Spektrum

Hintergrund

- 24.01.2022

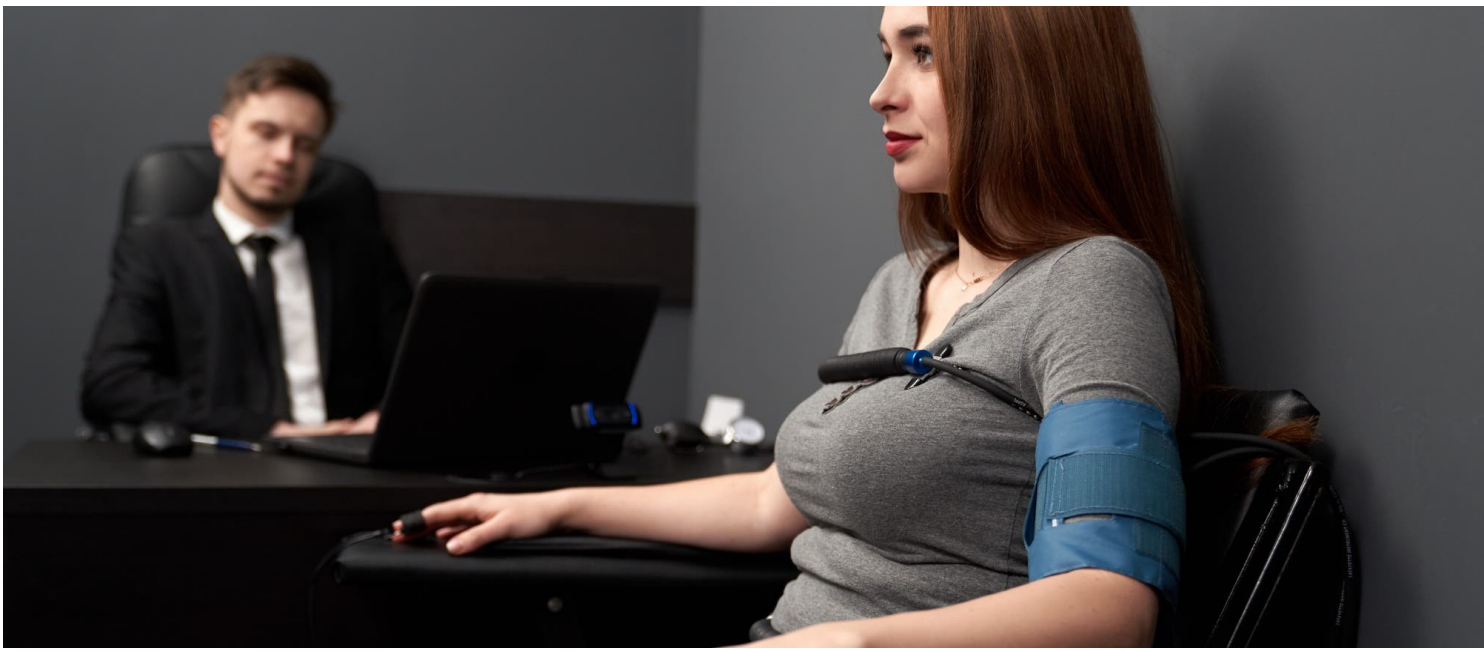
Forschungspraktiken: Wie sieht eine gute psychologische Studie aus?

»P-Hacking«, »cherry picking«: Nur wer solche Schummeleien erkennt, kann den Wert einer empirischen Studie richtig einschätzen. Wie auch Laien gute von schlechten Praktiken unterscheiden können - ein Leitfaden.

von Jan Schwenkenbecher

Jan Schwenkenbecher

Der studierte Psychologe arbeitet als Wissenschaftsjournalist im Rhein-Main-Gebiet.



Vor rund zehn Jahren geriet die Psychologie in die Schlagzeilen: Die Fälschungen des niederländischen Sozialpsychologen Diederik Stapel flogen auf, und Daryl Bems vermeintliche Belege für »Präkognition«, eine Art übersinnliche Vorahnung, ließen sich nicht bestätigen. In der Folge scheiterten zahlreiche weitere Versuche, klassische Studien zu wiederholen, ihre Befunde zu »replizieren«. Die so genannte Replikationskrise erfasste auch andere Fächer. Eine der Hauptursachen: schlechte Forschungspraktiken. Und die waren schon vor den beiden Skandalen weit verbreitet, wie mehrere einflussreiche Arbeiten deutlich machten.

Damit kam Bewegung in die Psychologie. Immer mehr Forscherinnen und Forscher verzichten seitdem auf zweifelhafte Methoden wie »p-Hacking« oder »cherry picking« und setzen stattdessen auf »Registered Reports« und »Open Data«. Was hinter diesen Begriffen steckt, erläutert der folgende Leitfaden. Er schildert gute Wissenschaft in vier Akten – von der Planung über Design und Auswertung bis zur Veröffentlichung.

1. Akt: Die Vorarbeit

Gutes wissenschaftliches Arbeiten beginnt schon mit der Formulierung der Forschungsfrage. Die erste Entscheidung steht an: »konfirmatorisch« oder »explorativ« forschen? Wer konfirmatorisch vorgeht, hat eine Hypothese und möchte sie überprüfen, im besten Fall: bestätigen (von lateinisch: confirmare = bestätigen). Eine explorative Arbeit hingegen erkundet unbekannte Gewässer (explorare = erkunden) und entwickelt aus den Entdeckungen im Nachgang Hypothesen, die es dann in weiteren Studien zu prüfen gilt.

Entscheidend bei der konfirmatorischen Hypothesen-Entwicklung: erst die Literatur lesen, dann daraus Hypothesen ableiten. Die Reihenfolge macht einen Unterschied, denn eine gut klingende Hypothese ist schnell erdacht und die passende Literatur gezielt herausgepickt, um die schöne Hypothese zu rechtfertigen. Dabei besteht die Gefahr, anders lautende Befunde beiseitezuschieben. Eine gute Hypothese muss zwar nicht zur gesamten verfügbaren Literatur passen, muss sie aber vollständig berücksichtigen.

Das könnte Sie auch interessieren: Spektrum Kompakt: Kuriose Experimente

Eine zunehmend beliebte Variante des konfirmatorischen Arbeitens ist, die Studien von anderen zu replizieren, also zu wiederholen. Wozu? Jeder Befund gilt zunächst nur für den Ort, den Zeitpunkt und das Setting der betreffenden Studie. Doch je öfter eine Studie mit dem gleichen Ergebnis wiederholt wird, umso wahrscheinlicher wird es, dass es auch wirklich stimmt und auf andere Kontexte übertragen werden kann.

Und schließlich kann man bei der Studienplanung noch auf die Team-Zusammenstellung achten: Wer programmieren muss, der zieht einen Programmierer oder eine Programmiererin hinzu. Wer höhere Statistik braucht, sucht sich einen Statistikprofi. Ohnehin gilt: Interdisziplinäres Arbeiten, Großprojekte mit weltweiten Erhebungen und der Aufbau gemeinsamer Datenbanken liegen im Trend. Das mag zwar aufwändiger sein als eine Online-Fragebogenstudie mit 100 Versuchspersonen aus dem Familien- und Freundeskreis, doch dafür sind die Ergebnisse belastbarer.

2. Akt: Das Studiendesign

Steht die Forschungsfrage, leitet sich daraus oft schon das grobe Gerüst für das Studiendesign ab. Dennoch gibt es einige Punkte, die Forschende berücksichtigen können. Einer davon: nicht unnötig viele Variablen zu erheben.

Man könnte zum Beispiel noch schnell einen Persönlichkeitsfragebogen dazulegen, einen IQ-Test oder, oder, oder. Doch je mehr Variablen, je mehr mögliche Effekte und Zusammenhänge analysiert werden, desto wahrscheinlicher wird einer von ihnen durch reinen Zufall signifikant. Forschende erhöhen so zwar ihre Chancen, ihre Arbeit in einem wissenschaftlichen Journal zu veröffentlichen. Falsch ist so ein Vorgehen trotzdem, denn den vermeintlichen Zusammenhang gibt es eigentlich gar nicht. Unterm Strich schadet ein solcher Zufallsbefund der Forschung, denn in Replikationsstudien wird er nicht erneut gefunden. Was dem Problem vorbeugt: nur diejenigen Variablen zu erheben, die wirklich nötig sind, um die Hypothese zu testen, oder vorab festlegen, auf welchen Variablen der Fokus liegen soll. Im Nachhinein lässt sich eine Analyse von vielen Variablen notfalls statistisch korrigieren.

> Eine Poweranalyse gibt an, wie viele Versuchspersonen nötig sind, damit sich der gesuchte Effekt höchstwahrscheinlich zeigt – wenn es ihn wirklich gibt

Stehen die Variablen fest, ist die nächste Frage, wie groß ein Effekt oder Zusammenhang ausfallen muss, damit er auch praktisch bedeutsam ist. Angenommen, es geht um die Wirksamkeit eines IQ-Trainings: Um wie viele Punkte müsste der IQ steigen? Ist der Mindesteffekt bestimmt, gibt eine so genannte Poweranalyse an, wie viele Versuchspersonen nötig sind, damit sich der gesuchte Effekt höchstwahrscheinlich in den Daten zeigt – wenn es ihn denn wirklich gibt.

Wer seine Stichprobengröße auf diese Weise festlegt, der gerät auch nicht in die Versuchung, zwei weitere schlechte Forschungspraktiken anzuwenden, wenn die Studie schon läuft. Erster Fehler: den Zwischenstand immer mal wieder anzuschauen und die Studie zu stoppen, sobald sich in den Daten der erhoffte Effekt abzeichnet – auch wenn die angedachte Versuchspersonenzahl noch gar nicht erreicht ist. Zweiter Fehler: so lange immer mehr Teilnehmer zu testen, bis schließlich ein signifikantes Ergebnis vorliegt, auch wenn die geplante Stichprobengröße überschritten ist.

Zwei weitere Wege schützen ebenfalls vor solchen (und anderen) schlechten Praktiken: Präregistrierungen und »Registered Reports«. Bei einer Präregistrierung laden Forschende ein Protokoll ihres Designs auf eine öffentlich einsehbare Präregistrierungsplattform. Dort wird es »eingefroren«, kann also nicht mehr nachträglich geändert werden. Spätere Abweichungen fallen auf und wollen gut begründet sein.

Die »Registered Reports« gehen noch eine Stufe weiter. Es handelt sich dabei um ein Format, das von immer mehr Journals angeboten wird. Forschende senden vor der Datenerhebung ihr Studiendesign ein, das Journal unterzieht es einem ersten Peer-Review, also einer Begutachtung durch Fachkolleginnen und -kollegen. Diese schlagen Änderungen vor, wenn sie das Design für ungeeignet halten, die Forschungsfrage zu beantworten. Eine weitere Peer-Review-Runde folgt, wenn die Studie durchgeführt und der Artikel geschrieben ist. Im Anschluss veröffentlicht das Journal die Studie – ganz gleich, ob

sich die vorab formulierten Hypothesen bestätigt haben. Registered Reports verschieben so den Fokus von aufregenden Ergebnissen auf methodisch gut gemachte Forschung.

3. Akt: Die Auswertung

Die größte Gefahr liegt im p-Hacking. Der Wert p gibt an, ob ein Ergebnis signifikant, also statistisch gesehen bedeutsam ist. Er sagt aber nichts darüber aus, ob ein Ergebnis praktisch bedeutsam ist. Ausgangspunkt ist die Nullhypothese, die davon ausgeht, dass es in Wahrheit gar keinen Effekt gibt, zum Beispiel keinen Unterschied zwischen zwei Gruppen. Wie oft würden – unter Annahme dieser Nullhypothese – die Unterschiede größer ausfallen als in den vorliegenden Daten, wenn die Studie unendlich oft in gleicher Weise wiederholt würde? Ist die Wahrscheinlichkeit kleiner als 5 Prozent, also p kleiner als 0,05, gilt ein Ergebnis in der Regel als statistisch signifikant. Die simple Zweiteilung in signifikante und nicht signifikante Ergebnisse wird der Realität aber nicht gerecht.

Um den magischen Wert zu unterschreiten, haben sich dennoch allerlei Tricks eingebürgert, in der Fachwelt bekannt unter dem Begriff p-Hacking. Darunter fallen das schon erwähnte vorzeitige Stoppen und Verlängern einer Datenerhebung. Andere Methoden: »störende« Versuchspersonen auszuschließen oder die Zusammensetzung von Gruppen so lange zu ändern, bis das Ergebnis passt, natürlich mit vorgeschobenen, vermeintlich guten Argumenten. Neben dem p-Hacking gibt es eine weitere fragwürdige Forschungspraxis: das HARKing, kurz für »Hypothesizing After the Results are Known«, das Aufstellen von Hypothesen, nachdem die Ergebnisse bereits bekannt sind.

> Veröffentlichung, Zitierungen, Forschungsgelder: All das steht und fällt mit der Frage, ob p kleiner ist als 0,05

HARKing bedeutet in der Praxis: Die Forschenden formulieren ihre Hypothesen, erheben Daten und stellen fest, dass die vorab vermuteten Effekte zwar nicht auftreten, andere aber schon. Also schreiben sie ihre Hypothese um, schon passt alles zusammen und einer Veröffentlichung steht nichts mehr im Weg. Ethisch ist das nicht einwandfrei. Explorative Überraschungsbefunde sind natürlich auch etwas wert; oft treiben gerade sie den wissenschaftlichen Fortschritt an. Aber sie müssen auch als explorativ dargestellt und im Anschluss konfirmatorisch abgesichert werden.

Das p-Hacking – und damit eine Ursache der Replikationskrise – wurzelt vor allem darin, dass das kleine p einen so großen Stellenwert hat. Zugespitzt formuliert: Veröffentlichung, Zitierungen, Forschungsgelder – all das steht und fällt mit der Frage, ob p kleiner ist als 0,05. Dabei wurde das Signifikanzniveau von fünf Prozent einst mehr oder weniger willkürlich festgelegt. Seit Jahrzehnten wird über eine Abkehr vom p-Wert diskutiert, Alternativen verbreiten sich aber nur langsam.

Das könnte Sie auch interessieren: Spektrum Kompakt: Statistik – Zahlenspiele mit Mehrwert

Mögliche Lösungen: den Fokus vom p-Wert weg und auf die Effektstärke zu richten, also wie groß der Effekt ist. Oder an Stelle eines einzelnen p-Werts das Konfidenzintervall für die Effektstärke angeben. Damit kann man die Präzision eines geschätzten Effekts angeben, zum Beispiel, ob ein IQ-Training den IQ wahrscheinlich um 5 bis 6 Punkte oder um 1 bis 10 Punkte steigert.

Nicht zuletzt gilt: Kein Ergebnis ist auch ein Ergebnis. Allerdings genügt ein p-Wert größer 0,05 nicht, um den Schluss zu ziehen, dass es zwischen zwei Gruppen keine Unterschiede oder zwischen zwei Variablen keinen Zusammenhang gibt. Diese Nullhypothese lässt sich nach Ansicht vieler Fachleute besser mit der bayesianischen Statistik überprüfen. Die Wissenschafts-Community wertschätzt das zunehmend – immer mehr Journals veröffentlichen Nullergebnisse oder setzen auf neue Analysemethoden.

4. Akt: Die Veröffentlichung

Beim Veröffentlichen ist Transparenz das höchste Gut. Das gilt zunächst mal für die in der Studie gewonnen Daten. Dazu sollten sie nach dem FAIR-Prinzip aufbereitet sein: FAIR, das steht für »findable, accessible, interoperable, re-usable« – also auffindbar, zugänglich, verknüpfbar und wiederverwendbar. Das gilt es schon bei der Planung zu berücksichtigen. Denn damit auch

Fachkolleginnen und -kollegen die Daten nutzen dürfen, müssen Teilnehmerinnen und Teilnehmer von Studien vorab dazu ihr Einverständnis geben.

Doch nicht nur die gewonnenen Daten sollten transparent sein. Ebenso umfassend sollten Studienautoren über Methoden, Rechenwege, widersprüchliche Ergebnisse, die Finanzierung ihrer Arbeit und etwaige Interessenskonflikte berichten. Und alles andere Berichtenswerte auch. Das Ziel ist, dass fremde Forschende alle nötigen Informationen haben, um die vorliegende Studie exakt nachstellen zu können. Tunlichst zu vermeiden ist das so genannte »cherry picking« (zu Deutsch in etwa: die Kirschen herauspicken). Damit ist das Auswählen von erwünschten und das Verheimlichen von unerwünschten Daten und Ergebnissen gemeint.

Ist alles erledigt und der Artikel geschrieben, steht der letzte Schritt an: die Arbeit bei einer Fachzeitschrift mit Peer-Review-Verfahren einzureichen. Aber bei welcher? Statt (nur) die Bedeutung des Journals in der wissenschaftlichen Community zu berücksichtigen, können Forschende darauf achten, dass in dem Journal eine »Open Access«-Publikation möglich ist – entweder als »Gold Open Access« (dann bezahlt man dafür) oder als »Green Open Access« (dann kann man den Artikel zweitveröffentlichen, zum Beispiel auf der eigenen Website). In beiden Fällen ist die Studie für jeden Menschen frei zugänglich.

Auf einen Blick
Dos und Don’ts in Psycho-Studien

1. **Planung und Design:** Am Anfang steht meist die Hypothese, und die sollte sich aus der Fachliteratur ableiten – nicht erst im Nachhinein aus den Daten (»HARKing«). Wer die Studie vorab registriert oder sogar das Design im Peer-Review prüfen lässt, sorgt für Transparenz. Wie viele Versuchspersonen nötig sind, zeigt eine Poweranalyse.
2. **Auswertung:** Geschummelt wird vorzugsweise mit »p-Hacking«. Dazu zählt unter anderem, Versuchspersonen aus der Analyse auszuschließen, die nicht zu den erhofften Ergebnissen beitragen, oder die Erhebung dann zu stoppen, wenn die Daten das gewünschte Ergebnis bringen. Weil der p-Wert, das Maß für die statistische Signifikanz der Ergebnisse, mit vielen Problemen verbunden ist, arbeiten Fachleute heute zunehmend mit anderen Maßen wie Effektstärken und Konfidenzintervallen.
3. **Veröffentlichung:** Gute Forschung bedeutet auch, die Studie in jeder Hinsicht transparent darzustellen. Das bedeutet zum einen, dass die Daten auffindbar, zugänglich, verknüpfbar und wiederverwendbar sein sollten. Zum anderen soll der Artikel auch unbequeme Ergebnisse und etwaige Interessenkonflikte offenlegen.