Kennwerte geben deshalb nicht immer die gleichen Ergebnisse und sind somit nicht austauschbar. (Schmidt-Atzert & Amelang, 2012)

Retest-Reliabilität
Eine Möglichkeit, die Reliabilität zu schätzen, ist die Messwiederholung. Ein Test wird ein zweites (oder weiteres) Mal mit der gleichen Stichprobe durchgeführt. Die Korrelation zwischen den zwei Durchgängen gibt die Retest-Reliabilität an. Besonders interessant ist die Retest-Reliabilität, wenn durch die Testung eine Prognose getroffen werden soll. Denn man kann eine Retest-Reliabilität nutzen, für die genauso viel Zeit zwischen den Messungen lag, wie durch den Test in die Zukunft vorhergesagt werden soll. Natürlich ist die Retest-Reliabilität auch von der zeitlichen Stabilität eines Merkmals abhängig. Zwischen den Messungen kann sich ein Merkmal, zum Beispiel Wissen oder eine Fähigkeit, ändern. (Schmidt-Atzert & Amelang, 2012)

Paralleltestreliabilität
Man gibt Probanden parallele Versionen von Tests und prüft die Korrelation zwischen den zwei Versionen. Es ist allerdings schwierig, tatsächlich parallele Tests zu konstruieren, die trotz unterschiedlicher Fragen oder Aufgaben das gleiche Merkmal messen. (Schmidt-Atzert & Amelang, 2012)

Testhalbierungsreliabilität
Zwei Hälften eines Tests, die von einer Person durchgeführt wurde, werden miteinander korreliert. Wenn die Items der beiden Hälften vergleichbar sind, kommen die Ergebnisse diese Methode der Idee der Reliabilität am nächsten, da die Motivation der Teilnehmer und die Veränderbarkeit der Eigenschaft über die Zeit keinen Einfluss auf die Messung nehmen können. (Schmidt-Atzert & Amelang, 2012)

3.3 Validität

Wenn ein Test valide ist, dann misst er tatsächlich auch das, was er vorgibt zu messen. Schmidt-Atzert und Amelang (2012) beschreiben zudem, dass es sich um ein Urteil darüber handelt, „wie angemessen bestimmte Schlussfolgerungen vom Testwert auf das Verhalten außerhalb des Tests oder auf ein Merkmal der Person sind.“
Somit ist die Validität also das wichtigste Gütekriterium. Wenn der DLR-Test nicht valide wäre, bedeutete dies, dass aller Aufwand unnötig ist. Man könnte auch zufällig Personen auswählen, die

Piloten werden sollen, da die Testung keine Aussagen über die Eignung der Kandidaten machen würde. (Schmidt-Atzert & Amelang, 2012)

Inhaltsvalidität

Inhaltsvalidität gibt an, wie repräsentativ die Items eines Tests für das zu messende Merkmal sind (Schmidt-Atzert & Amelang, 2012).

Beispielweise sind in einem Test, der fliegerische Fähigkeiten prüft, Fragen die die Nutzung von Kupplung, Gas- und Bremspedal betreffen, unpassend.

Kriteriumsvalidität

Bei der Kriteriumsvalidität vergleicht man das Testergebnis mit Verhalten oder Leistungen außerhalb der Testsituation, die mit dem gemessenen Merkmal zusammenhängen. Man kann die Leistungen von bereits arbeitenden Piloten mit ihren alten Ergebnissen im DLR-Test oder in Einstellungstests vergleichen und hat so ein Indiz für die Kriteriumsvalidität des DLR-Tests oder ebendieser Einstellungstests. Ein häufiges Problem ist, dass eine bestimmte Hälfte der Population aus der Stichprobe ausgeschlossen wird: Immer diejenigen Kandidaten, die die Testungen nicht bestanden haben. Jemand, der beim DLR-Test ein zu niedriges Ergebnis erzielt hat, wird nicht Pilot. Man weiß nicht, ob der Test nicht möglicherweise zu streng ist, da man niemals die beruflichen Leistungen dieser aussortierten „Probanden“ messen kann. (Schmidt-Atzert & Amelang, 2012)

Konstruktvalidität

Schmidt-Atzert und Amelang (2012) definieren Konstruktvalidität als „empirischen Beleg dafür, dass ein Test das Konstrukt erfasst, welches er erfassen soll“. Zu einem Konstrukt gibt es meist eine definierte Vorstellung. So muss ein Pilot, bei dem die Fähigkeit „Situationsbewusstsein“ hoch ausgeprägt ist, sich auf bestimmte Weise verhalten. Definitionen von Konstrukten enthalten also immer Verhaltenskomponenten. Ein Test, der konstruktvalide ist, muss genau diese Verhaltensweisen erfassen.

3.4 Zusammenhang zwischen den Hauptgütekriterien

Diese drei Hauptgütekriterien bauen aufeinander auf. Wenn ein Test nicht objektiv ist, dann kann er auch nicht reliabel sein. Wenn verschiedene Testleiter verschiedene Ergebnisse des Tests verursachen können, bedeutet dies, dass ein Test nicht bei jeder Messung zu jeder Zeit bei der gleichen Testperson das gleiche Ergebnis liefert. Sobald mehrere Messungen mit dem gleichen Test bei der gleichen Versuchsperson nicht das gleiche Ergebnis liefern, kann man natürlich auch nicht mehr behaupten, dass der Test valide ist.

Ein Test kann zwar objektiv und reliabel sein, muss aber deshalb nicht unbedingt eine hohe Validität besitzen. Sobald er jedoch valide ist, ist er immer auch objektiv und reliabel.

4. Einflüsse auf die Wahrnehmung von Menschen in Experimenten

In diesem Abschnitt wird sich nun genauer mit den Variablen beschäftigt, die die Messungen mit Menschen zu einer schwierigeren Aufgabe machen, als das Anlegen eines Maßbands an einen Holzbalken.

4. 1 Der Experimentalkontext

An den oben beschriebenen Kontrollmethoden sieht man, dass ein großer Aufwand betrieben wird, um den Einfluss von Störvariablen zu verhindern. Nur diesem kontrollierten Kontext ist zu verdanken, dass Experimente Kausalaussagen ermöglichen. Oft ist die Rede von einer *Isolation* des zu interessierenden Reizes (Sedlmeier & Renkewitz, 2013). Dies bedeutet, dass bestimmte Situationen ohne ihren komplizierten Kontext aus dem Alltag in eine Laborsituation übertragen werden, um ihre Zusammenhänge zu untersuchen. Die Verhaltenspsychologie legt jedoch nahe, dass die Wirkung eines Reizes durch seinen Kontext bestimmt wird. Dies kann man „Verhaltenskontrolle durch Kontextreize“ nennen (Domjan & Grau, 2003)).
Analog zum Paul Watzlawicks erstem Axiom der Kommunikation („Man kann nicht nicht kommunizieren, denn jede Kommunikation (nicht nur mit Worten) ist Verhalten und genauso wie man sich nicht nicht verhalten kann, kann man nicht nicht kommunizieren.“; Watzlawick, Beavin & Jackson, 1990) kann man sagen, dass es nicht möglich ist, einen Reiz ohne Kontext wahrzunehmen: Alles ist Kontext. Es gibt nicht keinen Kontext.
Für das Experiment bedeutet das, dass auch die von allen Reizen befreite Laborsituation einen Kontext darstellt. Was bedeutet dies für unser Ziel, Kausalzusammenhänge zu untersuchen?
Ein eindrucksvolles Beispiel für den Einfluss des Kontexts auf die Wahrnehmung von Reizen ist ein Experiment von Orne und Evans (1965). Probanden sollten mit bloßen Händen giftige Schlangen anfassen und Münzen aus schäumender Salpetersäure fischen – und taten dies auch. Später erklärten die Probanden, dass sie an ihren Schutz durch die Versuchsleiter glaubten. Hier sorgt der „Sicherheitskontext“ dafür, dass die Probanden sich anders verhalten haben, als sie es sonst tun würden. Der Reiz „ätzende Säure“ war also tatsächlich nicht isoliert, sondern eingebettet in eine Situation von Überwachung und Verantwortung (des Versuchsleiter den Teilnehmern gegenüber) und somit auch Sicherheit.
Ein anderes Experiment, auch von Gustaf Orne, zeigt ebenfalls, wie wichtig es ist, den Kontext einer Situation zu beachten: Bei Messung der Elektrodermale Aktivität[3] eines Probanden steigt diese üblicherweise dann an, wenn sie lügen. In einem Experiment haben Ellson, Davis, Saltzman und Burke (1952) herausgefunden, dass Probanden schwerer beim Lügen zu entlarven sind, wenn ihnen gesagt wurde, dass sie den „Lügendetektor“ nicht täuschen können, während Probanden leichter zu interpretierende physiologische Daten liefern, wenn sie denken, dass es ihnen in vorherigen Durchgängen gelungen ist, den Lügendetektor zu täuschen. Diese Ergebnisse waren sehr überraschend, da man üblicherweise in Lügendetektortests immer betont, dass der Lügendetektor unfehlbar ist und alle Lügen sichtbar macht. Aufgrund der neuen Ergebnisse müsste man diese alte Strategie jedoch plötzlich ändern. Gustafson und Orne wollten das Problem genauer untersuchen.
Als Erstes fanden sie in einer Replikation durch Fragebögen heraus, dass alle Probanden glaubten, ein Lügendetektor würde mit „normalen“ Menschen funktionieren, während notorische Lügner schwer zu entlarven seien.
In einem weiteren Experiment (Gustafson & Orne, 1965) wurden die Probanden in zwei Gruppen aufgeteilt: Eine wurde mit den folgenden Worten auf den Versuch vorbereitet: „Wir wollen testen, wie gut der Lügendetektor funktioniert. Wie Sie wissen, ist es im Falle von psychopathischen Persönlichkeiten oder notorischen Lügnern nicht möglich, Lügen aufzudecken. Wir wollen Sie bitten, in diesem Experiment ihr Bestmögliches zu geben, den Lügendetektor zu täuschen.“ Diese Bedingung ähnelt der im Versuch von Ellson et al., da die Probanden so die gleiche Vorstellung über die Täuschbarkeit von Lügendetektoren haben. Sie denken, dass nur Menschen mit schlechten Eigenschaften einen Lügendetektor täuschen können. Die Probanden wollen natürlich nicht, dass jemand von ihnen denkt, dass sie Psychopathen seien und sie nahmen mit der Erwartung und Hoffnung, vom Lügendetektor entlarvt zu werden, an der Studie teil. Die zweite Gruppe erhielt folgende Instruktionen: „Dies ist eine Lügendetektor-Studie. Es ist zwar sehr schwer, einen Lügendetektor zu täuschen, besonders intelligenten, emotional stabilen und reifen Personen gelingt dies allerdings.“ Hier wird die Vorstellung gezielt verändert. Die Probanden denken, ein gutes Bild von

sich zu geben, wenn sie den Lügendetektor täuschen. Den Lügendetektor zu täuschen, wird zum Ziel der Probanden.
Im tatsächlichen Versuch wurde den Versuchspersonen eine Zahl zwischen 1 und 9 vorgegeben, die sie sich merken sollten. Dann wurden ihnen der Reihe nach alle neun möglichen Zahlen präsentiert, wobei ihre elektrodermale Aktivität gemessen wurde. Die Probanden sollten über jede Zahl sagen, dass es nicht ihre sei, also einmal lügen. Nach dem ersten Lügendetektor-Durchgang wurden beide Gruppen wieder zweigeteilt: Es wurde ihnen entweder gesagt, dass ihre Lüge entdeckt wurde, indem ihnen ihre wahre Zahl präsentiert wurde, oder man präsentierte eine falsche Zahl, um ihnen zu zeigen, dass sie den Lügendetektor täuschen konnten. (Dies war möglich, da die Versuchsleiter durch heimliche Überwachung die Zahlen der Probanden kannten.)
Tabelle 4 stellt in einer Vierfeldertafel den weiteren Verlauf der Studie in Abhängigkeit von der Gruppenzugehörigkeit dar:

Tabelle 4. *Unterschiede zwischen den Bedingungen des Experiments*

	UV1	
UV2	Information: Täuschung misslungen	Information: Täuschung gelungen
Täuschungsfähigkeit „gut“	leicht zu entdecken	schwer zu entdecken
Täuschungsfähigkeit „schlecht“	schwer zu entdecken	leicht zu entdecken

Anmerkungen. Schwierigkeit, die elektrodermale Aktivität der Probanden zu messen, in Abhängigkeit von den Informationen, die die Probanden zu Beginn der Studie erhalten haben und der Rückmeldung zu ihrer Lügefähigkeit.

Das interessante Ergebnis der Studie ist, dass Probanden, deren Wunsch erfüllt wurde, schwerer zu entdecken waren, egal, um welche Art von Wunsch es sich handelt: durch geschicktes Lügen einen guten Eindruck zu machen, oder durch ungeschicktes Lügen zu zeigen, dass man kein Psychopath oder notorischer Lügner ist. Waren die Probanden frustriert, da sie in einer Situation waren, die sie schlecht dastehen ließ, waren sie auch schlechtere Lügner.

Auch in diesem Beispiel kann man eine Art Kontext für die unterschiedlichen Verhaltensweisen der Probanden verantwortlich machen. Dieser Kontext besteht aus den Werten und Vorstellungen, die die Probanden haben. Im Experiment von Ellson et al. brachten die Probanden selbst die Vorstellung (ihren „eignen Kontext“) mit, dass nur kriminelle Personen, einen Lügendetektor täuschen können und urteilten deshalb über die Situation, dass sie einen besseren Eindruck hinterlassen, wenn sie vom Lügendetektor beim Lügen erkannt werden. Diese Erwartungen und Werte wurden in Gustafsons und Ornes Experiment gezielt durch die Informationen in der Instruktion manipuliert. Die Probanden erhielten unterschiedliche Informationen hinsichtlich der Bedeutung ihrer Lügefähigkeit und dieser Unterschied führte zu zwei gegensätzlichen Effekten. Bliebe die Störvariable „Glaube über Lügendetektoren“ unentdeckt, könnte man fatale Folgen aus dem ersten Experiment von Ellson et al. ziehen. Dank Gustafson & Ornes Studie wissen wir aber, dass Befragte, die nicht durchschaut werden wollen, leichter zu durchschauen sind, wenn sie denken, dass sie ihr Ziel *nicht* erreichen.

4.2 Soziale Erwünschtheit und Ideale der Probanden

Ein weiterer Einflussfaktor auf Experimente wird in Gustafson & Ornes Studie sichtbar: Der Einfluss von Experimentalbedingungen ist von den Motiven der Probanden abhängig. Ein Proband kann möglichst gute Ergebnisse erzielen wollen, um intelligent oder leistungsfähig zu wirken, er kann aber auch durchschnittliche Daten produzieren wollen, wenn er einen gesunden und „unbescholtenen“

Eindruck machen möchte. Diesen Effekt der *sozialen Erwünschtheit,* also das Bedürfnis, vor anderen gut dazustehen (Sedlmeier, & Renkewitz, 2013).
Selbst wenn sich Menschen nicht beobachtet fühlen oder keine Bewertung durch andere erwarten, überwachen sie ihr eigenes Verhalten, denn sie wollen ihrem Selbstbild entsprechend handeln (Fries & Grawe, 2006). Das bedeutet, dass man durch Gewährleistung von Anonymität und die Abwesenheit eines Versuchsleiters diese Effekte nur mindern, niemals ausschalten kann. Umso mehr sollte man sich bei der Planung eines Experiments Gedanken darüber machen, welche subtilen Hinweise für „gutes" Verhalten die Untersuchungssituation liefert und ob diese Einfluss auf den zu untersuchenden Effekt nehmen.

4.3 Unbewusstes Verhalten

Durch den Begriff soziale Erwünschtheit sollte man sich nicht täuschen lassen und denken, dass Probanden nur ihr Verhalten bewusst ändern. Natürlich können sich Menschen bewusst verstellen, wenn sie meinen, dadurch vorteilhafter zu erscheinen. Allerdings sieht man an dem Lügendetektor-Experiment, dass Kontextreize oft wirken, ohne dass der Mensch es verhindern kann. Die Elektrodermale Aktivität ist eine Körperreaktion, die nicht kontrollierbar ist.

4.4 Demand characteristics

Orne führte seine Experimente durch, um ein Konzept zu veranschaulichen, das er selbst entwickelt hat: Die *demand characteristics*. Dies sind Hinweise im experimentellen Kontext, die der Proband nutzt, um zu erraten, was von ihm erwartet wird (Sedlmeier, & Renkewitz, 2013). Situationen in psychologischen Versuchen sind meist künstlich, vereinfacht und eher „kontextarm". Das führt dazu, dass unklar ist, welche Bedeutung ein Reiz hat. Umso einfacher ist es dann für eigentlich kleine und vielleicht zufällige Reize, die Kontrolle über das Verhalten der Probanden zu gewinnen. Sie geben dem Reiz einen Sinn.

4.4.1 Verhaltenskontrolle durch einen kleinen Reiz

Das folgende Experiment soll darstellen, dass selbst eine unauffällige Information, ein unauffälliger Reiz, die Wahrnehmung einer Situation verändern kann.
Loftus und Palmer (1974) wollten in einem Experiment darstellen, wie die Art der Zeugenbefragung einen Einfluss darauf nehmen kann, welche Angaben die Zeugen machen werden. Sie zeigten Probanden ein Video von einem Unfall, bei dem zwei Autos ineinanderfuhren. Daraufhin sollten die Probanden beurteilen, wie schnell die Autos waren. Es stellte sich heraus, dass Personen, denen die Frage „Wie schnell waren die Autos, als sie zusammen*krachten*?" höhere Schätzungen abgaben, als Probanden, die die Formulierungen „kollidierten", „stießen" oder „in Kontakt kamen" hörten. Außerdem wurden die Probanden eine Woche später gefragt, ob Glassplitter auf dem Boden lagen. Die „krachten-„Gruppe beantwortete diese Frage eher mit „ja", obwohl in dem Video kein zerbrochenes Glas zu sehen war.
Das Experiment zeigt, dass Reize nicht nur Verhalten, sondern tatsächlich auch die Wahrnehmung von Menschen beeinflussen. Dies sieht man sehr schön daran, dass selbst die Erinnerung an die Situation verändert war, und das zu einem Zeitpunkt, zu dem die Probanden wahrscheinlich nicht mehr wiedergeben konnten, welches Verb in der Frage im Versuch benutzt wurde.

4.5 Geist + Realität = Wahrnehmung

Die Suche des Menschen nach demand characteristics, die eine Situation leichter interpretierbar machen, zeigt, dass in einem Experiment immer eine Wechselwirkung zwischen dem menschlichen Geist und dem objektiv vorhandenen Kontext besteht. Der Geist jedes Menschen besteht aus unterschiedlichen Erfahrungen und den darauf basierenden Werten und Erwartungen. Man kann sich vorstellen, dass allen Reizen, die auf den Menschen treffen, zunächst durch diesen persönlichen Filter eine bestimmte Bedeutung verliehen wird, bevor sie die Wahrnehmung des Menschen bilden. Aufgrund persönlicher Unterschiede in der Erfahrung entsteht das individuelle Erleben aller Menschen und letztlich ihr unterschiedliches Verhalten in objektiv gleichen Situationen. Gustafsons und Ornes Lügendetektorexperiment (1965) zeigt diesen Einfluss des bewertenden Filters sehr schön: Probanden, die Lügefähigkeit als etwas Gutes ansahen, zeigten ganz andere Ergebnisse als Probanden, die Lügefähigkeit für etwas Schlechtes hielten.
Menschen unterscheiden sich also stark von anderen „Variablen“ einer Messung. Ein Holzbrett kann einem Reiz ausgeliefert werden, zum Beispiel Druck, und reagiert daraufhin auf eine für Holzbretter typische Weise – es bricht. Ein Mensch lässt Reize allerdings nicht nur auf sich wirken, sondern verarbeitet sie auch. Daraus ergeben sich viele Komplikationen für Experimente. Versteht man die Wirkprinzipien des menschlichen Geistes, kann man allerdings mit diesen Effekten umgehen, da sie vorhersehbarer werden.

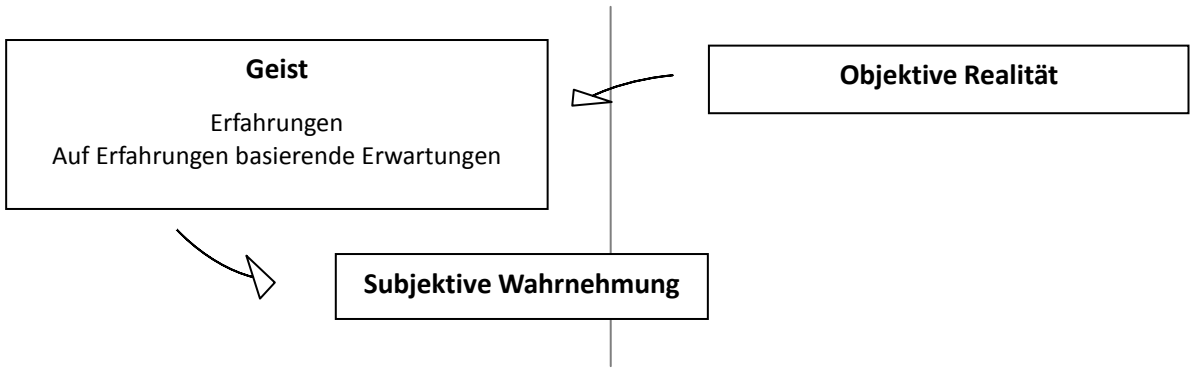

Abbildung 3. Zusammenhang zwischen der „objektiven“ Realität (wenn man davon ausgeht, dass es sie gibt), der geistigen Verarbeitung des Menschen und seiner subjektiven Wahrnehmung. Man sieht, dass sich die subjektive Wahrnehmung nur zum Teil mit dem Bereich der objektiven Realität überschneidet.

5. Kognitive Mechanismen

Für alle Tiere und Menschen ist es wichtig, eine Situation schnell erfassen und bewerten zu können, um die sicherste und vorteilhafteste Reaktion zu zeigen. Der Mensch hat einige kognitive (also geistige) Mechanismen, die ihm das Leben erleichtern, allerdings auch dafür sorgen, dass wir niemals die Welt so wahrnehmen können, wie sie tatsächlich ist, sondern „nur“ eine für schnelle Schlussfolgerungen und Reaktionen optimierte Version.

5.1 Der Mensch als Wissenschaftler – Kellys persönliche Konstrukte

Eine Erklärung für eine Vielzahl von Effekt steckt in der von George A. Kelly eingeführten Formulierung „Der Mensch als Wissenschaftler“. Wieso ist der Begriff „Mensch als Wissenschaftler“ passend? Die Ziele der psychologischen Wissenschaften werden meist wie folgt formuliert: Menschliches Erleben und Verhalten soll **erklärt**, **vorhergesagt** und **kontrolliert** werden können (Hussy, Schreier, & Echterhoff, 2010). Der Wissenschaftler unterscheidet sich dabei überhaupt nicht

von allen anderen Menschen, vielmehr nimmt er diese ihm angeborenen Eigenschaft in seinen Beruf mit. Denn für Kelly strebt jeder Mensch danach, Kontrolle über sich und seine Umwelt zu haben, um sicherer zu sein, also länger und gesünder zu leben. In einer Zusammenfassung seines Wissens, "Personal construct theory and the psychotherapeutic interview" (Kelly, 1977), schreibt er: "A person lives his life by reaching out for what comes next and the only channels he has for reaching are the personal constructions he is able to place upon what may actually be happening." ("Eine Person lebt ihr Leben, indem sie danach reicht, was als nächstes passiert, und der einzige Weg, den sie hat, um danach zu greifen, sind die persönlichen Konstrukte, die sie in der Lage ist, über das drüberzustülpen, was möglicherweise tatsächlich passieren wird." (Übers. d. Verf.) In diesem Zitat sieht man, was Kellys größter Verdienst in der Psychologie war: Er beschreibt die *persönlichen Konstrukte*, die später die Grundlage der „kognitiven Wende" sein sollten. In den Begriff der persönlichen Konstrukte hat Kelly seine Annahme verpackt, dass die Dinge, die wir wahrnehmen, keine objektiven Repräsentationen der Welt sind, sondern selbst konstruierte Vorstellungen darüber, welche Zusammenhänge herrschen. Seine weitreichende Schlussfolgerung ist, dass es keinen wirklichen „Realismus" geben kann und dieser nur als eine Art Dogma existiert, das den Menschen einengt und ihm Angst macht – genauso wie auch die naturwissenschaftliche Suche nach absoluter Wahrheit.[2] Deswegen plädierte Kelly dafür, zu akzeptieren, dass wir unsere Realität, also unsere Wahrnehmung, selbst schaffen und somit selbst der Ursprung unserer Gefühle und unseres Verhaltens sind. Kelly nutzte das Konzept der persönlichen Konstrukte vor allem als Therapeut, um zu verstehen, woher psychische Probleme rühren. Der oben erwähnte Artikel "Personal construct theory and the psychotherapeutic interview" (Kelly, 1977) bietet einen schönen Einblick in seine Sichtweise und den daraus resultierenden Therapieansatz.

5.1.1 Die Entwicklung eines Konstrukts

Der Mensch baut sich seine Konstrukte, indem er sich wiederholende Muster erkennt und sich diese merkt. Diese sich wiederholenden Muster bilden die Grundlage für Erwartungen. (Genauso nimmt man in der Forschung an, dass Ereignisse, die in der Vergangenheit am häufigsten vorgekommen sind, der beste Schätzer für zukünftige Ereignisse sind; Sedlmeier & Renkewitz, 2013). Konstrukte sind außerdem plastisch, sie werden laufend durch Erfahrungen angepasst. Die Konstrukte eines Menschen sind eine Persönlichkeitseigenschaft. Sie bestimmen nämlich, wie wir die Welt wahrnehmen. Sie sind sozusagen der Filter, der unsere Sichtweise formt, auf deren Grundlage wir wiederum handeln. Hierher rührt auch Kellys Vorstellung, dass es keinen Realismus geben kann. Schließlich versucht jeder Mensch mit seinem eigenen, „selbstgebauten" Filter, Eindrücke zu sortieren. Das Wort *sortieren* beinhaltet immer eine Wertung, ein System, nach dem Objekte einem Platz zugewiesen werden. Genauso steckt – gebildet durch unsere Erfahrungen – in unseren Köpfen ein System, das Sinn und Zusammenhang zwischen den Reizen unserer Umwelt herstellt. Eine objektive Betrachtung der Welt ist also nicht möglich.

Um das am eigenen Leib zu erfahren, muss man nur die zwei unteren Spalten lesen:

KATZE	**BEIN**
MAUS	**ARM**
VOGEL	**KOPF**
HAMSTER	**SCHULTER**
FISCH	**KNIE**
H?ND	**H?ND**

[2] Hier sieht man einen Zusammenhang zu der anfangs erwähnten Theorie von Kuhn (siehe Abschnitt 1). Auch wissenschaftliche Entdeckungen entstehen aus einem Paradigma heraus (Kuhn, 1970), sind also niemals frei von einem Kontext. Ohne dazugehöriges Theoriekonstrukt verlieren wissenschaftliche Vorstellungen ihren Wert.

Beim Lesen der ersten Worte erkennen wir ein Muster: Es handelt sich offensichtlich um Tiere oder Körperteile. Dies beeinflusst, was wir von den schwer zu entziffernden letzten Worten erwarten. Die Erwartung „hilft“ uns, das Rätsel zu lösen.

5.2 Kognitive Schemata

Tatsächlich zeigt das obere Beispiel eine hoch geschätzte Fähigkeit in unserer Gesellschaft: Mustererkennung und die Fähigkeit zu Abstrahieren. Eine Erklärung dafür bietet die Theorie der kognitiven Schemata.
Die Idee der kognitiven Schemata ist aus der Computerwissenschaft und Forschung zur Künstlichen Intelligenz entnommen. Schemata werden genutzt, um zu erklären, wie Menschen konzeptuelles Wissen speichern. Konzeptuelles Wissen stellt man sich als ein reiches Geflecht an Wissen vor, das Vorstellungen, die wir über die Ordnung der Welt haben, wiederspiegelt. Diese *Konzepte* gewinnen wir über Erfahrungen. Anderson (2007) beschreibt dies so: „Beispielsweise können wir von spezifischen Erfahrungen auf generelle Kategorisierungen der Merkmale der jeweiligen Erfahrungsklasse abstrahieren. Eine derartige Abstraktion bringt konzeptuelles Wissen hervor, das Kategorien beinhaltet; wie beispielsweise Stühle und Hunde. Sobald wir derartige Kategorien geschaffen haben, können wir diese zur abstrakten Repräsentation spezifischer Erfahrungen nutzen. Anstatt uns beispielsweise nur daran zu erinnern, dass wir von einem vierbeinigen pelzigen Objekt abgeleckt wurden, das um die 50 Pfund wog und einen wedelnden Schwanz hatte, erinnern wir uns daran, dass wir von einem Hund abgeleckt wurden.“
Menschen sind in der Lage, in einzelnen (vielleicht sogar lückenhaften) Reizen einen Zusammenhang zu erkennen, da sie wissen, dass diese häufig gemeinsam unterhalb eines im Schema höhergelegenen Konzeptes liegen. Auch wenn nicht alle Reize vorhanden sind, die nötig wären, um ein Konzept vollständig darzustellen, wird das wahrscheinlichste Konzept als Interpretation gewählt.

Ein anderes Beispiel dafür ist das folgende Bild:

Abbildung 4. Das Bild besteht aus einigen schwarzen Flecken, die zusammen jedoch von den meisten Menschen sofort als Dalmatiner erkannt werden. (Aus: Randommization, 2017).

Es fehlen noch sehr viele Reize, die dieses Bild zu einem tatsächlichen Dalmatiner machen würden, allerdings macht die Art der Ansammlung dieser Punkte es am wahrscheinlichsten, dass es sich um einen Dalmatiner handelt. Unser Leben ist voller solcher Auslassungen (Selbst den Baum draußen vor dem Fenster sieht man nicht vollständig von oben bis unten!) und nur durch die allgemeine Fähigkeit, in losen Reizen übergeordnete Kategorien zu erkennen, ermöglicht normales Leben.

Brewer und Treyens (1981) zeigten in einem Experiment, wie kognitive Schemata uns täuschen können. Sie baten Probanden, kurz (35 Sekunden) in einem Raum zu warten. Sie sagten ihnen, dass es sich um das Büro des Versuchsleiters handle und sie noch nicht ins Labor gehen könnten, weil nachgesehen werden muss, ob der vorherige Proband bereits fertig sei. Im Labor angekommen, sollten die Probanden dann alle Gegenstände aufschreiben, an die sie sich aus dem Büro des Versuchsleiters noch erinnern konnten. Brewer und Treyens vermuteten, dass die Probanden sich an

Dinge erinnern würden, die typisch für ein Büro sind. Wieder liefern kognitive Schemata die Erklärung: Die meisten Menschen besitzen ein Schema für das Konzept „Büro“, in dem sie bestimmte dazugehörige Gegenstände abgespeichert haben. Da die Probanden erfuhren, dass es sich bei dem Raum um ein Arbeitszimmer handelt, müssten sie sich seine Ausstattung auch unter dieser Annahme in Erinnerung rufen.

Abbildung 5. Raum, in dem die Versuchspersonen von Brewer und Treyens 35 Sekunden warten mussten. (Aus Brewer & Treyens, 1981.)

Brewer und Treyens Hypothese wurde bestätigt. Die meisten Versuchspersonen notierten „Stuhl“ und „Schreibtisch“. Viel weniger erinnerten sich an einen anatomischen Schädel. Es wurden sogar Objekte des Büro-Schemas aufgelistet, die sich gar nicht in dem Raum befunden haben, wie etwa ein Kaffeebecher und Stifte (Brewer & Treyens, 1981).

Auch in diesem Beispiel haben die Informationen, die die Versuchsleiter gegeben haben, ein kognitives Schema bei den Probanden aktiviert. Sie schufen einen Kontext für die Wahrnehmung der Probanden. An diesem kleinen Beispiel wird deutlich, wie relativ die Realität ist, die uns alltäglich eigentlich eindeutig erscheint.

6. Umgang mit dem unvermeidlichen Kontext

Bereits Gustafsons und Ornes Lügendetektor-Experiment (1965) zeigt, dass ein Kontexte bereits wirken kann, auch wenn der Versuchsleiter sich dessen nicht bewusst ist und meint, ein Experiment durchzuführen, in dem alle Reize abgesehen von der UV eliminiert wurden. Im Fall des Lügendetektor-Experiments haben die Menschen die Situation auf Grundlage ihrer eigenen Einstellungen bewertet. Ihrer Meinung nach war es schlecht, ein guter Lügner zu sein. Wäre diese Einstellung, also dieser „unsichtbare Kontext“ niemals erfragt worden, hätte man vollkommen falsche Schlüsse aus den Ergebnissen des Experiments gezogen. Erst in einem weiteren Experiment, in dem ein künstlicher Kontext geschaffen wurde (die Fähigkeit, einen Lügendetektor täuschen zu können, wurde offen als „gut“ oder „schlecht“ dargestellt) erlaubte es, den Kontext zu kontrollieren und den wahren Zusammenhang zwischen Frustration und Lügefähigkeit zu messen.

Hier zeigt sich, dass es nicht immer sinnvoll ist, Experimente mit einem möglichst neutralen, „leeren“ Kontext, man kann auch sagen „sterilen Laborbedingungen“ durchzuführen. Die Probanden werden dann umso sensibler für jeden Reiz, der auch nur zufällig im Experiment erscheinen kann (wie etwa die Wortwahl des Versuchsleiters oder seinen Gemütszustand, den er leider nicht immer

kontrollieren kann) und ziehen daraus Informationen über die Lage, die wiederum ihre *Wahrnehmung* (Beispiel: Wahrnehmung der Geschwindigkeit im Experiment von Loftus und Palmer) und ihr *bewusstes und unbewusstes Verhalten* (Beispiel: Elektrodermale Aktivität der Probanden in Gustafsons und Ornes Lügendetektor-Experiment) beeinflussen.
Jeder Kontext kann also eine Störvariable darstellen. Sofern man nicht weiß, wie Probanden eine Situation wahrnehmen und welcher Kontext wirkt, kann man versuchen, ihn zu erfassen oder selbst zu bestimmen. Orne hat den Kontext durch Befragungen herausgefunden: „Normale Menschen" wollen nicht lügen können. Daraufhin hat er einen Kontext geschaffen, der die Störvariable „Einstellung gegenüber Lügefähigkeit" kontrolliert. Man kann also eine Störvariable kontrollieren, indem man sie in einem Experiment als weitere unabhängige Variable hinzufügt.
Selbst wenn man sich nicht sicher ist, welcher Kontext in einem Experiment herrscht, kann man ihn einfach durch gezielte Anweisungen bestimmen. So stellt man immerhin sicher, dass alle Probanden die gleichen Informationen haben und die Situation gleich interpretieren.

7. Effekte, die psychologische Messungen verzerren

Hier sollen nun Effekte in Experimenten aufgelistet werden, die sich aus den kognitiven Prozessen der Probanden ergeben. Außerdem werden Möglichkeiten angeboten, die verhindern sollen, dass diese Effekte die Messungen verzerren.

7.1 Erwartungseffekte

7.1.1 Der Klassiker: Der Placebo-Effekt

„Bei einem Placebo handelt es sich um eine Scheinsubstanz oder -behandlung, die die positiven Wirkungen eines Arzneimittels oder einer Intervention nachzuahmen vermag, ohne aber dabei das spezifische Arzneimittel oder die spezifische Behandlungsmethode zu beinhalten", definiert die DFG-Forschergruppe für Placeboforschung (Placebo & Nocebo, 2017). Menschen können sich bei einer Krankheit also besser fühlen, wenn sie denken, ein Medikament einzunehmen, obwohl die Substanz keinen Wirkstoff enthält.
Dafür gibt es verschiedene Erklärungen. Eine davon ist die *Erwartungshaltung*. Hier wirkt also wieder der menschliche Geist: In Erwartung an eine Besserung tritt auch eine Besserung ein. Betont werden muss dabei, dass der Placebo-Effekt nicht nur subjektive Bewertungen betrifft (etwa Angaben in einem Wohlbefindens-Fragebogen). Es werden auch objektive Besserungen des Zustandes beobachtet, wenn auf Probanden ein Placebo-Reiz wirkt. Umgekehrt gibt es allerdings auch einen Nocebo-Effekt: Haben Menschen die Erwartung, ein Medikament oder ein anderer Reiz könnte ihnen schaden, dann lassen sich negative Veränderungen messen (Placebo & Nocebo, 2017)

Kontrolle des Placebo-Effekts
Der Placebo-Effekt ist besonders aus der Medizin bekannt. Bei der Testung von Medikamenten möchte man natürlich wissen, inwieweit der Wirkstoff einen Effekt verursacht. Daher ist der Placebo-Effekt eine Störvariable. Um diese zu kontrollieren, werden sogenannte **Doppelblindstudien** durchgeführt (Sedlmeier & Renkewitz, 2013). „Blind" nennt man diese Studien, weil die Probanden nicht wissen, zu welcher von zwei Gruppen sie gehören. Die, die das tatsächliche Medikament verabreicht bekommt, oder die Gruppe mit dem Placebo, also der Scheinsubstanz- oder behandlung. Es kann dank des Placebo-Effekts durchaus passieren, dass in beiden Gruppen Verbesserungen auftreten. Aus den Unterschieden zwischen den Gruppen kann man dann aber auf den tatsächlichen Effekt des Wirkstoffes schließen.

„Doppelt blind“ ist eine Studie dann, wenn auch alle restlichen am Experiment beteiligten Personen nicht wissen, wer zu welcher Versuchsgruppe gehört. Dies soll den sogenannten Rosenthal-Effekt verhindern (siehe Abschnitt 7.1.3).

7.1.2 Der Hawthorne-Effekt

Roethlisberger und Dickson (1964) versuchten in einer Fabrik (der Hawthorne-Fabrik des Elektronikunternehmens Western Electric) herauszufinden, wie man die Leistung von Arbeitern erhöhen kann. In einem der Experimente wurden in der Experimentalbedingung die Lichtverhältnisse verbessert, während in der Kontrollgruppe alles so blieb, wie es war. Wie man vermutet hatte, veränderte sich die Produktivität der Arbeiter in der Experimentalgruppe. Allerdings waren die Veränderungen unsystematisch, mal arbeiteten die Arbeiter mehr, als in der Kontrollgruppe, mal weniger. In einer Gruppe verrichteten die Probanden sogar weiterhin mehr Arbeit, als die Beleuchtungsstärke wieder auf das Niveau der Kontrollbedingung verringert wurde (Roethlisberger & Dickson, 2003).
Um diese unklaren Ergebnisse zu erklären, wurde ein weiteres Experiment durchgeführt. Zunächst wurde jeden Tag das Licht etwas heller eingestellt. Dies gefiel den Probanden. In einer Phase danach sahen sie, wie jeden Tag die Glühbirnen gegen vermeintlich stärkere Glühbirnen ausgetauscht wurden. Tatsächlich wurden immer wieder Birnen der gleichen Stärke eingedreht. In den Befragungen gaben die Arbeiter trotzdem an, die Arbeit „bei mehr Licht“ sei Tag für Tag angenehmer. Sie ließen sich täuschen, nahmen also helleres Licht wahr und waren dabei zufriedener.
Dann wurde der Arbeitsplatz wieder jeden Tag verdunkelt. Nach einiger Zeit kam wieder nur noch ein Handwerker, um Glühbirnen auszuwechseln und den Eindruck zu erwecken, dass weiterhin die Lichtstärke verringert wird. Trotzdem gaben die Probanden weiterhin an, dass das abnehmende Licht nicht so angenehm beim Arbeiten sei. Interessant ist: Die Produktivität der Probanden blieb in allen Phasen des Experiments gleich (Roethlisberger & Dickson, 2003).
Schon die ersten Lichtexperimente zeigen, dass in den Hawthorne-Studien nicht nur die Lichtintensität die gemessene Arbeitsleistung und –zufriedenheit beeinflusst hat. In Anbetracht des zweiten Versuches wird klar, dass auch „psychologische Einflüsse“ gewirkt haben müssen.
Mittlerweile hat sich in der Psychologie der Begriff „Hawthorne-Effekt“ für die Fälle eingebürgert, in denen Effekte in einem Experiment zustande kommen, weil die Testpersonen wissen, dass sie Teil einer experimentellen Manipulation sind (Lück, 2009 oder Schreier & Odağ, 2010). Sie erwarten also, dass sich etwas verändern muss und dass bei ihnen auch eine bestimmte Reaktion darauf stattfinden wird.

7.1.3 Der Rosenthal-Effekt

In den Hawthorne-Experimenten spielte die Tatsache, dass Menschen (die Forscher), die Arbeiter beobachteten und mit ihnen in Interaktion traten, eine große Rolle. Viele Studien legen nahe, dass die Wirkung von Placebos stark von der Beziehung zwischen Arzt und Patient abhängig ist (Adler & Van Buren, 1973; Benedetti, 2013). Der in diesem Abschnitt behandelte Rosenthal-Effekt beschreibt die Wirkungen der Interaktion zwischen Versuchsperson und Versuchsleiter.
Robert Rosenthal untersuchte in beeindruckenden Experimenten wie die Erwartungen und Vorurteile von Versuchsleitern die Leistung und Entwicklung von Versuchspersonen beeinflussen. In einem Experiment (Rosenthal & Fode, 1963) wurden Laborratten an Studenten ausgeteilt, welche die Ratten nur pflegen sollten. Einer Versuchsgruppe wurde gesagt, dass es sich um besonders intelligente Ratten handle, die andere Gruppe erfuhr, dass ihre Ratten daraufhin gezüchtet wurden, besonders dumm zu sein. In späteren Versuchen, in denen die Ratten möglichst schnell den Weg durch ein Labyrinth finden sollten, schnitten die Tiere besser ab, deren Pfleger dachten, sie wären nach Intelligenz gezüchtet worden.

In einem ähnlichen Experiment (Rosenthal & Jacobson, 1966) wurden in normalen Schulklassen Listen an Lehrer ausgeteilt, die die Namen von angeblich besonders vielversprechenden Schülern enthielten. Acht Monate später machten diese Schüler tatsächlich größere Fortschritte in IQ-Tests als ihre Mitschüler, obwohl sie nur zufällig ausgesucht wurden.
Erklärungen für diese Phänomene beinhalten immer, dass der Versuchsleiter (oder Lehrer) unwillentlich durch sein (nonverbales) Verhalten seine Erwartungen ausdrückt (Huber, 2005). Vor allem für den schulischen Kontext wird überlegt, ob die Schüler mit der Zeit dieses Fremdbild als Selbstbild annehmen, also die Erwartungen auch selbst an sich stellen.[3] Eine andere Erklärung ist, dass der Versuchsleiter durch seine Einstellung die Bedingungen des Experiments (wenn auch nur minimal) ändert. Bei vielversprechenden Probanden gibt er mehr Zeit zum Antworten oder ist freundlicher und schafft somit eine Atmosphäre, die günstiger für Konzentration und Leistung ist, während er die „schlechten" Probanden mit kühlerem Verhalten verunsichert oder weniger gütig bei der Befragung ist.

Kontrolle des Rosenthal-Effekts
Wie verhindert man einen solch subtilen Effekt? Dafür gibt es die zweite Seite der „Blindheit" eines Doppelblindversuchs: Je weniger der Versuchsleiter über das Experiment weiß, desto weniger kann er auch durch sein Verhalten kommunizieren. Wenn der Versuchsleiter zum Beispiel nur Fragebögen oder auch Medikamente austeilt, kann man dies leicht umsetzen.
Weiß der Versuchsleiter aufgrund der Aufgaben, die er selbst zu erledigen hat unvermeidlich auch, welches Ziel bei den Probanden erreicht werden soll, hilft nur ein Training. Der Versuchsleiter soll dabei lernen, mit allen Probanden gleich umzugehen. Dies soll zu einer maximalen Standardisierung der Versuche beitragen.
Letztlich ist es allerdings immer ratsam, jede Interaktion zwischen Eingeweihten und Probanden minimal zu halten (Sedlmeier & Renkewitz, 2013).

7.2 Oberserver-Bias

Wird in einem Experiment eine Verhaltensbeobachtung durchgeführt, muss man sich den kognitiven Prozess auf der Seite des Beobachters genauer anschauen. Nicht nur in Experimenten, auch in der Pilotenausbildung baut der gesamte Lehrprozess auf Verhaltensbeobachtungen der Flugschüler auf. Es ist also wichtig, dass die Beobachtungen objektiv und reliabel (siehe Abschnitt 3) sind. Ein Effekt, der systematische Verfälschung der Wahrnehmung des Beobachters beschreibt, nennt man Observer-Bias (Sedlmeier & Renkewitz, 2013). Es kann verschiedene Gründe dafür geben, dass eine Beobachtung nicht reliabel ist. Einige davon werden hier aufgelistet.

7.2.1 Halo-Effekt

Beim Halo-Effekt „überstrahlt" eine besondere Eigenschaft des Beobachteten oder der Gesamteindruck, den der Beobachter vom Beobachteten hat, alle anderen Merkmale. Wenn der Beobachter eine Versuchsperson sympathisch findet, ist es wahrscheinlicher, dass er auch andere Merkmale positiver wertet. (Sedlmeier & Renkewitz, 2013)

[3] Wird von Lehrern gesprochen, deren Erwartungen einen Einfluss auf die Schüler nehmen, so ist auch der Begriff *Pygmalion-Effekt* geläufig. Die mythologische Figur Pygmalion erschuf sich eine Frauenstatue nach seinen Idealen und verliebte sich in sie. Diese wurde irgendwann lebendig.

7.2.2 Observer-Drift

Beim Observer-Drift verändert sich die Sichtweise des Beobachters über die Zeit hinweg. Dies kann an nachlassender Aufmerksamkeit liegen. Es kann aber tatsächlich auch passieren, dass die Beobachtungen mit der Zeit genauer werden, da der Beobachter mit zunehmender Erfahrung ein geschulteres Auge für den Beobachtungsgegenstand entwickelt. (Sedlmeier & Renkewitz, 2013)

7.2.3 Ankereffekt

Werden nacheinander verschiedene Personen oder Objekte beobachtet und auf bestimmte Weise bewertet oder kategorisiert, kann es passieren, dass die erste Kategorisierung die folgenden bestimmt. Man sagt, dass sie als Anker wirkt. Wird zum Beispiel ein Proband, der eine Aufgabe mit sehr wenigen Fehlern gelöst hat, als mittelmäßig eingestuft, müssen die nächsten Probanden noch weniger Fehler machen, um als sehr gut eingestuft zu werden, da sie (vom Beobachter bewusst oder unbewusst) mit dem ersten verglichen werden. (Sedlmeier & Renkewitz, 2013)

7.2.4 Kontrasteffekt

Der Kontrasteffekt beschreibt, dass nur kleine Unterschiede zwischen zwei Probanden als viel größer wahrgenommen werden könnten, als sie tatsächlich sind und sich dies in der Bewertung durch einen großen Unterschied wiederspiegelt. (Sedlmeier & Renkewitz, 2013)

7.2.5 Einfluss von anderen Merkmalen

Pritz (1981) beschreibt das Phänomen, dass schnell sprechende Schüler in einer mündlichen Prüfung systematisch bessere Noten bekamen, als langsam sprechende Schüler, obwohl die Inhalte des Sprechens gleich waren. Dies ist nur ein Beispiel, das verdeutlichen soll, dass wir versuchen aus Merkmalen, die eigentlich nichts mit dem zu beobachteten Merkmal zu tun haben, Informationen über das gemessene Merkmal zu ziehen.

7.2.6 Kontrolle des Observer-Bias

Verschiedene Beobachter-Effekte kann man am besten eingrenzen, indem man Beobachter über die möglichen Fehler aufklärt, sie schult und die Beobachtungskategorien vorher genau festlegt. Dies bedeutet, dass klar definiert sein sollte, welches Verhalten welcher Kategorie zugeordnet wird.
Gegen den Observer-Drift aufgrund von Müdigkeit helfen häufige Pausen (Sedlmeier & Renkewitz, 2013).

7.3 Interpretationsspielraum in Fragebögen

7.3.1 Sprache in Fragebögen

Ein Problem von Fragebögen ist, dass diese meist durch Sprache „abfragen“. Die Formulierungen im Fragebogen müssen zunächst von den Gestaltern des Fragebogens gefunden werden. Dann werden sie von den Probanden interpretiert. Daraufhin wählen diese die Formulierungen aus, die nach ihrer

Interpretation am besten ihr Empfinden wiedergeben. Bei der Auswertung des Fragebogens weiß man allerdings nicht, wie die Probanden die sprachlichen Formulierungen verstehen.

Sprachfreie Einstellungsabfrage: Die Kunin-Skala
Ein Versuchs diesem Problem zu begegnen, wurde von Theodore Kunin unternommen (1955). Er entwickelte eine Skala zur Einstellungsabfrage, die anstatt aus verbalen Formulierungen zu bestehen, Gesichter nutzte. In einer aufwändigen Studie ließ er Probanden die „Glücklichkeit" von Gesichtern bewerten und wählte anschließend die Gesichter aus, bei denen die Probanden die größte Übereinstimmung hatten, sodass man das Ankreuzverhalten von späteren Probanden gut vergleichen konnte. Dieses Verfahren soll die Reliabilität (siehe Abschnitt 3.2) einer solchen Skala gewährleisten. Die Probanden zeigten bei der Bewertung bei diesen fünf Gesichtern die größte Übereinstimmung, weshalb Kunin sie zur Nutzung empfahl:

Abbildung 6. Die von Kunin empfohlenen fünf Gesichter zur Einstellungserfassung. (Aus: Kunin, 1955.)

Heute werden allerdings auch die verschiedensten anderen „Zeichentrick-Gesichter" in Fragebögen eingesetzt.

7.3.2 Interpretation mithilfe der Darstellung

Bei der Abbildung der „Einstellungs-Gesichter" fällt auf, dass das "neutrale" Gesicht, also dasjenige ohne Krümmung des Mundes, in der Mitte der Skala steht. Ein Proband kann also nun aus mindestens zwei Quellen seine Interpretation der Gesichter ziehen. Beides, die Krümmung des Mundes und die Lage des Gesichts in der Reihe deuten daraufhin, dass man „etwas Mittelmäßiges wählt", wenn man das dritte Gesicht ankreuzt.
In diesem Abschnitt soll es nun um den Einfluss der Skalengestaltung auf das Verhalten der Probanden gehen. Denn auch nicht-sprachliche Darstellungen können durch die Interpretation von Probanden an Bedeutung gewinnen.

7.3.2.1 Tendenz zur Mitte

Bei Befragungen stößt man immer wieder auf ein Phänomen: Versuchspersonen neigen dazu, Antwortmöglichkeiten anzukreuzen, die sich in der Mitte der Anordnung befinden. Eine Erklärung ist, dass Menschen meinen, dass die Mitte auch die Norm darstellt. Weiterhin geht man davon aus, dass Menschen nicht durch die Wahl einer „extremeren" Antwort als andersartig hervorstechen wollen.
Interessant ist, dass selbst in einer Studie, in der dieser Effekt an Multiple-Choice-Fragen untersucht wurde, bei denen die Antwortmöglichkeiten nicht nach Größe oder Ausprägung sortiert werden konnten, ihre Anordnung also zufällig war, die Probanden die in der Mitte stehenden Antworten häufiger ankreuzten (Sedlmeier, 2006).

Beispiel-Aufgabe
Frage: Wenn Sie die Güte einer Lehrveranstaltung einschätzen würden: Welches der folgenden vier Kriterien wäre für Sie am wichtigsten (bitte nur eines ankreuzen)?

Version 1:

Prüfungsbezug	Praxisbezug	Schwierigkeit	Strukturiertheit

Version 2:

Praxisbezug	Prüfungsbezug	Strukturiertheit	Schwierigkeit

In diesem Beispiel wurde der Praxisbezug viel häufiger in der *Version 1* als wichtigstes Kriterium genannt.

7.3.2.2 Zahlen sind nicht objektiv

7.3.2.2.1 Positive und negative Zahlen

Bei einer Befragung von Studenten zur empfundenen Relevanz einer Lehrveranstaltung (Seldmeier, 2006) bekam eine Gruppe eine Skala mit den Endpunkten 0 (irrelevant) bis 100 (sehr relevant), eine andere Gruppe eine Skala mit den Endpunkten -50 (irrelevant) bis +50 (sehr relevant) vorgelegt.

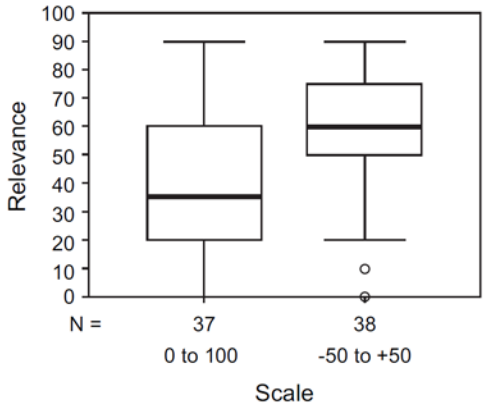

Abbildung 7. Boxplot der Daten der Umfrage zur Relevanz der Lehrveranstaltung. Für die Graphik wurde zu den Daten der Gruppe 2 (-50 bis +50) jeweils 50 addiert. (Aus: Sedlmeier, 2006.)

In der Graphik sieht man, dass in der zweiten Gruppe viel häufiger in den oberen Bereichen der Skala angekreuzt wurde. Sedlmeier sieht als möglichen Grund, dass die Probanden ungern negative Werte ankreuzen wollten. Er vermutet, dass in den Augen der Probanden die Punktzahl -10 eher Irrelevanz andeutet, als die Punktzahl 40, obwohl beide Werte an der gleichen Stelle ihrer jeweiligen Skala stehen.

7.3.2.2.2 Intervallbreiten

In einer anderen Studie sollten Studenten schätzen, wie viel Zeit sie durchschnittlich zur Vorbereitung einer Vorlesungssitzung benötigen würden. Dabei wurden zwei Gruppen von Probanden Skalen unterschiedlicher Breite (also mit unterschiedlich großen Zeitintervallen zur Auswahl) vorgelegt.

Feine Skala der ersten Gruppe:

< 5 Min.	5 – 15 Min.	16 – 30 Min.	30 Min. – 1 h	> 1 h

Grobe Skala der zweiten Gruppe:

< 30 Min.	30 Min. – 1 h	1 – 1,5 h	1,5 h – 2 h	> 2 h

Interessant ist, dass in der ersten Gruppe nur 8,1% der Teilnehmer schätzten, mehr als eine Stunde zur Vorbereitung zu benötigen, während es in der zweiten Gruppe 30,5% waren!

Hier wirkt als Erklärung wieder, dass Probanden davon ausgehen, dass der mittlere Wert einer Skala auch den mittleren Wert einer Population, also die Norm darstellt und sie sich selbst auch als normal sehen und darstellen wollen.

7.3.2.2.3 Abgefragter Zeitraum

Stellt man Piloten die Aufgabe „Nennen Sie mir die anstrengendsten Flüge innerhalb Ihres Berufslebens“ so wird man wahrscheinlich andere Antworten erhalten, als bei „Nennen Sie mir die anstrengendsten Flüge im letzten Jahr“. Es kann sogar sein, dass kein einziger der Flüge, die bei der zweiten Frage genannt wurden, bei der ersten Version vorkommt.
Paradox wird es, wenn die Piloten die Häufigkeit anstrengender Flüge auf einer Skala angeben sollen. Menschen gehen bei der Vorgabe von langen Zeiträumen davon aus, dass seltene und besonders intensive Erlebnisse gefragt, bei kürzeren Zeiträumen aber auch weniger intensive, häufigere Erlebnisse von Interesse sein können (Winkelman, Knäuper & Schwarz, 1998). Es kann also sein, dass Piloten rückblickend über ihr bisheriges Berufsleben sagen, dass es nicht besonders anstrengend war und ihnen ja auch nie etwas wirklich Schlimmes passiert ist. Infolgedessen geben sie auf einer Skala von 1 (keine anstrengenden Flüge) bis 10 (sehr viele anstrengende Flüge) durchschnittlich eine **3** an. Probanden, die sich an das letzte Jahr erinnern sollen, haben es einfacher, sich einzelne Flüge in Erinnerung zu rufen, da der Zeitraum kürzer ist. Sie blicken mit einem anderen „Filter“ auf die Vergangenheit und picken bereits leicht aufregende Erlebnisse hervor, von denen es mehr gibt. Als Konsequenz ist der Mittelwert der gewählten Zahlen in dieser Gruppe M = **6**. Diese zweite Gruppe hatte trotz des höheren Wertes natürlich *nicht* in einem Jahr mehr anstrengende Flugerlebnisse als eine andere Gruppe in ihrem bisherigen Berufsleben. Man sollte also immer darauf achten, was womit verglichen wird, um nicht zu Fehlschlüssen zu gelangen.

7.3.2.3 Antwortvorgabe

Wenn man sich in einer Befragung dafür entscheidet, den Probanden mögliche Antwortalternativen vorzugeben, dann wird man natürlich auch nur Antworten in diesem Spektrum erhalten. Man kann den Befragten *zusätzlich* die Möglichkeit für freie Antworten geben. Diese wird meist aber kaum genutzt.
Fragebögen mit Antwortvorgabe sind am sichersten interpretierbar, wenn man zunächst eine Vorstudie durchgeführt hat, in der man alle möglichen Antworten herausfindet und den Gegenstandsbereich besser kennenlernt (Sedlmeier, & Renkewitz, 2013).

Beispiel:
Eine Airline möchte herausfinden, welche Situationen den Piloten am meisten Stress bereiten, um diese später verstärkt zu trainieren. Da die Auftraggeber nicht wissen, welche Situationen überhaupt als Stressoren in Frage kommen, wird zunächst eine Vorstudie durchgeführt, in der die Piloten frei von den anstrengendsten Tätigkeiten berichten.
Daraufhin kann man in der Hauptstudie mit den am häufigsten angegebenen Situationen Experimente durchführen, in denen physiologische Daten der Probanden gesammelt werden, die als Indikatoren für Stress nutzbar sind – etwa den Puls oder die elektrodermale Aktivität (siehe Abschnitt 4.1).
Eine zweite schriftliche Umfrage *mit* Antwortvorgabe könnte man in diesem Beispiel einbauen, um vor den aufwändigen Experimenten die Zahl der zu untersuchenden Situationen noch einmal zu reduzieren. Es würde sich auch anbieten die Piloten direkt zu fragen, welche dieser vorsortierten Stresssituationen sie üben wollen.

7.3.2.4 Vergleichsurteile

Manchmal möchte man von Probanden ein Vergleichsurteil hören. Zum Beispiel sollen sie angeben, welche Einstellung eines Simulators realistischer wirkt.
Dabei kann man fragen:

1. „Verglichen mit der Einstellung A, wirkt die Einstellung B realistischer oder weniger realistisch auf Sie?“

Man kann aber auch fragen:

2. „Verglichen mit der Einstellung B, wirkt die Einstellung A realistischer oder weniger realistisch auf Sie?“

Dieser Unterschied wirkt auf den ersten Blick trivial, kann aber einen Effekt bewirken.
Wänke hat in einem Experiment genau diese Auswirkungen der „Vergleichsrichtung“ untersucht (1996). In seiner Studie sollten die Probanden allerdings einschätzen, wer mehr zur Luftverschmutzung beiträgt, der Verkehr oder die Industrie. War der Verkehr die Ausgangsbasis für den Vergleich (so wie die Einstellung A in unserem ersten Satz), wurde der Verkehr als mehr verantwortlich gesehen, war die Industrie die Ausgangsbasis, so sprachen die Probanden ihr mehr Schuld zu. Der Effekt, dass die Ausgangsbasis immer den höchsten Wert in einem Vergleich davonträgt, wurde auch gefunden, wenn es sich nicht um einen direkten Zweiervergleich handelte, sondern wenn Probanden den relativen Beitrag zur Luftverschmutzung auf einer Rating-Skala angeben sollten (Sedlmeier & Renkewitz, 2013).
Der gemessene Effekt ist zwar nur klein, kann aber trotzdem bei sehr ähnlichen Vergleichsalternativen die Bewertung umdrehen (Sedlmeier, & Renkewitz, 2013). Stattdessen kann man auch fragen:
„Welche Einstellung des Simulators wirkt realistischer auf Sie: A oder B?
Aber auch hier besteht die Möglichkeit, dass die Reihenfolge der Aufzählung die Antworten beeinflusst, ähnlich wie bei der Tendenz zur Mitte bei Multiple Choice-Aufgaben.

7.3.2.5 Kontrolle von Darstellungseffekten

Man verfährt immer sicher, wenn man seine Fragebögen in der Gestaltung variiert, also nicht an alle Probanden die gleiche Version des Fragebogens austeilt. Dies ermöglicht im Nachhinein zu prüfen, ob die Gestaltung einen Einfluss auf die Antworten hatte (Sedlmeier & Renkewitz, 2013).

8. Empfehlung

Dieser Bericht hat sein Ziel bereits erreicht, wenn der Leser sich von seiner Vorstellung verabschiedet hat, dass in den psychologischen Wissenschaften jemals absolute Aussagen gemacht werden. Diese Erkenntnis ist natürlich einerseits wichtig, um sich einen kritischen Blick auf die Forschung zu bewahren, andererseits ist sie unumgänglich, um selbst forschend tätig zu sein. Bei der Erhebung von psychologischen Größen stößt man bereits bei der Definition von Konzepten und Begriffen auf die Tatsache, dass latente Variablen nichts Greifbares sind. Im gesamten Forschungsprozess, von der Entwicklung eines Konzepts bis zur Interpretation von erhobenen Daten, steht man auf einem selbst konstruierten Grund. Es ist wichtig, diesen Grund nicht aus den Augen zu lassen, wenn man am Ende Daten haben möchte, die interpretierbar sind. Denn: Die Interpretation von Daten fußt auf dem zu Beginn konstruierten Konzept und bezieht sich darauf. Die Gütekriterien sind Werkzeuge, um sicherzustellen, dass man mit jedem Schritt weiterhin die Nähe zu diesem Konzept behält. Je genauer die Vorstellung von den Begriffen, die man nutzt, desto einfacher ist es, diese Begriffe in einer Messung auch wiederzufinden. Wer vorher nicht operational definiert hat, was er sich bei dem Konzept „Intelligenz“ denkt, der kann „Intelligenz“ auch nicht messen. Eine intensive Auseinandersetzung mit dem Gegenstandsbereich hilft auch dabei, zu planen und zu beobachten, ob

Probanden die Experimentalsituation gemäß dem Konzept wahrnehmen oder auf dem „falschen Dampfer“ unterwegs sind und dementsprechend unbrauchbare Daten liefern, also sich in ihrem wahrgenommene Kontext verhalten, nicht in dem durch das Experiment angestrebten.

Literatur

Adler, H. M. & Van Buren, O. H. (1973). The doctor-patient relationship revisited: an analysis of the placebo effect. *Annals of Internal Medicine*, *78*(4), 595-598.

Anderson, J. (2007). *Kognitive Psychologie* (1. Aufl.). Heidelberg: Spektrum, Akad. Verl.

Benedetti, F. (2013). Placebo and the new physiology of the doctor-patient relationship. *Physiological reviews*, *93*(3), 1207-1246.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual review of psychology*, *53*(1), 605-634.

Brewer, W. F. & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive psychology*, *13*(2), 207-230.

DLR - Institut für Luft- und Raumfahrtmedizin - Luft- und Raumfahrtpsychologie. (2017). *Dlr.de*. Verfügbar unter http://www.dlr.de/me/desktopdefault [20.01.2017]

Domjan, M. & Grau, J. (2003). *The principles of learning and behavior*. Australia: Thomson/Wadsworth.

Eggert, F. (2014). Experimentelle Designs. (Vorlesung, 25.11.2014). Braunschweig: Technische Universität Braunschweig.

Ellson, D. G., Davis, R. C., Saltzman, I. J. & Burke, C. J. (1952). A report on research on detection of deception DDC Technical Report ATI-168902, 1952, Indiana University. *Contract No. Nonr-60nr-18011. Office of Naval Research*.

Fries, A., & Grawe, K. (2006). Inkonsistenz und psychische Gesundheit: eine Metaanalyse. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie*, *54*(2), 133-148.

Gustafson, L. A. & Orne, M. T. (1965). Effects of perceived role and role success on the detection of deception. *Journal of Applied Psychology*, *49*(6), 412.

Huber, O. (2005). *Das psychologische Experiment: Eine Einführung* (4. Aufl.). Bern: Huber.

Hussy, W., Schreier, M. & Echterhoff, G. (2010). *Forschungsmethoden in Psychologie und Sozialwissenschaften - für Bachelor*. Springer-Verlag.

Kelly, G. A. (1955). *The psychology of personal constructs*. New York: Norton.

Kelly, G. A. (1977). Personal construct theory and the psychotherapeutic interview. *Cognitive therapy and research*, *1*(4), 355-362.

Kuhn, T. (1970). *The structure of scientific revolutions* (1. Aufl.) Chicago: University of Chicago Press.

Kunin, T. (1955). The Construction of a New Type of Attitude Measure. *Personnel psychology*, *8*(1), 65-77.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological reports*, *3*(3), 635-694.

Loftus, E. F. & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of verbal learning and verbal behavior*, *13*(5), 585-589.

Lück, H. E. (2009). Der Hawthorne-Effekt–ein Effekt für viele Gelegenheiten?. *Gruppendynamik und Organisationsberatung*, *40*(1), 102-114.

Nagel, T. (1974). What is it like to be a bat?. *The philosophical review*, *83*(4), 435-450.

Orne, M. T. & Evans, F. J. (1965). Social control in the psychological experiment: Antisocial behavior and hypnosis. *Journal of Personality and Social Psychology*, *1*(3), 189.

Placebo & Nocebo | Placeboforschung.de. (2017). *Placeboforschung.de. Verfügbar unter* http://placeboforschung.de/de/placebo-nocebo.aspx/tabid-5046/. [20.01.2017]

Pritz, V. (1981). Der Einfluss von Sprechflüssigkeit und Vorinformation auf die Leistungsbeurteilung in der mündlichen Reifeprüfung. *K. Ingenkamp (Hg.), Wert und Wirkung von Beurteilungsverfahren. Weinheim*, 50-96.

Something Made Your Brain Think These Dots Are a Dog [pic] - Randommization. (2017). *Randommization*. Verfügbar unter http://randommization.com/2013/05/21/something-made-your-brain-think-these-dots-are-a-dog-pic/. [20.01.2017]

Roethlisberger, F. J. & Dickson, W. J. (1964). *Management and the worker: an account of a research program conducted by the Western Electric Company, Hawthorne Works, Chicago, by FJ Roethlisberger and William J. Dickson, with the assistance and collaboration of Harold A. Wright*. Harvard Univ. Press.

Roethlisberger, F. J. & Dickson, W. J. (2003). *Management and the Worker* (5. Aufl.). Psychology Press.

Rosenthal, R. & Fode, K. L. (1963). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, *8*(3), 183-189.

Rosenthal, R. & Jacobson, L. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological reports*, *19*(1), 115-118.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik (Lehrbuch mit Online-Materialien)*. Springer Science & Business Media.

Schreier, M. & Odağ, Ö. (2010). Mixed Methods. In *Handbuch qualitative Forschung in der Psychologie.* VS Verlag für Sozialwissenschaften.

Sedlmeier, P. (2006). The role of scales in student ratings. *Learning and Instruction*, *16*(5), 401-415.

Sedlmeier, P. & Renkewitz, F. (2013). *Forschungsmethoden und Statistik für Psychologen und Sozialwissenschaftler* (1. Aufl.). München: Pearson, Higher Education.

Wänke, M. (1996). Comparative judgments as a function of the direction of comparison versus word order. *Public Opinion Quarterly*, *60*(3), 400-409.

Watzlawick, P., Beavin, J. H. & Jackson, D. D. (1990). *Menschliche Kommunikation*. Bern, Stuttgart: H. Huber.

Winkielman, P., Knäuper, B. & Schwarz, N. (1998). Looking back at anger: Reference periods change the interpretation of emotion frequency questions. *Journal of Personality and Social Psychology*, *75*(3), 719.

Zucker, I. & Beery, A. K. (2010). Males still dominate animal studies. *Nature*, *465*(7299), 690-690.