Dr. Daniel Wessel ◇ http://danielwessel.org ◇ danwessel@me.com

Vorbemerkungen

Über die kritische Betrachtung von Studien[1] kann man ganze Bücher schreiben. Entsprechend ist dieser Fragebogen nur als erste Handreichung zu verstehen. Wann immer in diesem Fragebogen eine Frage als zutreffen angekreuzt wurde, liegt möglicherweise eine Schwäche vor.

„You keep using that survey.
I do not think it means what you think it means."
Inigo Montoya (modified)

Nicht jede Schwäche ist eine bewusste Täuschung: Auch kompetente Forscher/innen können Fehler machen und keine Studie ist perfekt. Andere gehen allerdings grob irreführend vor: Sie degradieren die Wissenschaft zur Handlangerin einer Ideologie und betreiben Augenwischerei.

Insbesondere wenn Studien von Personen und Organisationen durchgeführt werden, die primär einer persönlichen oder institutionellen Agenden folgen, ist Vorsicht geboten. Hier werden zum Teil Ergebnisse bewusst oder unbewusst zum eigenen Vorteil hin "geformt". Zum Beispiel indem allgemeine Probleme als spezifisch für Teilgruppen dargestellt oder tatsächliche Lösungen verhindert werden. Wären auch andere davon betroffen würde der "*unique selling point*" der Organisation wegfallen. Und wenn man das Problem effektiv angehen und lösen würde, dann würde sich diese Person/Organisation überflüssig machen. Das trifft keineswegs auf alle, und vermutlich nicht mal auf die meisten, Organisationen zu. Aber es kommt vor.

Die kritische Haltung zu Studien soll auch nicht bedeuten, dass Studien per se schlecht oder irreführend sind, oder dass alle Statistiken gefälscht sind. Im Gegenteil. Gut durchgeführte und ausgewertete Studien ermöglichen Einblicke jenseits von bloßen Gefühlen oder Vermutungen. Sie erlauben es uns, bessere Entscheidungen zu treffen und unsere begrenzten Ressourcen dort einzusetzen, wo sie den größten Effekte haben. Und sie machen diese Effekte nachweisbar.

Unabhängig von der Intention haben die Schwächen einer Studie Konsequenzen für die Qualität der Ergebnisse und damit auch für die Qualität der möglichen Lösungen. Wird die Sachlage verzerrt, beeinflusst dies die Qualität der Entscheidungen.

Gerade ideologisch verzerrte Interventionen führen oft zu den Problemen, die sie eigentlich beheben sollten. Das ist für die Personen gut, die von diesen Konflikten profitieren, weil diese Konflikte ihre Infrastruktur und Stellen finanziert, ist aber langfristig schädlich sowohl für den Bereich als auch für die betroffenen Personen.

Wenn etwas wirklich ein Problem ist, dann sollten die Beteiligten keine Angst davor haben, qualitativ hochwertige Studien durchzuführen. Und für alle bis auf Interessengruppen ist es eher ein Vorteil als ein Nachteil, wenn ein Problem allgemeiner statt gruppenspezifischer Natur ist. Man hat mehr Personen die ein Interesse daran haben, das Problem zu lösen und es stellt eine Gruppe nicht als defizitär dar.

[1] Der Fragebogen bezieht sich vor allem auf Umfragen. Bei Experimenten kämen weitere Punkte hinzu, u.a., wie gut die Interventionen waren und ob sie gegriffen haben, ob die Personen wussten, in welcher Experimentalgruppe sie waren, die Effekte rein durch die Situation selbst, und viel mehr.

Quellen (u.a.):

Krämer, W. (1998). *Statistik verstehen. Eine Gebrauchsanweisung.* (3rd ed.). Frankfurt: Campus Verlag.

Krämer, W. (2000). *So lügt man mit Statistik.* München: Piper.

Perry, M. J. (2015, January 12). No polling organization would ever be taken seriously if its sample size was 73, and neither should this "study" on college rape. *www.aei.org*. Retrieved April 28, 2015, from http://www.aei.org/publication/polling-organization-ever-taken-seriously-sample-size-73-neither-study-college-rape/

Simmons, R. (2011). *Odd Girl Out. How to help your daughter navigate the world of friendships, bullying and cliques -- in the classroom and online.* London: Piatkus.

Dr. Daniel Wessel ◇ http://danielwessel.org ◇ danwessel@me.com

Oberflächliche Betrachtung

☐ Ändert sich die Glaubwürdigkeit der Ergebnisse, wenn die Gruppenzuordnungen vertauscht werden?

Man kann die Rolle der eigenen Vorurteile schnell sehen, wenn man z.B. die Hauptfarbe von "weiß" auf "schwarz", die Religion von "muslimisch" auf "christlich", das Geschlecht von "männlich" auf "weiblich", die sexuelle Präferenz von "heterosexuell" auf "homosexuell", oder die Nationalität von "amerikanisch" auf "deutsch" ändert. Wenn man die Ergebnisse dann kritischer sieht, dann entsprechend mit dieser Einstellung im Detail auf die Qualität der Studie schauen.

☐ Ist ein Interessenskonflikt zu erwarten?

> "Jede Statistik, die von einer interessierten Seite selbst erstellt und verbreitet wird, ist bis zum Beweis des Gegenteils als manipuliert zu betrachten."
> *Krämer, 1998*

Folge dem Geld: Wenn die Ergebnisse nicht oder sogar anders herum eingetreten wären, hätte dies negative Konsequenzen (Fördermittel, Einfluss, Status, *raison d'être*, etc.) für die Personen, die diese Studie durchgeführt haben? Sind die gefundenen Ergebnisse "gewünscht", im Sinne von: "Das Problem ist schlimmer geworden, aber wir leisten gute Arbeit und brauchen deswegen mehr Geld."? (Auch mit den besten Absichten ist dies ein Problem.) Müssen die Ergebnisse aufgrund einer zugrundeliegenden Ideologie eintreten, z.B. auf der Annahme, dass es keine Unterschiede zwischen den Gruppen in Interessen oder Leistung geben *darf* und entsprechend alle *gefundenen* Ergebnisse eine Konsequenz von unfairer Diskriminierung sein *müssen*?

☐ Wurde die Studie außerhalb einer wissenschaftlichen (oder in einer inzestuösen "wissenschaftlichen") Community durchgeführt?

Studien die ausschließlich von außerwissenschaftlichen Organisationen durchgeführt und veröffentlicht wurden, d.h. die weder an Universitäten/wissenschaftlichen Instituten durchgeführt wurden noch in peer-reviewten wissenschaftlichen Fachzeitschriften veröffentlicht wurden. Peer-review hat seine Schwächen, aber ist besser als "privat" durchgeführte Studien. /!\ Einige Studien wurden zurückgezogen oder korrigiert. Die "retraction notices" sind leider selten deutlich gemacht. Vorsicht vor wissenschaftlichem Inzest. Einige Subgruppen referieren sich gegenseitig und winken ihre verzerrten Studien durch.

☐ Gibt es kritische Reaktionen auf die veröffentlichte Studie?

Nicht jede Kritik ist auch berechtigt. Allerdings gibt es als Reaktion auf miserable Studien häufig Kritik in Blogs und in sozialen Medien. Wichtig sind hier Argumente, nicht Kritik an der "Erwünschtheit" der Ergebnisse.

Forschungslage

☐ Ist das theoretische Fundament mit ideologischen Steinen gebaut?

Häufig werden Konzepte verwendet die ideologisch geprägt sind und nicht auf klaren empirischen Definitionen beruhen (= woran kann man es konkret sehen?). **✘ Showstopper: Hiermit wird oft nur ein Vorurteil "bestätigt" indem die Untersuchung so gestaltet wird, dass nur das ideologisch passende herauskommt.**

☐ Wurde die zugrundeliegende Literatur sehr selektiv ausgewählt?

Man kann mit entsprechend ausgewählter Literatur die Basis für eine verzerrte Wahrnehmung schaffen. Z.B. indem man nur Literatur zitiert, welche die eigene Sichtweise unterstützt. Eine kritische Auseinandersetzung mit einem Thema setzt voraus, dass man auch Literatur zitiert, die für eine andere Position spricht. Man muss die Literatur dann entsprechend widerlegen können.

☐ Wurde die widersprechende Literatur verzerrt wiedergegeben?

> "The question is, do you think you're big enough to read a book by David Irving and make up your own mind about it, or do you think that someone else should do that for me."
> *Christopher Hitchens*

Insbesondere bei kontroversen Quellen lohnt es sich, persönlich einen Blick in die Literatur zu werfen. Oft wird diese sehr verzerrt und vereinfacht wiedergegeben. Und nur weil man eine kontroverse Position nachvollziehen kann, heißt dies nicht, dass man dieser auch zustimmt.

☐ Wurde bei der zustimmenden Literatur geheuchelt?

Es gibt einen "*woozle effect*" bei dem eine miserable Studie so lange über diverse Ecken weiter zitiert wird, bis die Zitation der letzten Quelle in der Kette glaubwürdig wirkt, weil keiner mehr bis zur miserablen Originalstudie kommt. Entsprechend muss man kritisch überprüfen, ob die Behauptungen in der Studie auch wirklich *empirisch* belegt sind. Ebenso wenn die schwachen Teile einer Studie zitiert werden (z.B. reine Vermutungen der Autoren/innen), aber nicht die empirischen Ergebnisse. Hier kann aus einem "könnte" oder "möglicherweise" schnell eine "gesicherte Erkenntnis" oder ein "Befund" werden. **✘ Showstopper (allerdings nur mit großem Aufwand nachzuweisen): Das Fundament ist damit fehlerhaft und "Befunde", die vorausgesetzt werden, sind nur Vorurteile. Gute Studien setzen gute Literaturkenntnisse voraus.**

Dr. Daniel Wessel ◇ http://danielwessel.org ◇ danwessel@me.com

Stichprobe

☐ Wurde nur die "interessante" Gruppe erhoben?

Obwohl ein Vergleich (Gruppe *x* hat es besser/schlechter als Gruppe *y*) behauptet wird, was zwangsläufig ein Vergleich zweier Gruppen voraussetzt, wird nur eine Gruppe erhoben. Oft werden hier Vorurteile ausgenutzt und einfach angenommen, die andere Gruppe hat es besser/schlechter, oder leider schlicht und ergreifend einfach nicht unter dem Problem. Emotional involvierende Geschichten werden hier häufig eingesetzt. **✘ Showstopper: Eine solche Studie ist komplett wertlos wenn es um einen Vergleich geht.**

☐ Wurde eine falsche Dichotomie aufgemacht?

In einigen Fällen scheint es nur so, als würde es nur zwei Gruppen zum Vergleich geben -- z.B. beim Vergleich "Schwarze" vs. "Weiße". Dass es noch Asiaten, Latinos, etc. gibt wird einfach übersehen.

☐ Wurden unterschiedliche Kriterien für Gruppendefinitionen verwendet?

Beliebt ist hier, die erfolgreichen Mitglieder/innen einer Gruppe mit allen Mitglieder/innen einer anderen Gruppe zu vergleichen. Zum Beispiel die Personen einer Gruppe, die es geschafft haben, in einem Bereich erfolgreich zu sein, mit allen Mitgliedern/innen einer anderen Gruppe. Für die "Unerfolgreichen" aus der "erfolgreichen" Gruppe interessiert sich üblicherweise niemand. **✘ Showstopper: Eine solche Studie ist wertlos wenn es um einen fairen Vergleich geht.**

☐ Fielen Personen selektiv bei der Messung raus?

Zum Beispiel weil Personen einer Gruppe aufgrund mangelnder Förderung in dieser Gruppe nicht so weit kommen. Auch Eindrucksvoll: Ein Krankenhaus, welches Krebspatienten kurz vor dem Tod nach Hause schickt, was die Sterblichkeitsrate im Krankenhaus stark reduziert (Krämer, 1998).

☐ Werden Gruppen auf eine/n prototypische/n Repräsentanten/in reduziert?

Es gibt üblicherweise eine ziemliche Heterogenität innerhalb einer Gruppe. Es gibt nicht nur die "coolen" Jungs, sondern auch die Nerds und die normalen Jungs. Ebenso hat nicht jeder Mann das Aussehen von Brad Pitt, das Geld von Rockefeller, und den Einfluss eines Kennedy's. Es ist also immer Vorsicht geboten, wenn Gruppen auf "Prototypen" reduziert werden oder nur spezifische Segmente einer Gruppe untersucht werden.

☐ Wurden die Beteiligten selektiv ausgewählt?

Eine gute Studie benötigt eine repräsentative Auswahl an Teilnehmer/innen. Es macht zum Beispiel wenig Sinn, eine Befragung zur Ehezufriedenheit oder zu Gewalt in Partnerschaften in einem Frauenhaus durchzuführen, oder eine Befragung zu den Folgen von Sexarbeit lediglich mit Aussteigern/innen. Insbesondere die Verwendung von Studierenden ist hier aufgrund des Alters/Bildungshintergrundes kritisch (siehe nächster Punkt). **✘ Showstopper: Verzerrungen bei der Stichprobe lassen sich nicht mit größeren Fallzahlen "ausgleichen" -- eine solche Studie ist außerhalb der selektiven Stichprobe wertlos.**

☐ Wurde eine bequeme aber weniger qualifizierte Stichprobe erhoben?

Studierende sind eine beliebte Stichprobe, weil sie leicht verfügbar sind und sie weitgehend kostenlos teilnehmen (z.B. wenn eine Teilnahme Voraussetzung für das Bestehen des Kurses ist). Es macht allerdings wenig Sinn, Studierende zum Beispiel wie Personaler zu behandeln und Personalentscheidungen treffen zu lassen. Noch schlimmer: Die Teilnehmer/innen sind oft nicht unabhängig voneinander. Oft waren sie alle im selben Kurs und beeinflussen sich gegenseitig (anstatt zufällig aus der Population aller Studierenden in Deutschland gezogen zu werden). Ganz bedenklich wenn das Thema des Kurses mit dem Thema der Untersuchung zu tun hat.

☐ Wurde nur eine sehr kleine Anzahl von Personen befragt?

Insbesondere bei aufwändigeren Studien (z.B. Interviews) ein Problem. Es gibt emotional sehr involvierende "Stories", aber die Anzahl ist zu gering, um davon auf die Gesamtbevölkerung zu generalisieren.

☐ Lag eine Selbstselektion vor?

Oft unvermeidlich bei Online-Befragungen von Nachrichten-Sites oder bei der Verwendung von #hashtags. Hier werden Personen angezogen, die nicht für die Gesamtbevölkerung repräsentativ sind (Wo "stolpern" sie über die Erhebung? In welchen Gruppen wird sie verbreitet?). **✘ Showstopper: Auch wenn es Aktivisten/innen gerne anders sehen, diese Methoden sind primär *outrage porn* und als Erhebungsinstrument wertlos.**

☐ Haben viele Personen die Studie abgebrochen/nicht geantwortet?

Zum Teil wird diese Angabe gerne verschwiegen, weil damit eine Art Selbstselektion vorliegt. Die Personen, die geantwortet haben, unterschieden sich neben ihrem Antwortverhalten häufig noch in anderen Merkmalen von denen, die nicht antworten. Damit ist die Stichprobe nicht mehr repräsentativ und die Ergebnisse können nicht mehr auf die Population generalisiert werden. Selbst wenn Tausende von ausgefüllten Fragebögen vorliegen, die entscheidende Frage ist: Wie viele (und wer) hat nicht mitgemacht?

Dr. Daniel Wessel ◇ http://danielwessel.org ◇ danwessel@me.com

Fragen

GIGO: Garbage In, Garbage Out.

☐ Werden die Fragen verschwiegen?

Zur Bewertung der Studie muss man Zugang zu allen tatsächlich gestellten Fragen haben. Das ist häufig ein Problem, weil oft nur Ergebnisse berichtet werden, die abstrakte Interpretationen der eigentlichen Fragen sind. Eine Gruppe hat einen "Wissensvorsprung", ist "bedrohter", "hat es schwerer". Die Frage ist immer: Was wurde konkret gefragt? Wenn die Fragen nicht angegeben oder zugänglich sind, ist höchste Vorsicht geboten. Mit "schlecht" konstruierten Fragen kann man wunderschön die Ergebnisse manipulieren. **✘ Showstopper: Sind die Fragen nicht zugänglich, ist die Studie nicht zu bewerten (und wenn sie von einer Interessengruppe durchgeführt wurde, als wertlos zu betrachten).** Tip: Die Befragung selbst einmal durchspielen: Hätte man als "normale" Person überhaupt die Chance dies im Fragebogen auch darzustellen?

☐ Wurden selbst entwickelte Fragebögen verwendet?

Es gibt gut entwickelte Fragebögen, deren Qualität bekannt ist. Man weiß, dass sie die dahinterstehenden Konstrukte gut abbilden. Vorsicht ist geboten, wenn Personen ihre eigenen Fragen entwickeln. Das sieht einfach aus (jede Billigzeitschrift hat Psycho-Tests), ist aber herausfordernd wenn man es richtig machen will.

☐ Wurden selektiv Vorteile/Nachteile abgefragt?

Zum Teil wird "*cherry picking*" betrieben und nur nach Vorteilen für die "bevorteilte" Gruppe und nur nach Nachteilen für die "benachteiligte" Gruppe gefragt. In vielen Fällen haben alle Gruppen **spezifische Vorteile und Nachteile**, die abgedeckt werden müssen um ein vollständiges Bild zu erhalten. Das betrifft sowohl Verhaltensweisen, als auch ganze Bereiche (z.B. STEM vs. non-STEM). **✘ Showstopper: Sind die Fragen "unfair", verzerren sie die Ergebnisse und die Befragung ist wertlos.**

☐ Wurden Beispiele verwendet, die nur/primär auf eine Gruppe zutreffen?

In Fragen werden oft konkrete Beispiele verwendet, z.B. bei Belästigungen ob eine Person "Schlampe" genannt wurde. Das trifft lediglich auf eine Gruppe zu (Männer werden üblicherweise nicht "Schlampe" genannt). Entsprechend ist die Verwendung der Beispiele kritisch (hier z.B. zusätzlich noch "Arschloch"). Andere Beispiele sind "Mädchen sind nicht gut in Mathe" vs. "Jungs sind nicht gut in Sprachen", die Vorstellungen von Attraktivität bei Frauen (dünn, Kurven) vs. bei Männern (groß, stark), etc.

☐ Wurden mehrdeutige Fragen verwendet?

"Die Manipulation und der Betrug beginnen da, wo wir die Begriffe so bestimmen, daß das Ergebnis uns ins Weltbild paßt."
Krämer (2000)

Doppelte Fragen sind hier sehr beliebt, ebenso Fragen, die so breit gefächert sind, dass unterschiedliche Fälle zusammengefasst werden. Als Beispiel die Frage, ob man jemals Geschlechtsverkehr hatte, während man "bewusstlos, unter Drogeneinfluss, betrunken, oder am Schlafen war". Bei einer ehrlichen Antwort würden wohl die meisten Erwachsenen zustimmen (wer hatte noch keinen betrunkenen Sex?). Das passt allerdings schlecht zu bewusstlos oder während man am Schlafen war. Wenn dann ein "Ja" auf diese Frage als Nachweis für "Sex ohne Einwilligung" (= Vergewaltigung) interpretiert wird, hat das wenig mit der Realität der Befragten zu tun. Ebenso bei häuslicher Gewalt wenn "festhalten" oder "schubsen" mit "schlagen" und "würgen" zusammen gefasst wird und später nur über extreme Verhaltensweise gesprochen wird. Falsch ist alles davon, aber es gibt Grade von Falsch.

☐ Wurden schwammige Begriffe verwendet?

Begriffe müssen klar definiert sein, Mehrdeutigkeit ist keine Tugend (z.B., Wie ist "sexuelle Belästigung" definiert? Ein Grüßen auf der Straße? Wiederholtes Ansprechen trotz Zeigen von Desinteresse? Ungewolltes Anfassen? Etc.; Was heißt "gleiche" Arbeit -- Jobbezeichnung? Output? Überstunden?). Insbesondere Interpretationen von Verhalten (siehe nächster Punkt) führt zu vielen Problemen.

☐ Wurde nach Interpretationen von Verhalten gefragt?

Das ist ein häufiges Problem, wenn sich Gruppen in ihrer Interpretation von konkreten beobachtbaren Verhaltensweisen unterschieden. Für einige Mitglieder/innen einer Gruppe kann eine Verhaltensweise (z.B. ein "retweet" auf Twitter) als Bedrohung aufgefasst werden, während die Mitglieder/innen der anderen Gruppe dies nicht als solches wahrnehmen. Probleme ergeben sich dann vor allem, wenn sich Leser/innen Vorstellungen von "Bedrohung" bilden, die sich nicht mit den als Bedrohung wahrgenommen Verhaltensweisen decken. Die Wahrnehmung einer Situation ist für die Betroffenen zwar real (und einigen Fällen sicherlich therapiebedürftig), Fragen sollten allerdings auf konkreten Verhaltensweisen beruhen die beide Gruppen wahrnehmen können.

☐ Ist der Antwortbereich eingeschränkt?

Zum Beispiel indem nicht der gesamte Skalenbereich angeboten wird: Eine Skala mit Startpunkt von "Männer und Frauen haben gleiche Chancen" zu "Männer haben Vorteile". Es verschweigt den eigentlichen Startpunkt "Frauen haben Vorteile" und verzerrt die Ergebnisse. Auch beliebt ist das Auslassen von Antwortalternativen wie "ist mir egal", "nicht zutreffend", etc. Gerade bei Online-Befragungen, bei denen man Fragen beantworten muss um die Befragung abzuschließen, kann man hier gut täuschen.

☐ Können Stereotype/Selbstwahrnehmungen die Befragung beeinflussen?

Ein großes Problem, wenn eine Gruppe aufgrund von vorherrschenden Stereotypen nicht über ihre Nachteile (oder Vorteile) sprechen möchte. Es ist zum Beispiel für Männer üblicherweise sehr negativ besetzt, über Probleme zu sprechen ("Bist du ein Mann oder eine Memme?", "Weichling!", "Warmduscher!", etc.). Ebenso können z.B. Manager/innen Nachteile für ihre Karriere vermuten, wenn sie sich mehr Zeit für die Familie wünschen. Hier ist es wichtig, dass der Befragung Vertrauen gegenüber gebracht wird, insbesondere was Anonymität betrifft. Überprüfbare Fragen nach konkretem Verhalten sind hier relevanter als Selbstwahrnehmungen, die zwar rational klingen, oft aber nur rationalisiert sind. Fragebögen führen hier vermutlich zu ehrlicheren Antworten als Interviews, in denen Befragte das Gefühl bekommen können, sich für Schwächen rechtfertigen zu müssen.

☐ Wurde nach den Erfahrungen anderer Personen gefragt?

Eine repräsentative Studie nimmt eine/n Teilnehmer/in als Repräsentant/in für Personen aus der Gesamtbevölkerung. Mit Fragen ob die Befragten Personen kennen, die etwas erlebt haben, wird diese Stichprobe aufgeweicht und damit verzerrt. Eine Person steht dann nicht mehr für einen Teil der Gruppe, auf die generalisiert wird. Noch schlimmer: Es wird Gedankenlesen oder "Wünsche lesen" vorausgesetzt ("meine Freunde/innen denken ...", "es wäre für andere hilfreich ..."). Einfaches Beispiel: Wenn ich wissen möchte, wie häufig Brillenträger sind, dann nehme ich eine repräsentative Stichprobe und frage die Personen, ob sie eine Brille tragen, nicht, ob sie eine Person kennen, die eine tragen. Der erste Fall führt zu einer guten Schätzung, der zweiten zu 99-100%.

☐ Wurde die Verursacher/innen-Frage unterschlagen?

Einige Studien nehmen einfach an, dass wenn eine Gruppe ein Problem hat (z.B. Diskriminierung), diese von "der anderen Gruppe" ausgeht. Wieso sollte sich eine Gruppe selbst diskriminieren oder schädigen? Das macht Sinn, wenn man die Welt mit dieser Schwarz-Weiß-Brille sieht. Aber die verwendeten Gruppendefinitionen täuschen darüber hinweg, dass Gruppen aus Individuen bestehen, die sich auch nach anderen Kriterien unterscheiden und für die diese Gruppenbezeichnung vielleicht nicht alles bestimmt bzw. überlagert. Neid, Missgunst, Konflikte basierend auf Statusunterschieden, Konkurrenz innerhalb der Gruppe -- es gibt viele andere Möglichkeiten, die zu bestimmten Problemen führen können. Entsprechend muss sauber erfasst werden, von wem problematisches Verhalten ausgeht. Ein "schönes" Beispiel für Konflikte unter Mädchen ist hier Simmons' (2011) "Odd Girl Out".

☐ Wurde die "typisches Verhalten vs. untypisches Verhalten"-Frage unterschlagen?

Es macht einen Unterschied von wem das Problem ausgeht -- z.B. ob eine Person 50 Personen belästigt oder ob 50 Personen jeweils eine Person belästigen. Hier wird häufig falsch generalisiert indem zum Beispiel nur nach Personen mit negativen Verhaltensweisen gefragt wird, nicht nach Personen mit neutralen oder positiven Verhaltensweisen. Oder wenn eine Person schon allein mit einem einmal gezeigten negativen Verhalten als Problemfall "abgestempelt" wird (was bei einigen, extrem negativen/kriminellen Verhaltensweisen Sinn macht, allerdings leicht zu weit gehen kann).

☐ Wurde bei der Einschätzung der Häufigkeit manipuliert?

Negative Ereignisse sind oft leicht abrufbar, das macht es schwer, die Häufigkeit gut zu schätzen. Idealerweise beziehen sich Fragen nach der Häufigkeit auf einen klar definierten, gut abrufbaren Zeitraum. Bei seltenen Ereignissen die letzten 6 oder 12 Monate, bei häufigeren Ereignissen kürzer. In diesen Fällen sollte man eine Option: "Ist passiert, aber außerhalb des Zeitraums" anbieten, um die Personen abzufangen, denen es wichtig ist, das Problem zu erwähnen und die sonst vielleicht falsche Angaben machen könnten.

Ergebnisse

☐ Ist der Datensatz "privat"?

Um die Datenauswertung und Interpretation zu überprüfen ist es häufig notwendig, in den Datensatz zu sehen. Unter anderem um sich im Detail die Verteilung der Antworten anzusehen. Nur wenige Organisationen stellen den Datensatz zur Verfügung -- unter anderem wird es unter Berufung auf Datenschutz verweigert. Zwar ist Datenschutz wichtig, aber es gibt z.B. die Möglichkeit, den Datensatz zu anonymisieren. **✘ Showstopper: Wird der Datensatz nicht geteilt (ggf. unter Auflagen), ist eine Studie -- insbesondere wenn sie von einer Interessengruppe durchgeführt wurde -- als wertlos zu sehen.** Das mag hart klingen, aber eine Studie die gut durchgeführt wurde hat nichts zu befürchten.

☐ Kannten die Auswerter/innen die Hypothesen/Gruppenzuordnungen?

Zum Teil müssen qualitative Aussagen in numerische Werte überführt werden. Es ist aber möglich, dass die selbe Verhaltensweise unterschiedlich wahrgenommen wird, je nachdem von welcher Gruppe sie kommt oder an welche sie gerichtet ist. Entsprechend sollten Auswertungen so weit wie möglich "blind" durchgeführt werden, d.h. ohne Kenntnis der Gruppenzuordnung. Das ist technisch leicht möglich (alle Datensätze der Personen in eine zufällige Reihenfolge bringen, durchnummerieren, unter anderem Namen speichern, Gruppenzuordnung wie z.B. Geschlecht, löschen, dann die Verhaltensbeschreibungen bewerten lassen, Ergebnisse nehmen und wieder mit der Zuordnung kombinieren). Wurde dies nicht gemacht, sind die Ergebnisse sehr zweifelhaft, weil die Auswerter/innen nicht neutral waren.

Dr. Daniel Wessel ◇ http://danielwessel.org ◇ danwessel@me.com

☐ Wurden Fragen fallen gelassen?

Eine beliebte fragwürdige Forschungsmethode: Man stellt mehrere Fragen/erhebt mehrere Variablen und lässt die fallen, bei denen nicht das gewünschte Ergebnis herausgekommen ist. Immer die gestellte Fragen mit den berichteten Ergebnissen vergleichen, und immer explizit nachfragen: Sind die berichteten Ergebnisse die Ergebnisse von allen gestellten Fragen? Oder wurden Fragen gestellt, die bei der Auswertung nicht mehr berücksichtigt wurden? Wenn ja, welche? Und warum? Schwerer zu überprüfen ist es, wenn Vorstudien durchgeführt wurden, bei denen rein aufgrund der Antworten "unpassende" Fragen aussortiert wurden. Eine Frage sollte zwar aussortiert werden, wenn sie theoretisch keinen Sinn macht oder nicht verstanden wird, wenn sie aber rein deswegen aussortiert wird, weil das Ergebnis nicht gefällt, dann hat man ein Problem.

☐ Wurden Fragen zusammengefasst?

In einigen Bereichen macht dies Sinn, dann sollte man allerdings nachweisen, dass diese Fragen auch dasselbe messen (Cronbach's Alpha, größer als .7 als Minimum). Hier muss man allerdings bei der Interpretation der Ergebnisse im Kopf behalten, was alles dieser Gruppierung zugrunde liegt (= nicht nur die extremsten Fragen).

☐ Wurden Antwortalternativen zusammengefasst?

Es ist leicht das Ausmaß und die Relevanz eines Problems zu verzerren, indem man es z.B. auf einer 6-stufigen Skala erhebt (starke Ablehnung bis starke Zustimmung) und die Ergebnisse nur als Zustimmung vs. Ablehnung betrachtet. Hier wird eine Einheitlichkeit in den Antworten vorgegaukelt, die nicht wirklich vorhanden war.

☐ Wurden nur bestimmte Stichproben bei der Auswertung verwendet?

Insbesondere wenn an mehreren Orten erhoben wurde ist es nicht ungewöhnlich, dass einige dieser Stichproben sich zufällig von anderen unterscheiden. Werden diese dann bewusst so ausgewählt, dass sich die gewünschten Unterscheide zeigen, hat das mit einer repräsentativen Studie nichts mehr zu tun. Und wenn dann auch noch so getan wird, als wäre dies eine zufällige Auswahl, dann ist man tief im Täuschungsbereich (Krämer, 2000).

☐ Wurden unpassende Personen aussortiert?

Es macht zum Teil Sinn, "Outlier" auszusortieren. Eine 72-jährige Studentin ist nicht typisch. Es gibt Personen, die die Fragen nicht verstanden haben oder bewusst Unsinn angekreuzt haben. Allerdings gibt es Forscher/innen, die unterschiedliche Kriterien für das "aussortieren" haben, je nachdem ob es die Ergebnisse in die gewünschte Richtung beeinflusst oder nicht. Datenbereinigung gehört zur Forschung dazu aber genau schauen: Was wurde warum bereinigt?

☐ Wird die (Un-)Sicherheit der Zahlen verschwiegen?

Insbesondere wenn sehr genaue Zahlen (nicht gerundet, viele Nachkommastellen) angegeben werden, ist Vorsicht geboten. Ist die Messung wirklich so genau? Auch relevant: Es gibt in der Statistik Konfidenzintervallen, die anzeigen, mit welcher Sicherheit man von der Stichprobe auf die Gesamtbevölkerung schließen kann. Häufig wird der Mittelwert der Stichprobe (z.B.: ich wähle zufällig 50 Deutsche aus, messe ihre Größe, und bilde davon den Durchschnitt) als Schätzung für den Mittelwert der Population verwendet (der Durchschnitt der Größe *aller* Deutschen). Und das ist mit Vorsicht zu genießen (ich kann Pech haben und viele kleine Deutschen erwischen). Ohne jetzt in Statistik einzutauchen, *über*vereinfacht gesagt, ein 95% Konfidenzintervall bildet einen Bereich ab, in denen der Mittelwert der Population mit 95%iger Wahrscheinlichkeit liegt. Der Durchschnitt meiner Stichprobe mag von dem Durchschnitt der Population abweichen, aber mit 95%iger Wahrscheinlichkeit liegt er irgendwo in diesem Bereich. Und der Bereich kann manchmal peinlich groß sein.

☐ Wird die statistische Signifikanz unterschlagen?

Nur weil sich Gruppen deskriptiv unterscheiden (Wert von Gruppe A ist anders als Wert von Gruppe B), heißt dies nicht, dass diese Unterschiede auch statistisch bedeutsam sind. Wir erheben Stichproben für Gruppe A und B und in diesen unterschieden sich die Ergebnisse der Personen. Die Unterschiede können rein zufällig zustande gekommen sein. Ein statistischer Signifikanztest liefert häufig einen p-Wert der aussagt: Angenommen es gäbe keine Unterschiede zwischen beiden Gruppen, wie wahrscheinlich sind dann die gefundenen (oder extremere) Daten? Ist diese Wahrscheinlichkeit unterhalb von .05 wird von einem statistisch signifikanten Unterschied gesprochen. Das heißt aber nicht, dass wir aufgrund dieser Daten schließen können, dass die Wahrscheinlichkeit, dass es keine Unterschiede gibt, weniger als .05 ist. Kurz gesagt: Diese Tests haben ihre Schwächen, aber wenn die statistische Signifikanz nicht einmal angegeben wird, dann ist Vorsicht geboten.

☐ Wird die Effektstärke unterschlagen?

Nur weil ein Unterschied statistisch signifikant ist, heißt dies noch lange nicht, dass ein solcher Unterschied auch bedeutsam ist. Statistische Signifikanz hängt u.a. davon ab, wie groß die Stichprobe ist. Je mehr Personen man erhebt, desto eher werden auch minimale Unterschiede zwischen Gruppen "statistisch signifikant" (nicht unbedingt in die gewünschte Richtung). Entsprechend ist die Effektstärke relevant. Es gibt diverse Maße für kleinen, mittleren oder großen Effekt. In jedem Fall lohnt es sich die Ergebnisse einmal praktisch zu übertragen: Für jedes X, was bekommen wir an Y? Und sich dann kritisch zu überlegen: Tritt dies wirklich ein, oder ist die Realität komplizierter und für jedes X bekommen wir auch Z (und vieles andere), was Y den Wert nimmt?

Dr. Daniel Wessel ◇ http://danielwessel.org ◇ danwessel@me.com

Diskussion der Ergebnisse

☐ Werden Ursachenbehauptungen aufgestellt?

Um Ursachenbehauptungen aufzustellen, muss man Experimente durchführen. Kurz: Man teilt Personen *zufällig* in mindestens zwei unterschiedlichen Bedingungen auf, in denen sie eine unterschiedliche Behandlung erfahren (z.B. Instruktion A vs. Instruktion B; Medikament A vs. Placebo B, etc.). Mit Befragungen kann man keine Ursachenbehauptungen aufstellen, weil die Personen nicht zufällig unterschiedlichen Bedingungen zugeordnet wurden (und noch diverse andere Probleme bestehen). **✗ Showstopper: Eine Befragung/Erhebung die Ursachenbeziehungen behauptet ist Quatsch.** Sie verwechselt eine statistischen Zusammenhang mit einer Ursachenbeziehung. Standardbeispiele: Männer und Frauen nach ihrem Einkommen befragen und dann zu schließen: Männer verdienen *aufgrund Ihres* Geschlechts mehr. Es ignoriert z.B. die ausgeübten Berufe, den Anteil von Vollzeit vs. Teilzeit, die Anzahl der Überstunden, die kumulative Arbeitszeit über die Lebenszeit, und viel mehr. Es mag plausibel erscheinen, ein sehr salientes Merkmal wie Geschlecht herauszugreifen und dies verantwortlich zu machen, es bricht bei genaueren Analysen aber üblicherweise zusammen. Außerhalb von streng biologischen Unterschieden ist Geschlecht eine *grottige* aber leider viel-gehypte Variable. Die Zeit als Variable ist häufig ebenso problematisch zu sehen.

☐ Geht die Diskussion nur selektiv auf die Ergebnisse ein?

Gerade bei längeren Berichten wird oft im "*Executive Summary*" oder in Pressemitteilungen nur auf gewünschte Ergebnisse eingegangen. So kann leicht der Eindruck erweckt werden, dass eine Gruppe besonders betroffen ist oder dass die Unterschiede größer sind, als die Studie tatsächlich gezeigt hat. Das ist insbesondere dann ein Problem, wenn die weiteren Verbreiter/innen nicht mehr als das "*Executive Summary*" lesen.

☐ Geht die Diskussion weit über die Ergebnisse hinaus?

Jeder interpretierende Aussage die über die Studie gemacht wird, muss in den Daten verankert sein. Wenn diese signifikanten und bedeutenden Unterschiede nicht vorliegen wird Augenwischerei betrieben. Autoren/innen können gerne spekulieren und weitere Forschung anregen, müssen die Spekulationen aber als solche kenntlich machen.

☐ Definiert die Diskussion die Begriffe neu?

Relevant wenn Fragen plötzlich in ihrer Bedeutung neu interpretiert werden oder sie als Anzeichen für tieferliegende Probleme verwendet werden. Die Diskussion wird dann fast unmerklich von den eigentlichen Daten getrennt. Es ist eine Sache, dass bestimmte Verhaltensweisen gefunden wurden, eine andere, was diese bedeuten. Auch beliebt: Befürchtungen werden unmerklich zu "Realitäten": "Gruppe x fühlt sich bedroht, also müssen wir die Sicherheit erhöhen", anstatt zum Beispiel die übersteigerte Angst von Gruppe x zu adressieren, wenn diese laut Statistik weniger häufig zum Opfer wird als Gruppe y.

☐ Wird eine künstliche Vergleichsbasis geschaffen?

Beliebt wenn die andere Gruppe überhaupt nicht erhoben wurde und über Rhetorik eine Vorstellung der anderen Gruppe als besonders erfolgreich geschaffen wird. Aber auch, wenn die Vergleichsbasis künstlich eingeschränkt wird. Eine insgesamt mittelmäßige Person kann zum Beispiel als "besten Fachvertreter unter 35 an der ganzen Fakultät." beschrieben werden (Krämer, 2000), oder ein Buch kann bei Amazon auf Platz 1 sein (im Kindle-Store im Bereich 44-64 Seiten unter "Selbsthilfebücher" bei kostenpflichtigen Angeboten). Insbesondere bei zeitlichen Vergleichen wird oft eine ideologisch passende Auswahl getroffen.

☐ Wird ein Goldstandard unreflektiert eingeführt?

Welche Ergebnisse wären wünschenswert gewesen und weswegen? Was ist das Ziel? Im Fall von zwei Gruppen meist keine (statistisch signifikanten und bedeutsamen) Gruppenunterschiede. Hier wird aber oft eine Ideologie als Standard gesetzt: "Gruppe A und Gruppe B sind 'gleich', also darf es keine Gruppenunterschiede geben, falls diese dennoch existieren deutet dies auf eine unfaire Benachteiligung hin", bei der andere Unterschiede in, z.B., Interessen, Sozialisation, biologischen Voraussetzungen, etc. pp. stillschweigend ignoriert werden. Chancengleichheit wird hier mit Ergebnisgleichheit verwechselt. Auch beliebt: Vertreter/innen einer Gruppe entscheiden für alle, was "normal" oder "erwünscht" ist (obwohl ein Vergleich auch Mitglieder/innen der anderen Gruppe betrifft) oder was Personen allein aufgrund einer nicht selbst gewählten Gruppenzugehörigkeit mit ihrem Leben anfangen sollen oder müssen.

☐ Wird verschleiert, wie viele Personen wirklich betroffen sind?

Es mögen in einer untersuchten Gruppe prozentual sehr viele Personen betroffen sein, aber diese Personen können nur einen kleinen Teil der Gesamtbevölkerung ausmachen. Das Gesamtbild ist hier entscheidend.

☐ Wurden Alternativerklärungen unterschlagen?

Selbst wenn alles auf eine "Ursache" hinaus läuft, kann diese "Eindeutigkeit" nur daran liegen, dass andere Erklärungen nicht in Betracht gezogen wurden. Andere Interpretationen sollten (fair) diskutiert und ausgeschlossen werden. Ist es z.B. möglich, dass Mitglieder der "benachteiligten Gruppe" aufgrund von mangelnden Kompetenzen oder schlechter Passung scheitern? Das heißt, wurden diese Variablen als mögliche Unterschiede zwischen beiden Gruppen berücksichtigt? Wurden situationale Einflussfaktoren berücksichtigt?

Dr. Daniel Wessel ◇ http://danielwessel.org ◇ danwessel@me.com

☐ Wird eine emotionale Sprache verwendet?

Insbesondere wenn emotional involvierende Fallgeschichten einer Gruppe hervorgehoben werden. Wenn Autoren/innen emotional "argumentieren" sind die Argumente oft zweifelhaft und eine kritische Diskussion soll im Keim erstickt werden. Berechtigte Kritik wird dann häufig als Unterstützung des Problems uminterpretiert. Auch beliebt ist der Versuch, Scham hervorzurufen. Zum Beispiel über den Vorwurf, wenn man die Schlussfolgerungen nicht unterstützt ist man Teil des Problems und nur über ~~unreflektierte~~ enthusiastische Unterstützung kann man sich von diesem Verdacht ~~freikaufen~~ befreien (Der Vorwurf ist Quatsch und man sollte ihn als emotionale Manipulation offenlegen, die in einer rationalen Diskussion unter Erwachsenen keinen Platz hat.).

☐ Wird mit Diagrammen getäuscht?

Häufig ist hier, dass Achsen abgeschnitten, gestaucht oder gedehnt werden, werden um Unterschiede größer darzustellen. Beliebt ist auch die Verwendung von zwei verschiedenen y-Achsen (wie wurden die Einheiten jeweils gewählt)? Auch mit Flächen oder Volumen wird häufig getäuscht, sie wirken oft größer, als sie sind, sowie mit Histogrammen und Dichtekarten (wie wurden die Wertebereiche bestimmt?). Auch beliebt: Man bildet nicht Entwicklungen ab sondern Wachstumsraten von Entwicklungen, oder Wachstumsraten von Wachstumsraten (vgl. Krämer, 2000).

☐ Sind die verallgemeinerten Zahlen unrealistisch/widersprüchlich?

Gerade wenn man von Stichproben auf die Population generalisiert, sollte man überlegen, ob die generalisierten Zahlen realistisch sind. *Zahlen müssen im Kontext Sinn machen.* Schönes Beispiel von Krämer (2000) sind die historischen Zahlen von Heeresgrößen im Vergleich zur Größe des Schlachtfeldes oder den verfügbaren Transportkapazitäten. Aber auch bezüglich der Vor- und Nachbedingungen können Zahlen schnell unglaubwürdig werden. Ereignisse haben Konsequenzen und die führen meist zu anderen, gut messbaren Ereignissen. Und diese Zahlen müssen konsistent sein. Beispiel von Perry (2015): Eine Quelle sagt aus, dass 1 von 5 Personen (= 20%) einer klar definierten Gruppe Opfer eines Verbrechens werden und gleichzeitig wird behauptet, dass nur 12% dieser Verbrechen angezeigt werden. Wenn eine andere Quelle jetzt auf 13 Anzeigen hinweist, bei einer Gruppengröße von 6355 Personen, dann widersprechen sich beide Quellen. Sind 13 Anzeigen 12%, dann sollten insgesamt (100%) 108.3 Verbrechen vorgelegen haben, davon 95.3 ohne Anzeige und die genannten 13 mit. Aber 20% von 6355 sind 1271 und das unterschiedet sich ganz gewaltig von den vermuteten 108.3 Verbrechen. Wenn 108 stimmen würde, wäre die Opferquote statt bei 1:5 bei 1:59. Jedes Opfer ist eines zu viel, aber mit verzerrten Zahlen wird man kaum passende Interventionen durchführen können. Noch schlimmer: *Man wird nicht überprüfen können, ob diese Interventionen auch tatsächlich funktionieren und die potentiellen Opfer schützen*!

☐ Haben die vorgeschlagenen Lösungen unerwünschte Konsequenzen?

Durch einen engen Fokus einer Studie kann schnell der Eindruck entstehen, als gäbe es nur dieses Problem. Und auch wenn die diskutierte Intervention ein Problem lösen kann, kann dies nur ein kurzfristiger Effekt auf ein spezifisches Symptom sein. Langfristig, bzw. im Kontext, kann es größeren Schaden anrichten. Die Frage ist hier, welche unintendierten Konsequenzen die Problemlösung hat, ob es nur auf Symptome eingeht aber Ursachen ignoriert, ob es zwar eine (wahrgenommene) Benachteiligung reduziert aber dafür (objektiv) andere Benachteiligungen schafft, etc. pp. Eine gute Lösung ist für alle Gruppen fair, im Sinne von Rawls' "*veil of ignorance*". Problematisch sind hier vor allem Lösungen, die andere Perspektiven sowie die Freiheit der Beteiligten ignorieren (im Sinne von Entscheidungsfreiheit, Interessen, etc.). Auch wenn es um Finanzierungen geht (nichts ist "kostenlos"), wird oft manipuliert ("Ein Freibetrag für Kinder freut die Steuerzahler sehr, eine Extrasteuer für kinderlose Arbeitnehmer regt die Steuerzahler auf.", Krämer, 2000)

☐ Betrifft die vorgeschlagene Lösung nur einen Teil der Betroffenen?

Die meisten (nicht-biologischen) Probleme betreffen Personen aus unterschiedlichen Gruppen, wenn auch zum Teil im unterschiedlichen Ausmaß. Es ist durchaus möglich, gezielt die Personen anzusprechen, die von einem Problem betroffen sind (z.B., Probleme damit haben, ihre Ideen in Diskussionen einzubringen). Das löst dieses Problem besser als Personen aufgrund einer Gruppenzugehörigkeit als "Problemträger/innen" zu definieren (z.B., ein Training für Frauen in Organisationen). Sobald eine Gruppenzugehörigkeit gewählt wird, die nicht 100% vs. 0% ist, schafft dies a) Ressentiment bei Personen, die aufgrund der Gruppenzugehörigkeit ausgeschlossen werden, obwohl sie dieses Problem haben, und b) Ressentiment bei Personen, die aufgrund der Gruppenzugehörigkeit als Problemträger/innen definiert werden, obwohl sie dieses Problem *nicht* haben. Und das führt zu berechtigten und unnötigen Konflikten. Sowohl das Arbeitsleben als auch der Alltag bieten genug Herausforderungen, die es zu überwinden lohnt, auch ohne sich für Stellvertreter/innen-Kriege von Ideologien herabzuwürdigen.