Direkt zum Inhalt, zur Navigation

Logo: DGPs - Deutsche Gesellschaft für Psychologie

Navigation

Suche

PsychSpider ZPID-Suche Suche im DGPs-Angebot Suchen

- Startseite
- Kontakt
- Sitemap
- English

Hauptmenü

- DGPs im Profil
- Fachgruppen
- Mitglieder
- Unsere Experten
- Jungmitglieder
- Der Kongress
- Dienstleistungen

- Psychologie studieren
- Faszination Forschung

- **Aktuelles und Termine**
  - **Diskussionsforum**
  - News
  - Termine
  - Aktuelle Mitteilungen der DGPs
  - Tagesordnungen
- Presse

Login für Mitglieder

Benutzeranmeldung

Geben Sie Ihren Benutzernamen und Ihr Passwort ein, um sich an der Website anzumelden:

Name:

Passwort:

Anmelden

Kennwort vergessen?

Mitglied werden

Aktuelles und Termine > Diskussionsforum

Diskussionsforum: Qualitätssicherung in der Forschung

Hier finden Sie Diskussionsbeiträge von DGPs-Mitgliedern zur Stellungnahme des Vorstands "Replikationen von Studien sichern Qualität in der Wissenschaft und bringen die Forschung voran".

DGPs-Mitglieder können ihre Diskussionsbeiträge an Dr. Bianca Vaterrodt (referentin(at)dgps.de) senden.

20.09.2015

Für eine bessere Passung von Gegenstand und Methode und mehr Qualität statt Quantität

PD Dr. Matthias Reitzle

Nach dem letzten Beitrag von Renkewitz scheint das Pendel wieder ein wenig in Richtung Relativierung der Ergebnisse des RP auszuschlagen. Fahrenberg misst dem Befund bei allen Einschränkungen einen Weckrufcharakter zu. Stroebe hingegen findet die Befunde des RP erwartbar, weil psychologische Studien naturgemäß für mehr nicht genug Power hätten. Auch er führt die metaanalytisch ermittelten 68 Prozent ins Feld, die in früheren Beiträgen (Fahrenberg, Renkewitz et al., Schönbrodt et al., Witte), zumindest für mich überzeugend, in Frage gestellt wurden. Für ihn liegt die Lösung des Problems in Metaanalysen, die Studien wie das RP überflüssig erscheinen lassen: „Since meta-analysis permits us to evaluate the validity of research without the need to collect new data, one can question whether the meagre results of this project justify the time investment of 270 researchers and thousands of undergraduate research participants."

Nun unterliegen Metaanalysen ebenfalls einer lebhaften Qualitätsdebatte, vor allem im Hinblick auf die massive Heterogenität der Einzelstudien (z. B. Maier & Möller, 2010; Brugha et al., 2012) sowie den Mangel an verbindlichen Qualitätskriterien für den Einschluss von Studien in eine Metaanalyse (Conn & Rantz, 2003) und nicht zuletzt für Metaanalysen selbst (Brugha et al., 2012). Erst in jüngster Zeit erarbeiteten Higgins et al. (2013) einen umfassenden Kriterienkatalog für die Güte von Metaanalysen. Kurzum, der (gewichtete) Durchschnitt aus zum Teil fragwürdigen Einzelstudien erbringt, auch unter Einbezug von Moderatoren, nicht zwangsläufig die pure Erkenntnis. Konträr zu Stroebes Position kamen Brugha et al. (2012) nach ihrer Analyse von 106 Metaanalysen und Reviews zu folgender Einschätzung: „Sometimes a single high-quality, well-reported study can be recommended instead of a statistical synthesis of heterogenous studies (p. 450)."

Fahrenberg stellt im RP Selektivität (experimentelle Studien, Kognitionspsychologie, Priming, computer-gestützte Versuche) fest, was zweifellos die Generalisierbarkeit der Befunde des RP einschränkt – aber in welche Richtung? Es handelt sich im RP überwiegend um jenen Typus von Studien, bei dem man eine höhere Replizierbarkeit erwarten könnte als bei den gemäß Fahrenberg unterrepräsentierten Korrelationsstudien, multipel bedingten Veränderungsmessungen und Kriterienvorhersagen.

Vereinfacht würde ich annehmen, je basaler und universeller der psychologische Prozess und je stringenter und theoriegeleiteter das experimentelle Paradigma, desto eher kann man Replizierbarkeit erwarten - wenn Durchführungsqualität gewährleistet ist. Im Hinblick auf komplexeres Geschehen merken Deutsch wie zuvor Jüttemann zutreffend an, dass Ontogenese wie auch Zeitenwandel psychologische Prozesse modifizieren und damit Replizierbarkeit massiv einschränken. Die Relation von physischen Stimuli und deren psychologischer Verarbeitung ist variabel und zwar besonders dann, wenn es um komplexe, systemische, in soziale Kontexte eingebettete Phänomene geht. Dann muss man mit kultureller, epochaler und entwicklungsbedingter Variation rechnen, die sich in unterschiedlichen Zusammenhangsmustern psychologischer Größen ausdrückt, den Zusammenhang zwischen experimenteller Intervention und Outcome eingeschlossen. Reflexe sind leichter replizierbar als meinetwegen die Bewältigungsprozesse von Verlusterlebnissen. Letztere variieren epochal, kulturell, interindividuell und intraindividuell.

Die intraindividuelle Variation wie die Frage nach der Konstruktvalidität von Operationalisierungen über die Ontogenese gehören zum Kerngeschäft der Entwicklungspsychologie. Messen der Strange Situation Test im Kleinkindalter und das Adult Attachment Interview die gleiche „Bindung"? Bleiben Zusammenhangsmuster invariant über die Entwicklung, gibt es

stationäre Prozesse, ist mangelnde Rangstabilität Ausdruck schlechter Messung oder Beleg für interindividuelle Unterschiede in intraindividueller Veränderung? Erfreulicherweise geht man solche Fragen zunehmend (wenn auch schleppend) „bottom-up“ an. Man untersucht Individuen und stellt dann Gemeinsamkeiten und Unterschiede fest, die zu einer Eingrenzung der Generalisierungsbasis verhelfen. Ein anschauliches Beispiel liefern Molenaar und Lo (2012). Sie analysierten intraindividuelle Zeitreihen von 16 Jungen, die jeweils nach Interaktionen mit ihren Vätern Fragebögen ausfüllten. Eine P-Technik Faktorenanalyse der Fragebogenitems ergab eine große Heterogenität in der Faktorenstruktur dieser individuellen Zeitreihen. Nur zwei der 16 Jungen wiesen eine Dreifaktorenstruktur auf, die mit den Etiketten „Involvement“, „Anger“ und „Anxiety“ umschrieben wurde. Das nichtstationäre dynamische Funktionieren wird an einer Beispielperson demonstriert. Der Lag-1-Effekt von „Anxiety“ auf „Involvement“ änderte sich bei ihr über die 80 Messzeitpunkte von -.19 zu T1 auf .12 zu T80. Zu Beginn hatte demnach Angst eine hemmende Wirkung auf die Involviertheit, am Ende eine steigernde. Genau das ist Lernen dynamischer Systeme und widerspricht der Stationaritätsannahme bei Entwicklungsprozessen innerhalb der Person. Zugleich wiesen nur zwei der 16 Personen gleichartig konfigurierte Konstrukte auf, was der Homogenitätsannahme menschlichen Funktionierens zwischen Personen widerspricht. Für sich genommen wie gemeinsam verletzen beide Sachverhalte die Ergodizitätsannahme (Molenaar, 2004), eine entscheidende Voraussetzung für die Replikation von Befunden unabhängig von der Zusammensetzung der Stichprobe und der unterschiedlichen Lern- oder Entwicklungshistorie ihrer einzelnen Elemente.

Das Fehlen von Ergodizität ist eine maßgebliche Ursache der Nichtreproduzierbarkeit. Die zumeist unausgesprochene Erwartung, man untersuche in Stichproben für alle Individuen gültige, statische, zeitinvariante, nomothetische Gesetzmäßigkeiten, ist, sofern es nicht um sehr basale Prozesse geht, überwiegend unzutreffend. Jaccard und Dittus (1990) lieferten mit ihrer konzeptuellen Trennung in den datengenerierenden Prozess und die analysierten Daten Beispiele dafür, dass die gemeinsame Analyse von Daten, denen unterschiedliche generierende Prozesse zugrunde liegen (was sehr wahrscheinlich ist, wir den Daten aber nicht ansehen), völlig irreführende Ergebnisse produziert. Folglich hängt Replizierbarkeit maßgeblich von der Stichprobenzusammensetzung im Hinblick auf unterschiedliche psychologische „Funktionstypen“ und ihre jeweiligen Entwicklungshistorien ab. Diese inter- wie intraindividuelle Heterogenität bleibt jedoch unentdeckt, solange man ganz selbstverständlich von nomothetischen Gesetzmäßigkeiten ausgeht. In diesem Sinne stellte Brandtstädter (1985) für die Entwicklungspsychologie fest:

„Considering the notoriously limited ‘generalizability’ of psychological findings, the realization of a nomological program in psychology seems to face great difficulties. Actually, alleged law formulations in psychology usually refer to contingent regularities, which hold only under specific and by no means invariant social, cultural, and historical context conditions. ... Premature nomological interpretations tend to block further inquiry into the formative and sustaining conditions of quasi-lawful regularities. ... If human development is a plastic, culturally regulated process, it might seem utopian to expect more from developmental psychology than just local generalizations of the quasi-law type (p. 246).”

Local Generalizations nähert man sich zielführender mit einer bottom-up-Strategie (Personenansatz) als top-down vermittels aggregierter Daten (Variablenansatz), vor allem dann, wenn eine bislang unausgereifte Theorie keine Aussagen zu den Bedingungen unterschiedlichen psychologischen Funktionierens macht. So richtig und wichtig die genannten Einschränkungen und Fehlannahmen für den Misserfolg direkter Replikationen sind, sie sind nicht dazu geeignet, das halbleere Glas in ein halbvolles umzudeuten. Zum einen irritiert, dass die zitierten Einsichten bereits vor Jahrzehnten publiziert wurden, ohne nennenswerte Auswirkungen auf die Forschungspraxis zu haben. Dazu gehört auch (siehe Werbiks Beitrag) der hartnäckige Glaube, die Psychologie gewinne einzig durch komplexe mathematisch-statistische Auswertungsverfahren, die oftmals auf recht krude Daten angewendet werden, an Wissenschaftlichkeit und Erkenntnis. Zum anderen erklären sie nicht das gesamte „1 - .36 Residuum“ der Replizierbarkeit. In diesen Rest gehen ohne Frage Faktoren wie mangelnde Power, Stichprobenbesonderheiten (Psychologiestudenten) u. ä. ein. Hinzu kommt aber eine zuweilen etwas saloppe Handwerklichkeit, die maßgeblich dem systemimmanenten Webfehler der rein quantitativen Produktivitäts- bzw. Erfolgskriterien geschuldet ist. Zur (hoffentlich) unterhaltsamen Illustration dieses Gedankens hänge ich die Antworten an, die sich bei mir auf Anfragen, Nachfragen und Vorschläge angesammelt haben:

„I generally save all documentation relevant to our analyses. Much to my consternation, in this case, we appear to have discarded the computer runs which formed the basis of the LISREL analyses reported in our article. Thus, the likelihood statistics are not readily available. The lack of these statistics in the article is a serious omission. Because of other pressing commitments, we cannot at this time rerun the models we published.” *(Nachfrage zu Auswertungsdetails in einem Artikel)*

“Again, our disciplines have different biases, and mine tends to pay a lot less attention to the structure of the data and more to the representativeness of the sample ... I realize that to some extent, I am saying I’m ok with messy data and we’ll find what we find. I’, having a tough time seeing the alternative speaking to my field.” *(Operationalisierungsvorschlag für einen Antrag)*

“Da ist mir wohl ein Fehler bei der Interpretation unterlaufen – die Zahlen stimmen aber! Aber jetzt ist es schon gedruckt ... „ *(Nachfrage zur Inkonsistenz zwischen Parametern und Text in einem Artikel)*.

„Sie erleben hier Wissenschaft live – das Projekt entwickelt sich weiter, Items werden optimiert, und man muss abwägen,

was der nächste sinnvollste Schritt nach vorn ist ;) ... zu den df: ich vermute, das liegt daran dass wir beim ... die residual variance eines ... items auf 0 setzen mussten“ *(Nachfrage zu Inkonsistenzen in den Operationalisierungen und Modellfreiheitsgraden innerhalb einer Publikation)*

„Thank you for your interest toward our work. I had little bit difficulties to find the LISREL inputs because during last year I had computer problems and, in that context, lost all my LISREL in-and outputs from the computer ...” *(Anfrage nach Output und ggf. Daten aus einem Artikel zu Lehrzwecken)*

“... vielen Dank für Ihre Mail und Ihr Interesse an unseren Papers. Zu den Items: Hier scheint tatsächlich ein editorischer Fehler vorzuliegen ...“ *(Nachfrage zu einem Artikel - der prominente Erstautor hat erst gar nicht geantwortet)*

„... unsere Daten sagen das so, aber natürlich wird man die Befundlage im Blick behalten und sicherlich auch, wenn sich das mal anbietet, die Auswertungen wiederholen“. *(Nachfrage zu Inkonsistenzen zwischen Forschungsbefunden und offiziellen Statistiken in einem Artikel)*.

„I passed your request to my colleague who knows about this, will try again, but may take weeks to retrieve, thank you for your interest” *(Anfrage nach verschwundenen Artikeln eines online Journals, bei denen der Angefragte Herausgeber und Koautor war – danach kam nichts mehr)*

„Actually, I tried to retrieve the articles. The stupid thing is that my backup drive with my old research projects was stolen last January and the copies are in a different country. I have asked my co-authors but nobody seems to have the papers (which is really strange, it WAS an online journal!)” *(Anfrage in gleicher Sache an die Autorin einer Metaanalyse, in der die verschwundenen Artikel verarbeitet wurden)*

„Thank you very much for your detailed inquiry. Let me look into the matter to see if I can shed any light on these questions arising from analyses that we performed more than a decade ago ... May I ask about the over-arching direction of your reading and discussions?” *(Nachfrage von Studierenden des Seminars “Studying the Studies” (2014) zu Inkonsistenzen zwischen Publikationen derselben Autoren; erste Antwort)*

„Thank you very much again for your interest in our project. I regret to report that too much time has passed for the surviving authors to be able to reconstruct the answer to the question you posed.” *(zweite Antwort in gleicher Sache; die Publikationen erschienen 2001 und 2002)*

Literatur

Brandtstädter, J. (1985). Individual development in social action contexts: Problems of explanation. In J. Nesselroade & A. von Eye (Eds.), *Individual development and social change* (pp. 243-264). Orlando, FL: Academic Press

Brugha, T. S., Matthews, R., Morgan, Z., Hill, T., Alonso, J. & Jones, D. R. (2012). Methodology and reporting of systematic reviews and meta-analyses of observational studies in psychiatric epidemiology: systematic review. *The British Journal of Psychiatry*, *200* (6), 446-453.

Conn, V. S. & Rantz, M. J. (2003). Research methods: Managing primary study quality in meta‐analyses. *Research in Nursing & Health*, *26*(4), 322-333.

Higgins, J. P. T., Lane, P. W., Anagnostelis, B., Anzures-Cabrera, J., Baker, N. F., Cappelleri, J. C. et al. (2013). A tool to assess the quality of a meta-analysis. *Research Synthesis Methods*, *4*, 351-366.

Jaccard, J. & Dittus, P. (1990). Ideographic and nomothetic perspectives on research methods and data analysis. In C. Hendrick & M. S. Clark (Eds.), *Research methods in personality and social psychology* (pp. 312-351). Newbury Park, CA: Sage.

Maier, W. & Möller, H. (2010). Meta-analyses: A method to maximise the evidence from clinical studies? *European Archives of Psychiatry and Clinical Neuroscience*, *260*(1), 17-23.

Molenaar, P. C. M. (2004). A manifesto on psychology as ideographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*, 201-218.

Molenaar, P. C. M. & Lo, L. (2012) Dynamic factor analysis and control of developmental processes. In B. Laursen, T. D. Little & N. A. Card (Eds.), *Handbook of developmental research methods* (p. 333-349). New York: The Guilford Press.

18.09.2015

Für einen Dualismus der Methoden - Konsequenzen aus der Replikationskrise

Prof. (em.) Dr. Hans Werbik

Die mangelnde Reproduzierbarkeit psychologischer Experimente hängt mit einer Überdehnung der naturwissenschaftlichen Methodologie zusammen. In Abhängigkeit von der Komplexität des zu untersuchenden Phänomens sollte zwischen naturwissenschaftlichen Methoden und geisteswissenschaftlichen Methoden unterschieden werden, ähnlich wie Wilhelm Wundts

Unterscheidung zwischen Physiologischer Psychologie und Völkerpsychologie.

Ergebnisse geisteswissenschaftlicher Methoden sind grundsätzlich nicht reproduzierbar, weil

das psychische System mit der Zeit nicht konstant bleibt. Beispielsweise ist eine autobiographische Erzählung ein Phänomenkomplex, der nicht wiederholbar ist; reproduzierbar sind lediglich die harten Daten des Lebenslaufs. Im Bereich der naturwissenschaftlichen Psychologie werden häufig probabilistische Modelle aufgestellt, deren Anwendungsbereich überdehnt wird. Denn der Begriff der mathematischen Wahrscheinlichkeit beruht auf dem Begriff des Zufallsexperiments.

Ein psychologisches Merkmal eine Zufallsvariable zu nennen setzt voraus, daß das Merkmal so betrachtet werden kann, als ob es ein Ergebnis eines Zufallsgenerators wäre. Diese Voraussetzung ist beispielsweise erfüllt, wenn man einen einfachen Durchstreichtest veranstaltet. Bei den meisten psychologischen Merkmalen ist diese Voraussetzung jedoch nicht erfüllt. Während in der Physik alle Aussagen, die nicht deterministisch sind, probabilistisch sind, gibt es in der Psychologie besonders häufig Zusammenhänge, die weder deterministisch noch probabilistisch sind: Es sind dies die bedeutungshaltigen und sinnhaften Verhaltensweisen. Bedeutung und Sinn kann nur mit geisteswissenschaftlichen Methoden erfaßt werden.

What have we learned from the Reproducibility Project?

Prof. Dr. Wolfgang Stroebe

Did a huge research replication exercise do a power of good?

When 270 researchers spend several years replicating 100 psychology experiments, one expects momentous insights. That only 36 per cent of results could be replicated, and that social psychological research was less reproducible than research in cognitive psychology is, on the face of it, shocking ("Majority of psychology papers are not reproducible, study discovers", News, 3 September). But are the findings of the Reproducibility Project: Psychology really that unexpected, and do they mean that we can no longer believe psychology textbooks?

Although these results have made headlines, they should not have been a surprise to research psychologists. In 1962, Jacob Cohen reported in the *Journal of Abnormal and Social Psychology* that the average statistical power of research in these fields was only 0.48. Several subsequent reviews indicated that the power of this type of research has not increased.

The power of a study, which determines whether we can identify valid and reject invalid hypotheses, can be compared to the magnification of a microscope: with too little magnification, we may not see things that are there or believe we see things that are not there. When conducting scientific studies, researchers look for "significant" results – in technical terms, for a "p-value" of less than 0.05. If one tried to replicate a study that had a p-value of less than 0.05 and power of 0.50, one would have only a 50 per cent chance of success. So why do psychologists not aim for a maximal power of 1? Power is determined by the study's sample size, the significance level and the strength of the effect (relationship) studied. In some cases that effect might be large statistically, as when we study the impact of study hours on grades; in other cases the effect may be weaker, as between class size and grades. Since in novel research one often does not know the effect strength, researchers often underestimate the large sample size they need to achieve acceptable levels of power.

The power of the original studies is not reported in the Reproducibility Project's new *Science* article. However, it notes that, for technical reasons, studies in cognitive psychology often have more power than social psychological studies. (In the latter field especially, researchers have to ensure that hypotheses are not obvious.) This would be one explanation why social psychology fared less well. Another is that whereas the variables investigated by cognitive psychologists are relatively unaffected by cultural norms and other social factors, variables studied by social psychologists are typically influenced by historical change and local context. The arguments used by US researchers to persuade US students in the 1980s would be unlikely to persuade European students, or US students tested 30 years later. Persuasive arguments change with time and context. So another factor in the lower percentage of successful replications for social versus cognitive psychological studies

was probably that “exact replications” may often have failed to capture the same theoretical variables manipulated in the original study.

Reactions to the Reproducibility Project will remind social psychologists of the replicability crisis of the 1970s. This was triggered by complaints that social psychological knowledge was not cumulative: for every study demonstrating some significant effect, there were replications that were not significant. Thus a reviewer who added up significant and non-significant effects in tests of some theory (known as the “box score method”) often found that non-replications outweighed successful replications. This led to the development of powerful new meta-analytic methods, which statistically combined the results of different experiments.

Even when many individual studies fail to yield significant results, meta-analysis may reveal a significant overall result. This prevents us from concluding that a finding is not reliable when in fact it is. Because non-significant findings typically remain unpublished, meta-analyses could be subject to publication bias, if they relied exclusively on published findings. To avoid this, meta-analysts attempt to trace and to include relevant unpublished studies. Furthermore, sophisticated methods have been developed to identify publication bias and even to correct it.

Although the Open Science Collaboration did not use simple box scores, the statement that only 36 per cent of the findings could be replicated is based on the same logic. But the Open Science Collaboration also conducted a meta-analysis, combining the effect sizes of the original studies with those of the replications to yield an overall effect size. Not surprisingly, the number of studies that had significant effects increased to 68 per cent. In other words, two-thirds of the results could be replicated when evaluated with a simple meta-analysis that was based on both original and replication studies.

As meta-analyses published in psychology journals typically combine the results of hundreds of studies, it is hardly surprising that they give a much more positive picture of the replicability of psychological research than the Reproducibility Project does. The conclusions of textbooks should be based not on single studies but on multiple replications and large-scale meta-analyses, so the results of the Open Science Collaboration will not undermine our faith in good psychology textbooks.

Reporting the percentage of successful replications is not very informative. More usefully, the project could have identified aspects of studies that predicted replication failure. But here the report disappoints. Since meta-analysis permits us to evaluate the validity of research without the need to collect new data, one can question whether the meagre results of this project justify the time investment of 270 researchers and thousands of undergraduate research participants.

Wolfgang Stroebe, Miles Hewstone

Replikationsforschung: Weiter, womöglich sogar noch besser?

Prof. Dr. Roland Deutsch

Liebe Kolleginnen und Kollegen,

die seit einigen Jahren laufende Debatte über die Zuverlässigkeit psychologischer Erkenntnisse, zweifelhafte Forschungspraktiken und offene Wissenschaft ist eine notwendige, zeitgemäße und in der Summe äußerst förderliche Entwicklung, welche einen kulturellen Wandel in unserem Feld anzeigt. Kulturelle Veränderungen gehen immer mit Dissens und zeitweiligen Konflikten einher. Wir stecken mitten in solchen Umbruchswirren. Es ist klar, dass das Ergebnis des Reproducibility Projects (RP) und andere Befunde zur Zuverlässigkeit unserer Forschungspraxis keinen Psychologen mit Stolz erfüllen. Aber die Tatsache, dass wir diese Befunde generiert haben und uns ihnen in einer großen Anzahl von Symposia, Special Issues und Diskussionsforen stellen, dass wir in zentralen Publikationsorganen unserer Disziplin nun Raum für Replikationsstudien bieten, dass wir nun über Zeitschriften mit ausschließlicher Publikation vorregistrierter Studien verfügen (z.B., Comprehensive Results in Social Psychology), all dies weist auf sehr begrüßenswerte Entwicklungssprünge hin. Wir sollten alles dafür tun, diese positiven Impulse zu stärken und dadurch die Psychologie als Wissenschaft voranzubringen. Ich selbst habe Replikationsprojekte aus voller Überzeugung unterstützt bzw. durchgeführt. Dazu gehört aber auch eine besonnene Vorgehensweise und Kommunikation. Sowohl Beschwichtigung als auch Dramatisierung werden eher zu einer Polarisierung der Einstellungen führen. Um die Ziele zu erreichen, die hinter den Initiativen stehen, ist aber ein gewisser

Konsens der Scientific Community nötig.

Aber, so mag man sich fragen, sind die Befunde des RP denn nicht ohne Zweifel dramatisch? Ja, die Befunde des RP weichen stark von dem ab, was sich vermutlich alle für unsere Forschungsergebnisse wünschen. Und dennoch gibt es einige (hier in vielen Forumsbeiträgen auch schon thematisierte) Schwierigkeiten im Zusammenhang mit Replikationsstudien (z.B. Repräsentativität von Stichproben, Regressionseffekte). Vor dem Hintergrund dieser Schwierigkeiten müssen wir uns die Frage stellen: Welches Ausmaß an Reproduzierbarkeit können wir eigentlich erwarten? Die Bewertung fehlgeschlagener Replikationsversuche hängt auch von der Antwort auf diese Frage ab. Ich möchte eine Schwierigkeit, die mir persönlich besonders wichtig ist, ein wenig näher diskutieren. Sie betrifft die Frage der sogenannten *direkten* Replikation.

Bei direkten Replikationen (auch Materialreplikationen) versucht man, das physische Versuchsmaterial der Originalstudie möglichst identisch zu verwenden. Psychologische Gesetze werden aber – insbesondere in kognitionspsychologischen Ansätzen – auf der Ebene theoretischer Konstrukte formuliert, die sich auf psychische Gegenstände beziehen (eine bedeutsame Ausnahme stellt das behavioristische Paradigma dar). Physische Gegenstände sind vor allem deshalb relevant, weil diese bestimmte innere Zustände auslösen. Solange Untersuchungsmaterialien nicht diejenigen inneren Zustände auslösen, die in den psychologischen Gesetzen beschrieben sind, ist keine Replikation des Effekts zu erwarten, selbst wenn das postulierte psychologische Gesetz zutrifft.

Der Zusammenhang zwischen physischen Gegenständen und inneren Zuständen ist je nach Forschungsbereich mehr oder weniger eindeutig und dauerhaft. So lässt sich zum Beispiel argumentieren, dass eine unveränderliche Funktion zwischen der physikalisch bestimmten Helligkeit einer Lichtquelle und der subjektiven Helligkeitswahrnehmung besteht. Es lässt sich aber beispielsweise viel schwerer argumentieren, dass bestimmte Körpermerkmale, die heutzutage mit hohem sozialem Status korrelieren, dies auch in 50 Jahren noch tun werden. Würde man mit den Materialien von heute in 50 Jahren erneut ein psychologisches Gesetz testen, in dem sozialer Status eine Rolle spielt, käme es womöglich zu einer Nicht-Replikation, obwohl das Gesetz gilt. Eine ausführliche Diskussion und empirische Demonstration der kulturell bedingten Veränderung der Validität von Versuchsmaterial findet sich unter: psych.stanford.edu/~michael/papers/Ramscar-Shaoul-Baayen_replication.pdf

Im Fall von Wissenschaften, die physische Untersuchungsgegenstände haben, liefern direkte Replikationen klare Ergebnisse. Die Relation zwischen Konstrukt einerseits und Messung bzw. Manipulation andererseits ist in der Regel eindeutig und kann grundsätzlich als stabil angesehen werden. Dies ist im Fall psychologischer Experimente häufig anders. Zu einem gewissen Teil mag dies darauf zurückzuführen sein, dass unsere Methoden noch nicht hinreichend entwickelt sind. Wir sollten deshalb weiter nachdrücklich danach streben, die Reliabilität und Validität unserer Messungen und Manipulationen zu verbessern. Es ist aber auch plausibel, dass eine geringere Eindeutigkeit und Stabilität zu einem gewissen Teil aus Eigenheiten unseres Untersuchungsgegenstandes resultiert. Eine Wandlung der Beziehung zwischen inneren Zuständen und physikalischen Reizen über die Zeit hinweg oder zwischen unterschiedlichen Lernumgebungen ist regelmäßig zu erwarten. Eine solche Instabilität der Beziehung zwischen inneren und äußeren Zuständen ist kein Methodenfehler. Sie folgt daraus, dass Menschen, vielleicht mehr als jedes andere Lebewesen, über eine hochgradig lernabhängige Flexibilität der Verhaltenssteuerung verfügen.

Vor diesem Hintergrund ist das Konzept der direkten Replikation zwar interessant, die daraus abzuleitenden Schlussfolgerungen über die Gültigkeit des psychologischen Gesetzes sind jedoch schwächer als bei anderen Replikationsformen. Die Wiederholbarkeit eines experimentellen Befundes ist nur dann zu erwarten, wenn die Bedingungen, die im zugrundeliegenden psychologischen Gesetz formuliert sind, im Replikationsexperiment auch tatsächlich gegeben waren. Dies kann man im Fall psychologischer Forschung in den meisten Fällen nicht a priori feststellen. Jedenfalls ist dies nicht durch die Verwendung desselben Materials sichergestellt.

Wie dramatisch sind vor diesem Hintergrund 36% fehlgeschlagene Replikationen bzw. 68% Beständigkeit in einer metaanalytischen Betrachtungsweise? Einerseits muss man sagen: Es ist ein besorgniserregender

Befund, denn man konnte nur in weniger als der Hälfte der Versuche ein publiziertes Ergebnis direkt wiederholen. Andererseits zeigt unsere Diskussion, dass womöglich ein Teil der beobachteten Unzuverlässigkeit auf Aspekte zurückzuführen ist, die in zukünftigen Replikationsarbeiten verbessert werden könnten. Einen dieser Aspekte habe ich im Detail beschrieben. Der im RP gewählte Ansatz der direkten Replikation führt dazu, dass zumindest manche der nicht gelungenen Replikationen darauf zurückzuführen sein könnten, dass die Operationalisierung der Konstrukte im gegebenen Kontext nicht gelungen ist. Die Methode der direkten Replikation ist nicht ausreichend, um zwischen grundsätzlich inadäquaten Operationalisierungen, nur im Kontext der Replikationsstudie inadäquaten Operationalisierungen und ungültigen Gesetzen als Ursache für Fehlschläge zu unterscheiden.

Daraus lassen sich zwei Implikationen für zukünftige Forschungsarbeiten ableiten. Erstens sollten Forschungsarbeiten die Generalisierbarkeit explizit reflektieren und diskutieren – die Generalisierbarkeit der vermuteten psychologischen Gesetze ebenso wie die Generalisierbarkeit der Operationalisierungen. Für beides können wir in der Psychologie nicht von axiomatisch universeller Gültigkeit ausgehen. Zweitens kann der Informationsgehalt von Replikationsstudien dadurch gesteigert werden, dass die Methode der direkten Replikation zugunsten einer geprüft konstruktvaliden Replikation aufgegeben wird. Das heißt, es muss sichergestellt werden, dass die Bedingungen, die in den psychologischen Gesetzen spezifiziert sind, auch wirklich gegeben waren. Dies kann erstens durch den konsistenten Einsatz von Manipulation Checks geschehen. Zweitens kann begleitend zu einer Materialreplikation eigenes, an der lokalen Grundgesamtheit validiertes Material verwendet werden.

Zusammengenommen möchte ich noch einmal bekräftigen, dass ich den einsetzenden kulturellen Wandel in der Psychologie als eine große Errungenschaft betrachte. Es ist eine wichtige Aufgabe, den Wandel weiter zu fördern und die Zuverlässigkeit psychologischer Forschungsergebnisse weiter zu stärken. Dabei spielen Replikationen eine große Rolle, dies wird durch das RP nochmals verdeutlicht. Publizierte und groß angelegte Replikationsversuche sind allerdings ein recht junger Teil unserer Forschungskultur, der in seiner Durchführung und Rezeption noch weiter optimiert werden sollte.

Roland Deutsch

14.09.2015

Der Befund des Reproducibility Projects kann nicht verallgemeinert werden, eine Theorie-bezogene kooperative Untersuchungsstrategie wäre überlegen

Prof. (em.) Dr. Jochen Fahrenberg

Liebe Kolleginnen und Kollegen

Das **Reproducibility Project** von Brian Nosek et al. (2015) hat offensichtlich ein hohes Anregungspotenzial, sich mit dessen Ergebnissen auseinanderzusetzen und über Folgen für das Bild der wissenschaftlichen Psychologie in der Öffentlichkeit nachzudenken. Bereits in der Planungsphase des **RP** wurde das Dilemma gesehen. Es gab pessimistische Erwartungen und ausdrückliche Warnungen, dass die Befunde voraussichtlich das Ansehen der Psychologie beeinträchtigen würden. Demgegenüber stand die Überzeugung, dass Wissenschaft auch systematische Selbstkritik verlangt. Gerade die methodisch kompetenten Psychologen könnten vorangehen und damit auch ein Vorbild für Nachbardisziplinen geben (siehe Siri Carpenter 2012 über diese „kühne Initiative“ ebenfalls in *Scienc*e; Literaturangaben in meinem Wikipedia-Artikel über Reproduzierbarkeit).

Hier ist auch ein Seitenblick auf die Auseinandersetzungen über die evidenzbasierte Medizin angebracht. Wenn es nicht allein um Grundlagenforschung, sondern auch um die wissenschaftlich begründete Angewandte Psychologie geht, haben die Bemühungen um Qualitätskontrolle eine zusätzliche berufsethische Perspektive. Diese Blickrichtung wird auch für die öffentliche Reaktion wichtig sein. Die abschliessenden Sätze des **RP**-Aufsatzes sind bemerkenswert: „We conducted this project because we care deeply about the health of our discipline and believe in its promise for accumulating knowledge about human behavior that can advance the quality of the human condition. Reproducibility is central to that aim. Accumulating evidence is the scientific community’s method of self-correction and is the best available option for achieving

that ultimate goal: truth."

In den bisherigen Diskussionsbeiträgen standen hier zunächst spezielle Fragen der statistischen Analyse im Vordergrund, statt eigene Erfahrungen und Schlussfolgerungen aus einem Replikationsversuch zu berichten oder grundsätzlicher zu fragen, d.h. Prinzipien der Evaluation und alternative Forschungsstrategien zum Reproduzierbarkeitsproblem zu diskutieren.

Repräsentativität und Generalisierbarkeit

Das **RP** basiert auf 100 Publikationen in drei amerikanischen Journals im Jahr 2008. Im Abstrakt ist von experimentellen *und* korrelativen Studien die Rede, doch wird später nicht nach Designtypen unterschieden. Die Auswahl aus den insgesamt 488 Artikeln wird als „quasi-random" bezeichnet, denn sie erfolgte nach einer Anzahl von a priori gesetzten Eignungskriterien und in einem stufenweisen Verfahren, welche Themen den potenziellen Projektmitarbeitern für den Replikationsversuch nach und nach angeboten wurden, wobei auch eine Kooperation mit den primären Autoren angestrebt war. Aufgenommen wurden jene 100 von 113 Replikationsversuchen, die rechtzeitig für den Ergebnisbericht fertiggestellt waren. Die möglichen Folgen dieser heterogenen Selektionsschritte sind nicht einzuschätzen.

Berücksichtigt wurden Studiencharakteristiken und Moderatoren, d.h. Einstufungen und formale Merkmale der primären Untersuchung und der Replikation, beispielsweise die pauschale „Data collection quality", Einstufungen der Erfahrung und Expertise der Untersucher, Prestige der Institution und Citation Index der Autoren (vgl. die Informationen im Begleitmaterial). Dagegen wurde offenbar nicht systematisch evaluiert, in wieweit die primären oder die sekundären Untersucher wichtige Kovariablen und typische psychologische Moderatoreffekte ihrer jeweiligen Experimente zu erfassen und auch zu kontrollieren versuchten.

Ist überhaupt eine *Population* psychologischer Experimente und Forschungsarbeiten zu definieren, so dass eine *Zufallsstichprobe* gezogen und das Ergebnis generalisiert werden kann? – Die Schlussfolgerungen gelten offensichtlich nur für diese eigentümliche Selektion, weder für „die experimentelle" Psychologie noch für „die" Psychologie überhaupt. Eine Aggregation der Ergebnisse aus der ersten und der zweiten Untersuchung war nicht explizit geplant und ist nachträglich nur als explorativer Beitrag anzusehen. Die Bilanz des Projektes mit den 36 bzw. 39 Prozent als erfolgreich bewerteten Replikationsversuchen kann folglich in keiner interessanten Hinsicht verallgemeinert werden, sondern bedeutet eine nicht unerwartete, aber herausragende Warnung und eine gewichtige Herausforderung.

Das RP folgte einer relativ einfachen, aber organisatorisch aufwendigen Strategie

Die Auswahl der Studien war schematisch. Es fehlt das gemeinsame Bezugssystem einer bestimmten psychologischen Theorie, damit die kooperative Überprüfung zentraler Deduktionen strukturiert, methodologisch geordnet und genau standardisiert werden kann. Wer sich die Mühe macht, die einzelnen Studien anzusehen, wird in vielen Fällen nicht gerade beindruckt sein von der theoretischen Ableitung des experimentellen Ansatzes, dem Niveau der Operationalisierung und dem vermutlichen Stellenwert für eine vertiefende theoretische Auseinandersetzung. Es dominieren gegenwärtig aktuelle Themen der „kognitiven Psychologie", u.a. elementare Untersuchungen zum Priming und zu den Auswirkungen spezieller Instruktion auf Einstellungen, häufig auf der Basis einfacher computer-gestützter Versuche. Die Themen sind also auch inhaltlich und methodisch hochselektiv. Anspruchsvollere Untersuchungen hinsichtlich Forschungsaufwand, Methoden, Apparatur und Teilnehmern, d.h. nicht nur Studierende der Psychologie, sind in der Minderzahl. Außerdem besteht der größere Anteil psychologischer Forschung wahrscheinlich nicht aus strikten Laborexperimenten, sondern aus quasi-experimentellen Studien, Korrelationsstudien, multipel bedingten Veränderungsmessungen, Kriterienvorhersagen und anderen Designs.

Bei genauerer Evaluation wird auch nach den Datenquellen und Aggregationsstufen zu unterscheiden sein. Welchen Anteil haben Selbstberichte, Selbstbeurteilungen (Fragebogendaten) gegenüber beobachteten oder gemessenen Verhaltensmerkmalen und physiologischen Messungen? Welche Mängel der internen Validität sind zu erkennen? Aus der Methodologie der Evaluationsforschung sind hierzu viele einschlägige Prinzipien und Kriterien zu entnehmen.

Zur Evaluation dieses herausragenden **RP** gehören Überlegungen, ob andere Strategien eventuell prägnantere Schlussfolgerungen ermöglichen. Zum Thema Reproduzierbarkeit (Replikation) existieren bereits seit Jahrzehnten einzelne Vorbilder und zahlreiche konzeptuelle Beiträge und Übersichten, übrigens auch beachtliche Beiträge deutscher Autoren (Schweizer, 1989; Schmidt, 2009).

Andere Untersuchungsstrategien

Die **direkte (genaue) Replikation durch denselben Untersucher** könnte der naheliegende erste Schritt sein, quasi als eine Selbstverpflichtung im naturwissenschaftlichen Arbeitsstil. Diese Strategie zum Standard zu machen, entspricht jedoch nicht dem oft kurzatmigen Forschungs- und Publikationsstil der heutigen Zeit. Das Verständnis vieler Kollegen und Gutachter müsste erst gewonnen werden.

Auf allen Gebieten psychologischer Forschung gibt es mehr oder weniger aufwendige Projekte. Wenn nun eine

Untersuchung, beispielsweise in der Psychophysiologie, mehrere spezialisierte Mitarbeiter und kostspielige Gerätetechnik erfordert, dann kann etwa eine Projektdauer von sechs Monaten – addiert – leicht mehrere Arbeitsjahre und beträchtliche Sachmittel erfordern. Wären diese Jahre und Forschungsgelder nicht besser in ein Anschlussprojekt oder ein anderes Forschungsthema statt in eine Replikation investiert? Wer für systematische Replikationen plädiert, muss sich mit dieser Frage auseinandersetzen. Wer einen missglückten teuren Replikationsversuch dieser Art selber erfahren hat, wird seine Erwartungen und Maßstäbe anpassen.

Die **möglichst ähnliche Replikation,** zumindest der theoriefundierenden Experimente durch *unabhängige* Untersucher ist eine generelle Forderung wissenschaftlicher Arbeit. Weiterhin kann zwischen der *näherungsweisen (approximativen)* Replikation, der *partiellen* Replikation und der *systematischen* Replikation mit geplanter Variation des Designs unterschieden werden. Bei einer relativ simplen Methodik (Fragebogen oder standardisierte Software computergestützter Experimente) sind Replikationsversuche leichter zu bewerkstelligen als bei einer anspruchsvollen Methodik, die eine spezielle Ausbildung verlangt. Generell ist zu fragen, ob wichtige Varianzquellen ausreichend kontrolliert oder standardisiert sind. Beispiele sind Erwartungseffekte und Vl-Vp-Interaktionen. Wie verschaffen sich die Untersucher eine hinreichende Evidenz, dass die durch ihre Instruktionen intendierte psychologische Bedingungsvariation (Einstellung, Emotion, Motivation) aktuell erfolgreich war, und nach welchen Kriterien werden Teilnehmer mit mangelnder Compliance ausgeschlossen?

Bei apparativ anspruchsvoller Methodik wird es Standards innerhalb eines Labors geben, aber nicht ohne weiteres zwischen den Laboratorien. Viele Publikationen enthalten nicht mehr hinreichend genaue Forschungsberichte mit präzisen Angaben über alle wichtigen Definitionsmerkmale der Aufgaben und des u.U. relevanten Kontextes, über die Messtechnik und über die genaue Parametrisierung der abhängigen Variablen, Ausreißer-Management usw. – erfahrungsgemäß kommt es vor, dass eine hinreichende Klärung selbst durch Laborbesuche nicht zu erreichen ist, da einzelne wichtige Parameter nicht gemessen oder protokolliert wurden oder nicht valide zu messen und zu standardisieren sind.

Die **konzeptuelle (konstruktive) Replikation** kann durch eine eigenständige Deduktion aus der zugrundeliegenden Theorie vorgenommen werden und ist eventuell mit einer eigenständigen, aber als adäquat postulierten Operationalisierung des theoretischen Konstrukts verbunden. Replikationen dieser Art sind häufiger zu finden, jedoch nicht unter dieser Bezeichnung, sondern als mehr oder minder freie Anlehnungen an vorausgegangene Untersuchungen. Diese Strategie ist wohl typisch für den Ablauf wissenschaftlicher Forschung, meist unkoordiniert, aber vielleicht kreativ zu neuen Einsichten führend – oder zur Konfusion. Wahrscheinlich kennt jeder Empiriker zu einer aktuellen Forschungsfrage seines Gebiets ein wichtiges Review, das „inconsistent results“ feststellt. Solche Knoten widersprüchlicher Ergebnisse oder komplexe Methodenprobleme werden in der fachlichen Auseinandersetzung diagnostiziert, eventuell zeitweise ausgeblendet und treten nach einiger Zeit erneut hervor.

Von der Programmatik der strukturalistischen Wissenschaftskonzeption wäre eigentlich zu erwarten, dass sich an die prägnante Formalisierung der theoretischen Annahmen auch eine ähnlich prägnante Verhandlung über adäquate Operationalisierungen und über kooperativ angelegte Replikationen anschließt. Wäre auf diese Weise zumindest in kleinen Bereichen eine konsistente Strategie für konstruktive Replikationen zu erreichen?

Eine **Generalisierbarkeitsstudie** nach der Konzeption von Cronbach et al. fordert die systematisch fortschreitende Verallgemeinerung eines als wichtig angesehenen Ergebnisses über verschiedene Personengruppen, über Termine (einfache Wiederholungen), Varianten der Methodik (Testformen) und über Versuchsbedingungen (Situationen). Die **Labor-Feld-Generalisierbarkeits-Studie** prüft die Übertragung von Untersuchungsergebnissen im Labor bzw. im Untersuchungszimmer auf das Feld bzw. auf Alltagsbedingungen. In der psychologischen Diagnostik und Assessmenttheorie ist der Begriff der *externen* Validität, in den Untersuchungen zum ambulanten Monitoring und Assessment auch der Begriff der *ökologischen* Validität üblich. Die von Kurt Lewin formulierte Grundfrage lautet: Handelt es sich um denselben Geschehenstyp, der im Labor und im Feld beobachtet wird? Die systematische **Prüfung der Kontextabhängigkeit** von Forschungsergebnissen ist ein bisher noch kaum entwickeltes Gebiet der schwierigen Methodologie unseres Faches.

Die Labor-Feld-Vergleichsstudien und die innovative Methodik des Ambulanten Monitoring und Assessment geben hier eine neue Erfahrungsbasis, die in diese Diskussion einbezogen werden sollte. – Der Eindruck, dass bisherige Replikationsstudien kaum einen Einfluss von Moderatorvariablen feststellten, stammt aus fragwürdigen Studien und mag zutreffen, wenn nur einzelne formale Charakteristika wie im **RP** gemeint sind. Die Variation eines einzigen Priming-Experiments in der „Many Labs“-Studie (siehe Diskussionsbeitrag von Frank Renkewitz) kann hier nicht ausreichen. Kontextabhängigkeit meint weitaus mehr.

Insgesamt ergeben sich kritische Fragen nach Adäquatheitsbedingungen, Operationalisierungen und Kontrollen von Moderatorvariablen, die gewiss noch weitaus schwieriger sind als die Details der statistischen Analyse und der Bewertung von Signifikanz und Effektgrößen. In wieweit das in der kritisch-rationalistischen Wissenschaftstheorie empfohlene Verhandlungsmodell bei kooperativen Replikationsprojekten trägt, müsste erprobt werden.

Die Befunde des **RP** sind nicht unerwartet und deshalb wäre es verfehlt, den Problemstand etwa als eine *neue* „Krise“ der Psychologie zu stilisieren. Die „kühne Initiative“ des Reproducibility Project von Nosek und Kollegen hat erneut die Reflexion von Grundsatzfragen angestoßen und motiviert zu kooperativen Programmen: jeweils auf eine bestimmte psychologische Theorie zentriert und methodologisch noch anspruchsvoller angelegt.

Jochen Fahrenberg

Anmerkung: Wikipedia-Artikel schreibe ich, weil ich die Idee dieser kooperativen Enzyklopädie hervorragend und zukunftsweisend finde. Geeignete Bearbeitungen des Artikels sind nötig und üblich. Wikipedia ist heute wahrscheinlich für viele Studierende der Psychologie zu einer zentralen Informationsquelle geworden.

09.09.2015

Nicht „Weiter so“, sondern „Weiter, aber besser“!

Prof. Dr. Mario Gollwitzer

Die bisherigen Beiträge in diesem Forum machen deutlich, wie wichtig es ist, die Diskussion um die Replizierbarkeit psychologischer Befunde jetzt zu führen. Mein Eindruck ist, dass die Diskussion in diesem Forum mit großer Ernsthaftigkeit, teilweise aber auch recht emotional geführt wird. Diese Diskussion muss weitergehen. Sie darf sich aber in diesem Forum nicht zu stark auf Fragen wie „Liegt die ‚korrekte‘ Replikationswahrscheinlichkeit eher bei 36% oder eher bei 68%?“ oder „Wer hat Recht: die Originalstudie oder die Replikation?“ konzentrieren. Bisweilen habe ich den Eindruck, es besteht die Gefahr einer solchen Konzentration auf methodische Detailfragen.

Ich möchte den Blick noch einmal etwas weiten und fragen: sind die Maßnahmen, die die psychologische scientific community in den vergangenen Jahren ergriffen hat (nur ein paar Beispiele: Einführung von „replication sections“ in psychologischen Fachzeitschriften; Einführung von Zeitschriften bzw. Sonderausgaben mit vorregistrierten Studien; grundlegende Veränderungen der editorialen Praxis bei der Einreichung und Begutachtung von Manuskripten u.v.a.), sinnvoll und wirksam? Muss noch mehr getan werden? Wie kann eine Fachgesellschaft wie die DGPs solche Maßnahmen fördern und damit zur Qualitätssicherung in der Forschung beitragen? Genau solche Fragen werden gerade im Vorstand intensiv diskutiert - hier brauchen wir auch die Rückmeldung aus der Mitgliederschaft. Ideen wie sie im Diskussionsbeitrag von Schönbrodt und Kollegen skizziert werden, sind daher hoch willkommen! (Weitere kritische Beiträge zu unserer Pressemitteilung sind natürlich auch willkommen...)

Ein Wort noch zur Pressemitteilung, auf die sich viele Kommentare beziehen: Es ist keineswegs so, wie in einigen Beiträgen vermutet, der Vorstand der DGPs wolle die Botschaft verbreiten „Es ist doch alles prima; wir können weitermachen wie bisher.“ Das Ziel der Pressemitteilung bestand vielmehr darin, den vielen Medienberichten über die Science-Studie, die ihrerseits über das vergangene Wochenende zu erschreckend destruktiven Leser(innen)kommentaren geführt hatten, eine konstruktive Botschaft hinzuzufügen: Replikationsstudien wie die der Open Science Collaboration sind eben nicht als Todesstoß für unsere Disziplin zu verstehen. Sie bergen vielmehr Chancen - beispielsweise die, unsere Publikationspraktiken, aber auch unseren Blick auf die Generalisierbarkeit unserer Befunde noch einmal kritisch zu überdenken. Diese Chancen sollten wir nutzen. An die Presse gewandt wollten wir die Botschaft aussenden: genau dies werden wir jetzt tun. Zu dieser Botschaft stehen wir als Fachgesellschaft. INNERHALB der Fachgesellschaft darf der Diskurs gerne (selbst)kritischer ablaufen als nach außen. Aber: Pressemitteilungen sind eben eine Form der Außenkommunikation.

Also: Was kann unsere Fachgesellschaft tun, um die Qualität der psychologischen Forschung - in Deutschland und darüber hinaus - zu sichern bzw. noch weiter zu steigern und damit die Replikationsrate psychologischer Effekte - wie auch immer sie operationalisiert und quantifiziert wird - zu erhöhen?

Mario Gollwitzer (u.a. Schriftführer der DGPs)

09.09.2015

Dr. Frank Renkewitz

Aufgrund der fehlerhaften Darstellung und Interpretation der Ergebnisse des Replikationsprojekts in der Pressemitteilung der DGPs und einiger inzwischen in diesem Forum geäußerter Kritikpunkte möchte ich die zentrale Aussage der Befunde unserer Studie hier nochmals erläutern.

Fehlinterpretation und Kritik scheinen mir zumindest teilweise auf einem Missverständnis über das Ziel des Replikationsprojekts zu beruhen. Das primäre Ziel lag nicht darin, den Anteil falsch positiver Befunde (und damit indirekt auch die Anteile falscher und korrekter Hypothesen) in psychologischen Top-Journals zu schätzen. Ein solches Unterfangen würde erfordern, dass man eine ausreichend große (und möglichst unverzerrte) Auswahl an Studien *mehrfach* repliziert – und hätte damit auch die Ressourcen einer sehr großen internationalen Kollaboration überfordert. Das Hauptziel des Replikationsprojekts war es hingegen, eine Schätzung der Replizierbarkeit von Befunden in psychologischen Top-Journals zu gewinnen. Replizierbarkeit ist eine fundamentale Anforderung an jede empirische Wissenschaft, die vermutlich in nahezu jedem einführenden Methodenlehrbuch auf den ersten Seiten erläutert und betont wird. Das zentrale Ergebnis unserer Studie ist nun, dass es um die Replizierbarkeit psychologischer Befunde schlecht und in Teilbereichen desaströs bestellt ist. Aus diesem Ergebnis mag man auch Aussagen oder zumindest besser fundierte Annahmen über den Anteil falsch positiver Befunde ableiten können. Zum Beispiel erscheint es mir wenig begründet, in Anbetracht eines Anteils von 64% scheiternden Replikationen (gemessen am ursprünglichen Evaluationskriterium, einem Signifikanztest) zu argumentieren, dass die überwiegende Mehrzahl der zugrundeliegenden Hypothesen wohl dennoch wahr sein wird. Spezifische Schlussfolgerungen über den Anteil falsch positiver Befunde werden letztlich aber notwendigerweise von Zusatzannahmen abhängig und damit strittig bleiben. Die unzureichende Replizierbarkeit psychologischer Befunde zeigt hingegen eindeutig, dass die in psychologischen Top-Journals berichtete und augenscheinlich in unserer *scientific community* weitgehend akzeptierte Evidenz schwach und unzuverlässig ist. Unsere Studie weist zudem nach, dass die veröffentlichte Evidenz systematisch zugunsten der geprüften Hypothesen verzerrt ist. Schwache, unzuverlässige und verzerrte Evidenz stehen einer empirischen Wissenschaft gewiss nicht gut zu Gesicht. Man wird auch kaum behaupten können, dass eine Wissenschaft mit schwacher und verzerrter Evidenz sonderlich effizient valides und belastbares Wissen generiert – selbst wenn einige der nicht-replizierten Befunde dennoch mit korrekten Hypothesen verbunden sein sollten. So uns also an einer Qualitätssicherung in der psychologischen Forschung liegt, ist die wesentliche Schlussfolgerung aus den Ergebnissen des Replikationsprojekts sehr simpel: Wir sollten dringlich etwas ändern!

Die Probleme schwacher und verzerrter Evidenz seien hier auch nochmal anhand einiger Kritikpunkte im Beitrag von Klaus Fiedler und der Fehlinterpretation unserer Ergebnisse in der Pressemitteilung dargestellt – ausführlichere Stellungnahmen zu manchen der Kritikpunkte finden sich allerdings bereits in anderen Forumsbeiträgen.

Regressionseffekt: Selbstverständlich unterliegen die Effekte in den Replikationsstudien einer Regression zur Mitte – dies ergibt sich in der Tat daraus, dass die Ergebnisse der Primärstudien nicht perfekt reliabel sein können. Wir sollten aufgrund dieser Regression zur Mitte beispielsweise erwarten, dass besonders große Effekte aus den Primärstudien in den Replikationsstudien kleiner werden, aber auch dass besonders kleine Originaleffekte in den Replikationsstudien größer ausfallen. Wenn sich die Ergebnisse aus Primär- und Replikationsstudien *allein* aufgrund des Regressionseffekts unterschieden, wäre zudem zu erwarten, dass (etwa) 50% der Replikationsstudien größere Effekte finden als die entsprechenden Primärstudien. Dies ist nicht, was wir beobachten. Tatsächlich sank der beobachtete Effekt in 83% der Replikationsstudien ab. Entsprechend bewegt sich in unserem Fall auch die „Mitte“: Der Mittelwert der Effektstärken der Primärstudien ist größer als der Mittelwert in den Replikationsstudien. Dieser Effekt fällt sehr deutlich aus: Die mittlere korrelative Effektgröße sinkt für das Gesamt aller Studien von .40 auf .20, für kognitionspsychologische Studien (aus JEP:LMC) von .47 auf .27 und für sozialpsychologische Studien (aus JPSP) von .29 auf .07 – ein Befund, der in der Pressemitteilung der DGPs leider gänzlich übersehen wird. Dieser Befund kann nicht durch einen Regressionseffekt erklärt werden (eine ausführliche Begründung geben Moritz Heene und Ulrich Schimmack in ihrem Forumsbeitrag). Um diesen Befund zu erklären, muss man annehmen, dass die ursprünglich veröffentlichten Ergebnisse einer Selektion unterlagen, wie sie etwa durch Publication- und Reporting Biases oder *p*-hacking entsteht.

Dass eine solche Verzerrung der veröffentlichten Evidenz zugunsten der geprüften Hypothesen existiert, kann man mit guten Gründen für wenig überraschend halten. Aufgrund unserer wohl weitgehend geteilten Erfahrung mit der Forschungs- und Publikations-Praxis ist nichts anderes zu erwarten. Daneben zeigen auch diverse theoretische Analysen, dass eine solche Verzerrung anzunehmen ist (z.B. Fiedler, 2011). Der von Klaus Fiedler – wie ich hoffe zurecht – vermutete besondere Einfluss des Replikationsprojekts scheint mir aber just darauf zurückzugehen, dass das theoretische Wissen, dass es so sein muss, weit weniger wirksam ist, als die empirische Demonstration, dass es tatsächlich so ist. Das Replikationsprojekt liefert zudem erstmals eine empirische Schätzung der Größe der Verzerrung in publizierter Evidenz.

Reliabilität: Ich teile die Einschätzung Klaus Fiedlers, dass unzureichende Reliabilität der abhängigen Maße möglicherweise ein wesentliches Problem vieler psychologischer Studien ist. Dieses Problem betrifft offensichtlich zunächst die Primärstudien. Da die Replikationsstudien als direkte Replikationen die Maße der Primärstudien übernehmen, ist anzunehmen, dass sie gegebenenfalls auch ihre schlechte Reliabilität „erben“. Da das Ziel des Replikationsprojekts aber nicht darin bestand, Schwächen und Methodenprobleme der Primärstudien zu korrigieren, sondern eben darin, die Replizierbarkeit dieser Studien zu evaluieren, schwächt dies die Aussage unserer Befunde nicht – es handelt sich hier eher um ein *feature* als um einen *bug* der Replikationsstudien. Möglicherweise liefern also sowohl Primär- als auch

Replikationsstudie schwache Evidenz, z.B. aufgrund unzureichender Reliabilität. Dies kann natürlich auch abweichende Befunde in beiden Studien erklären. Im Hinblick auf das Replikationsprojekt ist das kein Problem, sondern war gerade der Erkenntnisgegenstand. Das Problem besteht darin, dass die schwache und daher nicht-replizierbare Evidenz dennoch (in einem Top-Journal) veröffentlicht wurde. Mangelhafte Reliabilität kann zudem nicht die systematisch verringerten Effektgrößen in den Replikationsstudien erklären. Möchte man die Evidenz beider Studien relativ bewerten, wird man daher dennoch schließen müssen, dass die Replikationsstudien die bessere Evidenz liefern – diese Studien waren zumindest nicht den Selektionsmechanismen ausgesetzt, denen die Primärstudien offensichtlich unterlagen.

Sehr ähnliche Argumente gelten auch für diverse andere Kritikpunkte: Zum Beispiel mag ein Manipulation Check in einem Teil der betrachteten Studien tatsächlich sehr wünschenswert sein. Fand ein solcher Manipulation Check in der Primärstudie statt, wurde er auch in der Replikationsstudie wiederholt. Fand in der Primärstudie kein Manipulation Check statt, stellt sich erneut die Frage, warum dies (exklusiv) ein Makel der Replikationsstudie sein sollte. Wenn denn unklar oder strittig ist, ob die gewählte Operationalisierung tatsächlich die relevante theoretische Variable manipuliert, sollte natürlich auch die Primärstudie einen unabhängigen Nachweis erbringen, dass die Manipulation gelungen ist – und damit zumindest auch einen Hinweis darauf geben, wie die theoretische Variable denn erfasst werden kann. Man mag auch darüber streiten, ob Power nicht nur als Funktion der Stichprobengröße sondern auch als Funktion diverser Quellen unsystematischer Varianz aufgefasst werden sollte. Man kann hingegen nicht darüber streiten, dass die Replikationsstudien die Power der Primärstudien in der Regel verbessert haben, indem sie die Stichprobe vergrößert haben.

Moderatoren und Kontextabhängigkeit: Natürlich können noch unbekannte Moderatoren eine Ursache scheiternder Replikationen sein. Falsch ist hingegen die Interpretation der Pressemitteilung, dass scheiternde Replikationen bereits zeigen, dass solche Moderatoren wirksam sind. Offensichtlich gibt es mehr als eine mögliche Ursache für Nicht-Replizierbarkeit. Das Replikationsprojekt selbst kann keine Auskunft darüber geben, welche oder wie viele Befunde tatsächlich kontextabhängig sind – anhand von nur einer Replikation lässt sich die Wirksamkeit von Moderatoren nicht beurteilen. Da sich die Psychologie bislang eben nicht sonderlich für Replikationen oder die Generalisierbarkeit ihrer Befunde interessiert hat, ist die sonstige Evidenz zur Frage der Kontextabhängigkeit spärlich. Allerdings hat ein weiteres OSF-Projekt diese Frage untersucht (die „Many Labs“ Studie, Klein et al., 2014; s.a. den Forumsbeitrag von Moritz Heene). Das Resultat war, dass sich in zahlreichen direkten Replikationen verschiedener Befunde keine oder nur schwache Kontextabhängigkeit fand. Es gibt also keine Belege dafür, dass die mangelhafte Replizierbarkeit psychologischer Befunde tatsächlich auf unbekannte Moderatoren zurückgeht. Eindeutig ist hingegen, dass die Selektionsmechanismen, denen veröffentlichte Befunde unterliegen, erheblich zum Replikationsproblem beitragen.

Noch wichtiger als die Tatsache, dass dem Argument der Kontextabhängigkeit die empirische Grundlage fehlt, erscheint mir allerdings, dass dieses Argument stets *post hoc* vorgebracht wird. Die Studien des Replikationsprojekts sind unter Randbedingungen durchgeführt worden, auf die der ursprüngliche Befund laut Darstellung im Original-Artikel generalisieren sollte. Wenn das Scheitern einer Replikation nun auf bloß im Nachhinein vermutete Moderatoren zurückgeführt werden kann, ist es schlicht unmöglich, irgendeinen Befund jemals als falsch positiven Befund zu interpretieren – und damit ebenfalls ausgeschlossen, die zugrundeliegende Hypothese zu falsifizieren.

Qualitätsprobleme in der psychologischen Forschung sind offensichtlich. Sie waren das auch schon vor der Veröffentlichung der Befunde des Replikationsprojekts. Zahlreiche Belege finden sich in den Literaturhinweisen der Forumsbeiträge von Matthias Reitzle sowie Felix Schönbrodt und Kollegen. Das Replikationsprojekt demonstriert die Konsequenzen dieser Probleme empirisch und erlaubt es, einen Teil dieser Konsequenzen zu quantifizieren. Diese Befunde mögen für die Psychologie zunächst unliebsam sein. Es wird aber kaum helfen, die Augen zu verschließen. Die positive Nachricht liegt darin, dass wir es besser machen können. Die Botschaft, die von der DGPs meines Erachtens ausgesendet werden sollte, lautet daher, dass wir es auch besser machen werden. In der Außendarstellung wird uns die Tatsache helfen, dass wir weder mit Replikationsproblemen noch mit den zugrundeliegenden Ursachen allein sind (einen sehr lesenswerten Kommentar zum Zustand der bio-medizinischen Forschung gibt Horton, 2015). Die Psychologie hat also die Chance, eine Vorreiterrolle dabei zu übernehmen, die Forschungs- und Publikationspraxis zu verbessern – aufgrund diverser Initiativen und bereits eingeleiteter Veränderungen wäre sie dafür momentan ganz gut aufgestellt. Auch nach innen kann die Nachricht, dass wir es besser machen können und werden, aber gewiss motivierend wirken: Wir müssen gar nicht zwangsläufig durch ein Meer aus nicht-replizierbaren Befunden waten, unpassende Ergebnisse umdeuten oder vergessen oder uns für psychologische Belege extrasensorischer Wahrnehmung schämen. Das öffentliche Teilen von Materialien und Daten, generell mehr Transparenz im Forschungsprozess, vorregistrierte Studien, mehr Replikationen und veränderte Anreizstrukturen werden die Verzerrung in veröffentlichter Evidenz reduzieren und damit auch die Replikationsrate psychologischer Befunde verbessern. Auf der Ebene universitärer Institute weist die Initiative des Department Psychologie an der LMU den Weg. Derartige Maßnahmen und Initiativen sollte die DGPs mit all ihren Möglichkeiten fördern.

Frank Renkewitz

Fiedler, K. (2011). Voodoo correlations are everywhere – Not only in social neurosciences. *Perspectives on Psychological Science, 6*, 163-171.

Horton, R. (2015). Offline: What is medicine's 5 sigma? *The Lancet, 385,* 1380.

Klein R.A. et al. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology, 45,* 142-152.

09.09.2015

Die Reproducibility-Studie: Der empirische Test einer Theorie

Erich H. Witte

Liebe Kolleginnen und Kollegen,

offensichtlich war der Tranquilizer, die Pressemitteilung der DGPs, nicht völlig bei der inneren Diskussion wirkungsvoll. Sie geht weiter und hat wohl bereits vorher zu Konsequenzen an der LMU geführt. Diese Konsequenzen sind sehr zu begrüßen. Sie stellen eine verbesserte Diagnostik dar, aber über die Heilbehandlung werden wir noch Diskussionen führen müssen. Sie wird einschneidender sein, als wir es uns noch erhoffen und sie folgt nicht aus den getroffenen diagnostischen Maßnahmen.

Ich möchte es vermeiden, um möglichst viele KollegInnen zu erreichen, eine technische Diskussion über die Statistik-Schulen zu führen. Das können wir Experten intern tun, um dann mögliche Varianten in der Zukunft vorzuschlagen.

Die jetzt veröffentlichten Ergebnisse basieren nicht auf neuen Erkenntnissen noch einer neuen inferenzstatistischen Theorie. Diese Studie ist eine empirische Bestätigung dessen, was wir in der Vergangenheit bei unserer Forschung getan haben. Wir hätten die Ergebnisse recht genau vorhersagen können, nur hat sich bisher niemand diese Mühe der empirischen Prüfung gemacht. Eigentlich kann niemand mehr die Augen vor diesem Ergebnis verschließen.

Jeder von uns kennt die Diskussion um die POWER eines Tests. Jeder von uns kennt die Klagen um die geringe POWER unserer Untersuchungen. Cohen (1962) hat bereits auf diese Problematik hingewiesen. Er fand einen Wert von (1-ß) = 0.46 . Jetzt haben wir einen Wert von (1-ß)=0.35 berichtet bekommen (Bakker et al., 2012). Es ist in dem letzten halben Jahrhundert eher schlechter als besser geworden. Wenn wir die aktuelleren Daten nehmen, dann gibt es eine klare Prognose aus einer Neyman-Pearson-Inferenztheorie, die ja den Kern unserer Theorienprüfung darstellt: Es sind 35% der vom Zufall abweichenden Ergebnisse replizierbar. Genau das ist eingetreten. Wen kann das eigentlich wundern? Die POWER ist die Wahrscheinlichkeit eine „wahre“ Alternativhypothese in unserer Untersuchung entdeckt zu haben. Irgendwie haben wir das ja in einem Statistik-Kurs zu Beginn des Studiums gelernt. Dann aber schnell vergessen, weil uns vielfältige Umstände zu einer Publikation signifikanter Effekte gezwungen haben.

Beginnen wir mit dieser simplen Erkenntnis, dass sich unsere Inferenz-Theorie bewährt hat. Sie sagt sehr gut den Prozentsatz nicht-replizierbarer Ergebnisse vorher. Leider ist die Replizierbarkeit der Gold-Standard jeder empirischen Wissenschaft (auch die Replizierbarkeit der Veränderung, um auf Jüttemann einzugehen).

Bakker, M., van Dijk, A., & Wicherts, J.M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543-554.

Cohen, J. (1962). The statistical power of abnormal social psychological research: A review. *Journal of abnormal and social Psychology, 65,* 145-153.

Vielleicht nur ein Symptom

PD Dr. Matthias Reitzle

Liebe Kolleginnen und Kollegen,

Ich halte die Befunde des Replikationsprojektes eher für ein Symptom, nicht für das eigentliche Problem. Daher ist die Debatte um Regression zur Mitte bzw. die technische Kritik an der Kritik in meinen Augen ein Nebenkriegsschauplatz. Vielmehr leistet der von Detlef Rost beklagte und zutreffend erklärte Mangel an Replikationsstudien einem bisweilen nonchalanten Umgang mit der Qualität bei psychologischen Studien Vorschub. Wird nur publiziert, nicht rezipiert und nicht repliziert, ist die Gefahr, widerlegt oder auf Mängel in der eigenen Arbeit aufmerksam gemacht zu werden, gering.

Das Problem mangelnder Replizierbarkeit ist weder durch Kontext noch durch die unvermeidbare Stichprobenvariation hinreichend erklärt. Eine gewichtige Rolle spielt, einmal abgesehen von der paradigmatischen Fehlannahme der kontextuellen und zeitlichen Invarianz menschlichen Funktionierens (s. Gerd Jüttemanns Beitrag vom 7.9.), oftmals ad hoc produzierter Wildwuchs in Sachen Konstruktdefinition, Operationalisierung/Messung, unpassender und/oder fehlinterpretierter Methode und Verletzung statistischer Annahmen. Die ebenso ideosynkratische wie uneindeutige Benennung und Operationalisierung psychologischer Konstrukte hat Block (2000, fußend auf Thorndike und Kelley) humorvoll als "jingle fallacy" und "jangle fallacy" beschrieben. Der Qualitätsausweis von „Messung“ reduziert sich in der Regel auf Cronbach’s Alpha, die Validitätsfrage bleibt zumeist unterbelichtet, eine Grundsatzdebatte um psychologisches Messen (s. z.B. Michell, 2008) findet, wenn überhaupt, in kleinen Zirkeln fernab vom Forschungsalltag statt. Kognitive Fallen in Fragebogenerhebungen (z.B. N. Schwarz, 1999, 2007) finden in diesem Alltag ebenso wenig Beachtung wie kritische Debatten in der Psychological Inquiry (z. B. Smedslund, 1991, s. auch seine Beiträge in Theory and Psychology, 2009, 2012a und 2012b), wie provokante Beiträge von Toomela (2008), Valsiner (2009) M. Schwarz (2009) in Integrative Psychology & Behavioral Science oder wie Molenaars (z.B. 2004) unermüdliches Bemühen, dem forschenden Kollegium die eingeschränkte Aussagekraft von Aggregatbefunden (mangels "ergodicity") für den Einzelfall nahe zu bringen.

Entgegen landläufiger Meinung ist Peer Review nicht unbedingt Schutz gegen Unzulänglichkeiten. Denn viel Zeit und Hirnschmalz in eine gründliche Begutachtung einschließlich Recherche in zitierter Literatur etc. zu investieren dient ebenso wenig der eigenen Karriere wie es Rost für die Durchführung von Replikationsstudien feststellt. So diffundieren vielfach „billige Texte“ (Clemens Albrecht in Forschung & Lehre 5/14, S. 341) in den Markt. Im schlimmsten Fall wird heterogene und zum Teil fragwürdige Massenware zu Meta-Analysen verdichtet und erreicht auf diesem Wege, obwohl als Einzelstudien nahezu unrezipiert und unrepliziert, den Adelsrang unumstößlicher wissenschaftlicher Erkenntnis. Ich räume ein, dass das Gesagte wohl häufiger für Fragebogen-basierte Feldstudien als für gut geplante und durchdachte Experimente gilt. Außerdem steht mir keine Meinung außerhalb meiner Teildisziplin zu.

Dennoch: „business as usual“ und ein rein quantitatives Verständnis von sog. „Produktivität“ hat zu einer Lage geführt, die Harré (2000) bereits vor fünfzehn Jahren, ebenfalls in SCIENCE, folgendermaßen charakterisierte: „It is a remarkable feature of mainstream academic psychology that, alone among the sciences, it should be almost wholly immune to critical appraisal as an enterprise. Methods that have long been shown to be ineffective or worse are still used on a routine basis by hundreds, perhaps thousands of people. Conceptual muddles long exposed to view are evident in almost every issue of standard psychology journals. This is a curious state of affairs. New pathways and more realistic paradigms of research have been proposed, demonstrated, and ignored (p. 1303).“

Zum guten Schluss ein Kommentar zu Gerd Jüttemanns Beitrag: Sicher ist es gewagt, angesichts des hochkomplexen Mensch-Umwelt-Systems Zeit- und Kontextinvarianz menschlichen Funktionierens als Regelfall zu erwarten. Ähnlich abwegig ist die Annahme stationärer Prozesse im Verlauf menschlicher Entwicklung (Molenaar, 2004). Nur was heißt dies für die Replikation als Instrument der Qualitätssicherung empirischer Erkenntnis? Soll man jedwedes Ergebnis gleichermaßen ernst nehmen und der Spezifität des bio-psycho-sozio-historischen „Augenblicks“ zuschreiben? Das käme einem Persilschein für Zufallsergebnisse gleich. Soll man zu divergierenden Befunden ex post facto eine plausibel klingende „Story“ generieren? Das begünstigt Deutungswillkür. Auch wenn es nie vollständig gelingen wird, wäre ein lohnenswertes Ziel, charakteristische Grundzüge dieses „Augenblicks“ theoriegeleitet in das Untersuchungsmodell einzubeziehen? Darüber hinaus hilfreich ist unbedingte Transparenz und Überprüfbarkeit durch obligatorische Publikation der Rohdaten (Simonsohn, 2013).

Block, J. (2000). Three tasks for personality psychology. In L. R. Bergman, R. B. Cairns, L.-G. Nilsson & L. Nystedt (Eds.), Developmental Science and the holistic approach. Mahwah, NJ: Erlbaum.

Michell, J. (2008). Is psychometrics pathological science? Measurement, 6, 7-24.

Molenaar, P. C. M. (2004). A manifesto on psychology as ideographic science: Bringing the person back into scientific psychology, this time forever. Measurement, 2, 201-218.

Schwarz, M. (2009). Is psychology based on a methodological error? Integrative Psychological & Behavioral Science, 43, 185-213.

Schwarz, N. (1999). How the questions shape the answers. American Psychologist, 54, 93-105.

Schwarz, N. (2007). Cognitive aspects of survey methodology. Applied Cognitive Psychology, 21, 277-287.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. Psychological Science, 24, 1875-1888.

Smedslund, J. (1991). The pseudoempirical in psychology and the case for psychlogic. Psychological Inquiry, 2, 325-338.

Smedslund, J. (2009). The mismatch between current research methods and the nature of psychological phenomena: What researchers must learn from practitioners. Theory & Psychology, 19, 778-794.

Smedslund, J. (2012a). The bricoleur model of psychological practice. Theory & Psychology, 22, 643-657.

Smedslund, J. (2012b). What follows from what we all know about human beings. Theory & Psychology, 22, 658-668.

Toomela, A. (2008). Variables in psychology: A critique of quantitative psychology. Integrative Psychological & Behavioral Science, 42, 245-265.

Valsiner, J. (2009). Integrating psychology within the globalizing world: A requiem to the post-modernist experiment with Wissenschaft. Integrative Psychological & Behavioral Science, 43, 1-21.

Die Glaubwürdigkeitskrise - wie kann es weitergehen?

von PD Dr. Felix Schönbrodt, Prof. Dr. Moritz Heene, Prof. Dr. Markus Maier, PD Dr. Michael Zehetleitner, Prof. Dr. Markus Bühner

Anschließend an mehrere bisherigen Kommentare möchten wir betonen, dass wir die in der Pressemitteilung herausgehobene Zahl von "68% Replikationsanteil" als nicht plausibel ansehen. Wenn es verzerrte Effektschätzungen durch Publikationsbias gibt (und dieser ist bei dem Set an Originalstudien völlig unbestritten), dann erhält man üblicherweise eine bessere Schätzung des wahren Effekts, wenn man **nicht** beide Studien meta-analytisch zusammenfasst, sondern sich **nur** auf die bias-freie Studie verlässt (Nuijten, van Assen, Feldkamp, & Wicherts, 2015).

Die äußerst lesenswerte *Bayesianische Reanalyse* von Alexander Etz (http://alexanderetz.com/2015/08/30/the-bayesian-reproducibility-project/) erlaubt darüber hinaus eine differenziertere Beurteilung der Ergebnisse, indem die Frage beantwortet wird, die für die meisten wohl wirklich relevant ist: "Ist der gefundene Replikationseffekt kompatibel zu dem, was in der Originalstudie gefunden wurde, oder ist er eher gar nicht vorhanden?".
Wenn man diese Reanalyse auf wenige Zahlen herunterbrechen will, dann resultierte das RP:P in folgendem Ergebnis:

- 34% der Replikationsstudien zeigen Evidenz **für** den Originalbefund
- 38% der Replikationsstudien zeigen Evidenz **gegen** den Originalbefund und für die H0.
- Bei den verbleibenden 28% ist die Evidenz der Replikationen uneindeutig - dort gibt es weder starke Evidenz für den Originalbefund, noch starke Evidenz für die Abwesenheit des Effekts. Von der Tendenz her sprechen zwei Drittel dieser uneindeutigen Ergebnisse jedoch eher gegen den Originalbefund.

Das RP:P, die vielen weiteren kollaborativen Projekte wie ManyLabs 1-3 (centerforopenscience.org/communities/), sowie Befunde zu "questionable research practices" (Bakker, van Dijk, & Wicherts, 2012; Francis, Tanzman, & Matthews, 2014; John, Loewenstein, & Prelec, 2011; O'Boyle, Banks, & Gonzalez-Mulé, 2014; Simmons, Nelson, & Simonsohn, 2011; Schimmack, 2012) haben das Vertrauen in die Glaubwürdigkeit und Replizierbarkeit vieler psychologischer Befunde erschüttert.
Im Gegensatz zu der recht optimistischen Presseerklärung der DGPs sind wir persönlich der Meinung, dass ein einfaches "Weiter so!" nicht angemessen ist. Daher überlegen wir, was man konkret tun kann, um mit den Worten von J. Ioannidis (2014) das Ziel zu erreichen: "to make more published research true."

Am Department Psychologie der LMU haben wir als Reaktion darauf im Juli ein "*Open-Science-Komitee*" ins Leben gerufen (siehe auch www.nicebread.de/introducing-the-open-science-committee-at-our-department/). Die Ziele des Komitees sind:

- Aktuelle internationale Entwicklungen zu dem Thema zu beobachten und in das Department zu kommunizieren.
- Konkrete Vorschläge bzgl. tenure-track-Kriterien, Berufungsverfahren, Doktorandenbetreuung, etc. weiterzuentwickeln, so dass die richtigen Anreize für transparente und replizierbare Forschung gesetzt werden.
- Workshops und Fortbildungen zu organisieren (z.B.: Wie führe ich eine Präregistrierung durch? Was sind registered reports? Wie publiziere ich Open Data?).

- Die departmentinterne Diskussion über mögliche Konsequenzen aus der Replikationskrise zu bündeln und auszuloten, ob gemeinsame Standards entwickelt werden können.

Darüber hinaus haben zwei Forschungseinheiten eine *Selbstverpflichtung zu Transparenz und Forschungsstandards* entwickelt, die bereits
auf der Webseite öffentlich gemacht wurde (https://osf.io/mgwk8/). Diese definiert Standards zu eigener Forschung, Betreuung von Doktoranden, und peer reviews. Diese Selbstverpflichtung soll unter anderem nach Außen signalisieren, dass wir nicht zum "business as usual" zurückkehren werden, nachdem sich die Wogen vielleicht geglättet haben.

Mit diesen ersten konkreten Schritten hoffen wir, nicht nur auf individueller Ebene ein Umdenken zu erreichen, sondern auch auf struktureller Ebene Anreize zu setzen, die eine transparente Forschungspraxis belohnen (anstatt sie wie bisher eher zu bestrafen). Diese Schritte auf Departmentsebene kommen auch dem aktuellen Ruf nach, dass die *Institutionen ihren Beitrag zur Erhöhung der Reproduzierbarkeit* unserer Forschung leisten sollen (www.nature.com/news/robust-research-institutions-must-do-their-part-for-reproducibility-1.18259). Die Autoren schreiben:
"The systems needed to promote reproducible research must come from institutions — scientists, funders and journals cannot build them on their own. [...] Even if it is accompanied by an apparent decrease in productivity, the resulting increase in research quality will be well worth the costs."

Wie kann die DGPs als Institution dazu beitragen, neue Anreize für Transparenz und Replizierbarkeit zu setzen?

Referenzen

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. Perspectives on Psychological Science, 7(6), 543–554. doi.org/10.1177/1745691612459060

Francis, G., Tanzman, J., & Matthews, W. J. (2014). Excess Success for Psychology Articles in the Journal Science. PLoS ONE, 9(12), e114255. http://doi.org/10.1371/journal.pone.0114255

Ioannidis, J. P. A. (2014). How to Make More Published Research True. PLoS Medicine, 11(10), e1001747. doi.org/10.1371/journal.pmed.1001747

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices with Incentives for Truth-Telling. Psychological Science, Forthcoming. Abgerufen von http://dionysus.psych.wisc.edu/lit/articles/JohnL2011a.pdf

Nuijten, M. B., van Assen, M. A. L. M., Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. Review of General Psychology, 19, 172–182.

O'Boyle, E. H. J., Banks, G. C., & Gonzalez-Mulé, E. (2014). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. Journal of Management, 0149206314527133. http://doi.org/10.1177/0149206314527133

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. Psychological Methods, 17(4), 551. http://doi.org/0.1037/a0029487

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. Psychological Science, 22(11), 1359–1366.

07.09.2015

Prof. Dr. Gerd Jüttemann

Das problematische Unvereränderlichkeitspostulat der Psychologie und die positive Interpretation nicht reproduzierbarer Untersuchungsergebnisse

Die Tatsache, dass sich ein großer Teil von experimentell durchgeführten psychologischen nicht replizieren lässt, wird von mancher Seite als ein Skandal bewertet. Sie könnte aber auch als eine Art Beweis für die Richtigkeit einer lange gehegten Vermutung aufgefasst werden und den Anlass für eine grundlegende Neuorientierung bilden.

Diese Annahme hat den folgenden Hintergrund: Der niederländische Psychiater J. H. van den Berg (1914 - 2012) verwies bereits vor mehr als 60 Jahren auf das problematische Unveränderlichkeitspostulat der Psychologie:

„Ist für die Psychologie, die sich auf das Postulat der Unveränderlichkeit stützt, das Leben eines vorigen Geschlechtes eine Variation auf ein bekanntes Thema, so gestattet die Voraussetzung, daß das menschliche Leben ein veränderliches Leben ist, den Gedanken, daß frühere Generationen anders, und zwar wesentlich anders lebten“ (van den Berg, 1960, S. 11; Hervorh. G.J.).

Van den Berg wollte das Unveränderlichkeitspostulat durch ein Veränderlichkeitspostulat ersetzen oder zumindest ergänzen. Sein Vorschlag blieb jedoch bis heute unbeachtet. Doch diese Situation könnte sich ändern. Voraussetzung dafür wäre die Anerkennung der Wahrheit, dass sich das Menschlich-Psychische zu einem erheblichen Teil im Laufe der Zeit verändert, und zwar nicht nur innerhalb eines individuellen Lebens, sondern auch im im Rahmen des intergenerationellen Wandels. Ausgehend von dieser Erkenntnis erscheint es naheliegend, misslungene Replikationen – teilweise – auf „normale“ Modifikationen des Psychischen zurückzuführen und zugleich als einen Beweis für die Gültigkeit des Veränderungspostulats zu werten. Langfristig betrachtet betrifft die Variabilität des Psychischen auch das Genom und ist In erster Linie evolutionsbiologisch zu erklären. Kurz- und mittelfristig gesehen ist sie jedoch kulturell bedingt. Aber um welche Vorgänge handelt es sich hier eigentlich? Eine generelle Antwort auf diese Frage lässt sich aus der Überschrift eines inzwischen ebenfalls weitgehend vergessenen Buchs des Schweizer Philosophen Michael Landmann ableiten, das kurze Zeit nach der Veröffentlichung der deutschen Ausgabe der Monographie von van den Berg erschien. Der Titel lautet:

„Der Mensch als Schöpfer und Geschöpf der Kultur“ (Landmann,1961).

Die Metapher verweist auf zwei grundlegende Prozesse, die beide auch für die Psychologie eine hohe Bedeutung besitzen, weil sich nicht nur das ganz elementare menschliche Erleben und Verhalten, sonders auch die Techniken und Themen (Thomae 1996) personaler Daseinsgestaltung in mancher Hinsicht ständig verändern. Das gilt sowohl im Hinblick auf die „produzierenden“ als auch auf die „konsumierenden“ Teilnehmenden an der Kultur.

Mit dem von Landmann zum Ausdruck gebrachten Gedanken wurde – im Wortlaut leicht abgewandelt – ein kürzlich erschienener Sammelband eingeleitet (Jüttemann 2014). Das Vorwort beginnt mit den Sätzen:

„Der Mensch verändert seine Welt, und die veränderte Welt verändert den Menschen. Das sind zwei zentrale Vorgänge, die die Entwicklung der Menschheit kennzeichnen und die sowohl für unsere Selbstvergewisserung als auch im Hinblick auf grundlegende Planungen einer besonderen Beachtung bedürfen.“

Die hier gemeinten Bewegungen verdeutlichen zugleich die – manchmal nicht direkt erkennbare – psychogenetische Relevanz vieler Fragestellungen, mit denen wir es in unserer Disziplin zu tun haben.

Der Begriff „Psychogenese“ bezieht sich in diesem Zusammenhang auf drei grundlegende Prozesse, die sich drei verschiedenen Zeithorizonten zuordnen lassen (vgl. Jüttemann 2013 und 2014, S. 32.: „Das Stufenmodell der Entwicklung des Menschen“). Besitzt der Gegenstand eines psychologischen Forschungsprojekts eine verdeckte psychogenetische Relevanz, dann sind im Falle einer Wiederholung der Untersuchung gleichlautende Ergebnisse nicht zu erwarten. Möglicherweise lassen sich in diesem Fall aber abweichende Resultate, ähnlich wie bei einem Vergleich zweier Zeitpunkte im Rahmen einer Längsschnittstudie, historisch-psychologisch interpretieren.

Berg, J. H. van den (1960). Metabletica. Über die Wandlungen des Menschen. Grundlinien einer historischen Psychologie (Übersetzung aus dem Holländischen). Göttingen: Vandenhoeck & Ruprecht. Landmann, M. (1961). Der Mensch als Schöpfer und Geschöpf der Kultur. Geschichts- und Sozialanthropologie. München / Basel: Ernst Reinhardt. Jüttemann. G. (2013) (Hg.). Die Entwicklung der Psyche in der Geschichte der Menschheit. Lengerich: Pabst Science Publishers Jüttemann. G. (2014) (Hg.). Entwicklungen der Menschheit / Humanwissenschaften in der Perspektive der Integration. Lengerich: Pabst Science Publishers. Thomae, H. (1996). Das Individuum und seine Welt. 3, erweiterte und verbesserte Auflage. Göttingen: Hogrefe

Gerd Jüttemann

07.09.2015

Erneute Verdeutlichung des Kernbefunds des Replikationsprojekts Psychologie

Liebe Kolleginnen und Kollegen,

am vergangenen Dienstag hat die DGPs eine Pressemitteilung zur Replikationsstudie der Open Science Collaboration veröffentlicht. Inzwischen hat sie zudem auf ihrer Homepage ein Forum zur Diskussion dieser Studie und der Pressemitteilung eingerichtet. Wir begrüßen diese Maßnahme. Um die Diskussion auf eine solidere Grundlage zu stellen,

möchten wir als Mit-Autoren der Studie darauf hinweisen, dass die Pressemitteilung der DGPs die Befunde zumindest an einer Stelle missverständlich darstellt: Anders als in der Pressemitteilung dargelegt, wurde in unserer Untersuchung kein Replikationsanteil von 68% ermittelt. Diese Zahl ergibt sich aus einer meta-analytischen Betrachtung der Effektstärken aus Original- und Replikationsstudien (wie auch kurz in der Pressemitteilung erwähnt). Sie beruht also auf einer Kombination der Evidenz beider Studien und ist daher nicht als reine Analyse der Replikationsdaten zu verstehen. Tatsächlich lag die Replikationsrate je nach verwendetem Evaluationskriterium lediglich zwischen 36% und 47%. Entsprechend haben die Mitglieder der Open Science Collaboration in eigenen Pressemitteilungen und diversen Pressegesprächen darauf hingewiesen, dass unabhängig von der verwendeten Analysemethode weniger als die Hälfte der ursprünglichen Befunde erfolgreich repliziert werden konnte.

Wir freuen uns auf die Diskussion und stehen selbstverständlich im Forum zur Verfügung, um weitere Fragen zu klären.

Frank Renkewitz, Andreas Glöckner, Susann Fiedler, Felix Henninger, Andreas Cordes, Angela Dorrough, Georg Jahn, Marc Jekel, Tim Kuhlmann, Johannes Meixner, Stephanie Müller, Franziska Plessow, René Schlegelmilch, Stefan Stieger, Carina Sonnleitner, Martin Voracek

06.09.2015

Eine Antwort auf Klaus Fiedler.

von Prof. Dr. Moritz Heene und Prof. Ulrich Schimmack

Wir möchten im Folgenden auf zwei zentrale Argumente von Hr. Fiedler eingehen:

1), dass die im Durchschnitt deutlich geringeren beobachteten Effekgrößen im Replikationsprojekt ein Artefakt der Regression zur Mitte darstellen sowie

2), dass Unreliabilität in den Replikationsstudien dort zu geringeren beobachteten Effekten geführt habe.

ad 1) Wie bereits hier von Heene ausgeführt, impliziert die Regression zur Mitte (Ergebnisse, die in einer ersten Messung extrem waren, "tendieren" in einer zweiten Messung zum Mittelwert), dass die originalen Effekte verzerrte, d.h. extreme Schätzungen des Populationseffektes darstellen, weil selektiv publiziert wurde. Dies wird in der Ausführung von Herr Fiedler zur Regression zur Mitte aber nicht genannt sowie missinterpretiert, wenn Hr. Fiedler in seinem Kommentar schreibt:

"(2) The only necessary and sufficient condition for regression (to the mean or toward less pronounced values) is a correlation less than zero. … One can refrain from assuming that the original findings have been over-estimations."

Man kann es also eben *nicht* unterlassen davon auszugehen, dass die originalen Ergebnisse Überschätzungen waren, denn selektives Publizieren ist eine notwendige Bedingung für die sehr deutlichen Reduktion der beobachteten Effektgrößen.

a) Herr Fiedler irrt, wenn er sich für seinen Beleg auf Furby (1973) bezieht, indem er schreibt: "The only necessary and sufficient condition for regression (to the mean or toward less pronounced values) is a correlation less than zero. This was nicely explained and proven by Furby (1973)". Interessanterweise verweist Furby (1973) in seinem Beispiel explizit auf die Notwendigkeit einer Selektion ober- oder unterhalb des Populationsmittelwertes, wenn er schreibt: "Now let us choose a certain aggression level at Time 1 (any level other than the mean)".

Nur, um diesen Sachverhalt zu illustrieren: Das erwartete Ausmaß der Regression zur Mitte ist gegeben durch (1 – r)*(mu – M), wobei r: Korrelation zwischen der ersten und zweiten Messung, mu: Mittelwert in der Population, M: Mittelwert in der selegierten Gruppe. Wenn bspw. r = .80 (also kleiner eins wie von Hr. Fiedler vorausgesetzt) und der Mittelwerte der selegierten Gruppe gleich dem Populationsmittelwert, also M = mu, also bspw. M = mu = .40, dann tritt *kein* Regressionseffekt auf, denn (1 – .80)*(.40 – .40) = .20*0 = 0. Folglich ist die Bedingung r < 1 *keine* notwendige und hinreichende Bedingung für die Regression zur Mitte. Nur wenn r < 1 *und* M ungleich mu, tritt dieser Effekt auf.

b) Der Regressionseffekt kann positiv wie auch negativ sein. Wenn M < mu und r < 1, dann sind die Messwerte der zweiten Messung *größer* als der ersten Messung, die Tendenz zur Mitte also positiv. Wenn hingegen M > mu und r < 1, dann ist dieser Regressionseffekt negativ. In den Replikationstudien wurde ein negativer Effekt beobachtet, da man im Schnitt kleinere Effekte beobachtete. Das bedeutet, dass die beobachteten originalen Effekte im Schnitt über dem Populationseffekt gelegen haben müssen, also einen positiven bias aufgrund von publication bias (und evtl. auch durch p-hacking etc.) aufwiesen.

Die niedrigeren Effektgrößen in den Replikationsstudien lassen sich daher gut durch Publikationsbias zusammen mit

Regression zur Mitte erklären. Die OSF-Ergebnisse erlauben es zu schätzen, wie stark Publikationsbias zu einer Überschätzung berichteter Mittelwerte führte. Für die Sozialpsychologie fallen die durchschnittlichen Effektgrößen von Cohen's d = .6 auf d = .2 ab. Das zeigt eine Inflation von 200%. Es ist daher nicht überraschend, dass die Replikationsstudien so wenige signifikante Ergebnisse ergeben haben, da die geringe Erhöhung der Stichprobengröße den starken Abfall der Effektgrößen nicht ausgleichen konnte.

ad 2) Für den Fall der einfachen Regression gilt, dass der beobachtete Regressionskoeffizient b eines messfehlerbehafteteten Prädiktors X eine Funktion des Regressionskoeffizienten b_T für eine messfehler*freie* Variable X ist:

b = b_T*Rel(X).

(den mathematischen Beweis hierfür kann bei Interesse Hr. Heene per Email zusenden).

Die Formel impliziert, dass der beobachtete Regressionskoeffizient immer kleiner als derjenige für messfehlerfreie Variablen ist, wenn die Reliabilität kleiner als 1 ist. Das impliziert auch, dass die beobachtete Effektgröße, hier nämlich das Regressionsgewicht b, kleiner als die wahre Effektgröße b_T ausfällt. Dies führt, wie Hr. Fiedler korrekt bemerkt, natürlich zu einer Reduktion der statistischen Power (weil der Standardfehler des Regressionskoeffizienten größer wird). Dieses statistische Argument einer Reduktion der Power durch Unreliabilität in den Messungen trifft aber, und das ist der entscheidende Punkt, sowohl auf die originalen als auch die Replikationsstudien zu und würde daher zu keiner systematischen Reduktion der beobachteten Effektgrößen alleine in den Replikationsstudien führen. Kurz gesagt ist das Unreliabilitätsargument also gar keine stichhaltige Erklärung für die geringe Erfolgsquote in den Replikationsstudien.

Referenzen

Furby, L. (1973). Interpreting regression toward the mean in developmental research. Developmental Psychology, 8(2), 172-179. doi:10.1037/h0034145

04.09.2015

Prof. Dr. Erich H. Witte

Liebe KollegInnen,

die Reaktion des Vorstandes auf die höchst seriöse, fundamentale und differenzierte Reproducibility-Studie in Science ist sehr verständlich, denn es geht hierbei um eine Außendarstellung. Jede Mitteilung wird immer auch abhängig vom Adressaten sein.

Also beruhigen wir doch zuerst einmal die Öffentlichkeit. Im Tenor heißt das, wenn wir so weiter machen und uns nur noch ein wenig mehr anstrengen, dann werden wir auch besser. Ich hoffe, die Fachgruppe Methoden wird die notwendige interne Diskussion intensivieren und anleiten.

Herr Kollege Heene hat ja schon reagiert. Ich stimme ihm völlig zu. Da ich selber diese 36% in einem Rundbrief hervorgehoben habe, möchte ich nur zwei untechnische Bemerkungen machen:

1. Bei den 68% Replikationsrate sind 27 Studien ausgeschlossen worden von 100.

2. Wenn ich dem Sachverstand der Kolleginnen traue, und ich habe bei dieser Studie keinen Zweifel, dann werden 39% der Effekte als repliziert angesehen.

3. Die Zählmethode bei der 68%-Replikationsrate hat Herr Heene ja schon erwähnt. Hoffentlich bleiben wir den eigenen Ergebnissen wissenschaftlich so kritisch, wie wir es gegenüber fremden Ergebnissen zu sein haben.

Erich H. Witte

04.09.2015

Prof. Dr. Klaus Fiedler (Universität Heidelberg)

Besten Dank für die Mitteilung der DGPs bezüglich Brian Nosek's reprodicibility project. Ich finde auch, dass die neuen Facetten der wissenschaftlichen Arbeit – wie Replication, Pre-Registration etc. – bereichernd wirken könnten. Ich finde aber auch, dass auch so ein Replikations-Projekt ein durchaus anspruchsvolles Forschungsfeld darstellt, auf dem man viele Fehler machen kann und das mit methodischem Verstand beackert werden muss. Gerade weil eine solche Studie so eine

enorme Aufmerksamkeit bekommt und so bedeutsame Konsequenzen haben kann für das Ansehen der gesamten Psychologie, finde ich, die Regeln guter wissenschaftlicher Arbeit sind auf eine solche Studie (mindestens) ebenso streng anzuwenden wie auf gewöhnliche empirische Studien.

Für diejenigen in der DGPs, die es interessiert, schicke ich unten einen Kommentar, den ich einem amerikanischen Journalisten (Bruce Bowers) überlassen habe. Darin erkläre ich, warum der Nosek Report nicht „state-of-the art" ist.

Beste Grüße Klaus Fiedler

First of all I want to make it clear that I have been a big fan of properly conducted replication and validation studies for many years – long before the current hype of what one might call a shallow replication research program. Please note also that one of my own studies has been included in the present replication project; the original findings have been borne out more clearly than in the original study. So there is no self-referent motive for me to be overly critical.

However, I have to say that I am more than disappointed by the present report. In my view, such an expensive, time-consuming, and resource-intensive replication study, which can be expected to receive so much attention and to have such a strong impact on the field and on its public image, should live up (at least) to the same standards of scientific scrutiny as the studies that it evaluates. I'm afraid this is not the case, for the following reasons …

The rationale is to plot the effect size of replication results as a function of original results. Such a plot is necessarily subject to regression toward the mean. On a-priori-grounds, to the extent that the reliability of the original results is less than perfect, it can be expected that replication studies regress toward weaker effect sizes. This is very common knowledge. In a scholarly article one would try to compare the obtained effects to what can be expected from regression alone. The rule is simple and straightforward. Multiply the effect size of the original study (as a deviation score) with the reliability of the original test, and you get the expected replication results (in deviation scores) – as expected from regression alone. The informative question is to what extent the obtained results are weaker than the to-be-expected regressive results.

To be sure, the article's muteness regarding regression is related to the fact that the reliability was not assessed. This is a huge source of weakness. It has been shown (in a nice recent article by Stanley & Spence, 2014, in PPS) that measurement error and sampling error alone will greatly reduce the replicability of empirical results, even when the hypothesis is completely correct. In order not to be fooled by statistical data, it is therefore of utmost importance to control for measurement error and sampling error. This is the lesson we took from Frank Schmidt (2010). It is also very common wisdom.

The failure to assess the reliability of the dependent measures greatly reduces the interpretation of the results. Some studies may use single measures to assess an effect whereas others may use multiple measures and thereby enhance the reliability, according to a principle well-known since Spearman & Brown. Thus, some of the replication failures may simply reflect the naïve reliance on single-item dependent measures. This is of course a weakness of the original studies, but a weakness different from non-replicability of the theoretically important effect. Indeed, contrary to the notion that researchers perfectly exploit their degrees of freedom and always come up with results that overestimate their true effect size, they often make naïve mistakes.

By the way, this failure to control for reliability might explain the apparent replication advantage of cognitive over social psychology. Social psychologists may simply often rely on singular measure, whereas cognitive psychologists use multi-trial designs resulting in much higher reliability.

The failure to consider reliability refers to the dependent measure. A similar failure to systematically include manipulation checks renders the independent variables equivocal. The so-called Duhem-Quine problem refers to the unwarranted assumption that some experimental manipulation can be equated with the theoretical variable. An independent variable can be operationalized in multiple ways. A manipulation that worked a few years ago need to work now, simply because no manipulation provides a plain manipulation of the theoretical variable proper. It is therefore essential to include a manipulation check, to make sure that the very premise of a study is met, namely a successful manipulation of the theoretical variable. Simply running the same operational procedure as years before is not sufficient, logically.

Last but not least, the sampling rule that underlies the selection of the 100 studies strikes me as hard to tolerate. Replication teams could select their studies from the first 20 articles published in a journal in a year (if I correctly understand this sentence). What might have motivated the replication teams' choices? Could this procedure be sensitive to their attitude towards particular authors or their research? Could they have selected simply studies with a single dependent measure (implying low reliability)? – I do not want to be too suspicious here but, given the costs of the replication project and the human resources, does this sampling procedure represent the kind of high-quality science the whole project is striving for?

Across all replication studies, power is presupposed to be a pure function of the size of participant samples. The notion of a truly representative design in which tasks and stimuli and context conditions and a number of other boundary conditions are taken into account is not even mentioned (cf. Westfall & Judd).

Frankly speaking, as a journal editor (I am an Associate Editor of the Journal of Experimental Psychology: General) I would not accept such a report for publication. I wonder how it was accepted for Science, and what review process did not notice all these weaknesses.

Kind regards, Klaus

P.S. Es gab dann in Facebook eine Menge Echo. Darauf habe ich in einem zweiten Kommentar geantwortet, der wie auch der erste von meinem Kollegen Mickey Inzlicht ins Netz gestellt wurde. Ich schreibe das nur, um bestimmten Reaktionen vorzubeugen:

Having read the echo to my earlier comment on the Nosek report, I got the feeling that I should add some more clarifying remarks.

(1) With respect to my complaints about the complete failure to take regressiveness into account, some folks seem to suggest that this problem can be handled simply by increasing the power of the replication study and that power is a sole function of N, the number of participants. Both beliefs are mistaken. Statistical power is not just a function of N, but also depends on treating stimuli as a random factor (cf. recent papers by Westfall & Judd). Power is 1 minus β, the probability that a theoretical hypothesis, which is true, will be actually borne out in a study. This probability not only depends on N. It also depends on the appropriateness of selected stimuli, task parameters, instructions, boundary conditions etc. Even with 1000 participant per cell, measurement and sampling error can be high, for instance, when a test includes weakly selected items, or not enough items. It is a cardinal mistake to reduce power to N.

(2)The only necessary and sufficient condition for regression (to the mean or toward less pronounced values) is a correlation less than zero. This was nicely explained and proven by Furby (1973). We all “learned” that lesson in the first semester, but regression remains a counter-intuitive thing. When you plot effect sizes in the replication studies as a function of effect sizes in the original studies and the correlation between corresponding pairs is < 1, then there will be regression. The replication findings will be weaker than the original ones. One can refrain from assuming that the original findings have been over-estimations. One might represent the data the other way around, plotting the original results as a function of given effects in the replication studies, and one will also see regression. (Note in this connection that Etz’ Bayesian analysis of the replication project also identified quite a few replications that were “too strong”). For a nice illustration of this puzzling phenomenon, you may also want to read the Erev, Wallsten & Budescu (1994) paper, which shows both overconfidence and underconfidence in the same data array.

(3) I’m not saying that regression is easy to understand intuitively (Galton took many years to solve the puzzle). The very fact that people are easily fooled by regression is the reason why controlling for expected regression effects is standard in the kind of research published here. It is almost a prototypical example of what Don Campbell (1996) had in mind when he tried to warn the community from drawing erroneous inferences.

(4) I hope it is needless to repeat that controlling for the reliability of the original studies is essential, because variation in reliability affects the degree of regressiveness. It is particularly important to avoid premature interpretations of seemingly different replication results (e.g., for cognitive and social psychology) that could reflect nothing but unequal reliability.

(5) My critical remark that the replication studies did not include manipulation checks was also met with some spontaneous defensive reactions. Please note that the goal to run so-called “exact” replications (I refrain from discussing this notion here) does not prevent replication researchers from including additional groups supposed to estimate the effectiveness of a manipulation under the current conditions. (Needless to add that a manipulation check must be more than a compliant repetition of the instruction).

(6) Most importantly perhaps, I would like to reinforce my sincere opinion that methodological and ethical norms have to be applied to such an expensive, pretentious and potentially very consequential project even more carefully and strictly than they are applied to ordinary studies. Hardly any one of the 100 target studies could have a similarly strong impact, and call for a similar degree of responsibility, as the present replication project.

Klaus Fiedler

04.09.2015

Prof. Dr. Moritz Heene

Sehr geehrte Mitglieder des Vorstandes der DGPs,

Zunächst Dank an Sie für das Bemühen, die Ergebnisse des OSF-Replikationsprojektes der Öffentlichkeit klarer zu machen. Angesichts dieser Stellungnahme der DGPS möchte ich jedoch persönlich meinen Widerspruch dazu ausdrücken, da ich als

Mitglied der DGPS durchdiese Stellungnahmen in keiner Weise eine ausgewogene Sichtweise ausgedrückt sehe, sie im Gegenteil als sehr einseitig ansehe. Ich sehe diese Stellungnahme vielmehr als einen Euphemismus der Replikationsproblematik in der Psychologie an, um es milde auszudrücken, bin davon enttäuscht und hatte mir mehr erwartet. Meine Kritikpunkte an ihrer Stellungnahme:

1. Zum Argument 68% der Studien seien repliziert worden:

Der Test dazu prüft, ob der replizierte Effekte im Konfidenzintervall um den originalen Effekt liegt, ob diese also signifikant voneinander verschieden sind, so die Logik der Autoren. Lassen wir mal großzügig beiseite, dass dies kein Test über die *Differenz* der Effektgrößen ist, da das Konfidenzintervall um den zuerst beobachteten Effekt gelegt wird, nicht um die Differenz. Wesentlicher ist, dass dies ein schlechtes Maß für Replizierbarkeit ist, denn die originalen Effekte sind upward biased (sieht man in dem originalen paper auch), und vergessen wir den publication bias nicht (siehe density distribution der p-Werte im originalen paper). Anzunehmen, dass die originalen Effektgrößendie Populationseffektgrößen sind, ist wirklich eine heroische Annahme, gerade angesichts des positiven bias der originalen Effekte. Nebenbei: In einem offenen Brief von Klaus Fiedler auf Facebook dazu publiziert wurde, wird argumentiert, die Regression zur Mitte habe die im Schnitt geringeren Effektgrößen im OSF-Projekt produziert, könne diesen Effekt erklären. Dieses Argument mag teilweise stimmen, impliziert aber, dass die originalen Effekte extrem (also biased, weil selektiv publiziert wurde) waren, denn genau das ist ja das Charakteristikum dieses Regressionseffektes: Ergebnisse, die in einer ersten Messung extrem waren, "tendieren" in einer zweiten Messung zum Mittelwert. Die Tatsache, dass die originalen Effekte einen deutlichen positiven bias aufweisen, wird in Ihrer Stellungnahme ignoriert, bzw. gar nicht erst erwähnt.

Das Argument der 68%-Replizierbarkeit wird im übrigen auch vom Hauptautor in Antwort auf ihre Stellungnahme ganz offen in ähnlicher Weise kritisiert:

https://twitter.com/BrianNosek/status/639049414947024896

Kurzum: Sich genau diese Statistik als Unterstützung dafür aus der OSF-Studie herauszusuchen, um der Öffentlichkeit zu bedeuten, dass in der Psychologie im Grunde alles in Ordnung ist, sehe ich als "cherry picking" an.

2. Das Moderatoren-Argument ist letztlich unhaltbar, denn erstens wurde dies insbesondere im OSF-Projekt 3 intensiv getestet. Das Ergebnis ist u.a. hier zusammengefasst:

https://hardsci.wordpress.com/2015/09/02/moderator-interpretations-of-the-reproducibility-project/

Siehe u.a.:

> In Many Labs 1 and Many Labs 3 (which I reviewed here), different labs
> followed standardized replication protocols for a series of
> experiments. In principle, different experimenters, different lab
> settings, and different subject populations could have led to
> differences between lab sites. But in analyses of heterogeneity across
> sites, that was not the result. In ML1, some of the very large and
> obvious effects (like anchoring) varied a bit in just how large they
> were (from “kinda big” to “holy shit”).Across both projects,
> more modest effects were quite consistent. Nowhere was there evidence
> that interesting effects wink in and out of detectability for
> substantive reasons linked to sample or setting.

Länger findet man es hier

https://hardsci.wordpress.com/2015/03/12/an-open-review-of-many-labs-3-much-to-learn/

zusammengefasst:

> The authors put the interpretation so well that I’ll quote themat
> length here [emphasis added]:
> A common explanation for the challenges of replicating results across
> samples and settings is that there are many seen and unseen moderators
> that qualify the detectability of effects (Cesario, 2014). As such,
> when differences are observed across study administrations, it is easy
> to default to the assumption that it must be due to features differing
> between the samples and settings. Besides time of semester, we tested
> whether the site of data collection, and the order of administration
> during the study session moderated the effects. None of these had a
> substantial impact on any of the investigated effects. This
> observation is consistent with the first “Many Labs” study (Klein
> et al., 2014) and is the focus of the second (Klein et al., 2015). The
> present study provides further evidence against sample and setting
> differences being a default explanation for variation in
> replicability. That is not to deny that such variation occurs, just
> that direct evidence for a given effect is needed to demonstrate that
> it is a viable explanation.

Zweitens schreiben Sie In ihrer Stellungnahme:

> Solche Befunde zeigen vielmehr, dass psychologische Prozesse oft
> kontextabhängig sind und ihre Generalisierbarkeit weiter erforscht
> werden muss. Die Replikation einer amerikanischen Studie erbringt
> möglicherweise andere Ergebnisse, wenn diese in Deutschland oder in
> Italien durchgeführt wird (oder umgekehrt). In ähnlicher Weise
> können sich unterschiedliche Merkmale der Stichprobe
> (Geschlechteranteil, Alter, Bildungsstand, etc.) auf das Ergebnis
> auswirken. Diese Kontextabhängigkeit ist kein Zeichen von fehlender
> Replizierbarkeit, sondern vielmehr ein Zeichen für die Komplexität
> psychologischer Phänomene und Prozesse.

Nein, das zeigen diese neuen Befunde eben nicht, denn dies ist eine (Post-hoc-)Interpretation die durch die im neuen OSF-Projekt erhobenen Moderatoren *nicht* unterstützt wird, da diese Moderatorenanalysen gar nicht durchgeführt wurden. Die postulierte Kontextabhängigkeit wurde zudem im OSF-Projekt #3 nicht gefunden. Was man zwischen den labs als Variationsquelle fand war schlicht und einfach Stichprobenvariation, wie man sie nun mal in der Statistik erwarten muss. Ich sehe für Ihre Behauptung also gar keine empirische Basis, wie sie doch in einer sich empirisch nennenden Wissenschaft vorhanden sein sollte.

Was mir als abschließende Aussage in der Stellungnahme deutlich fehlt ist, dass die Psychologie (und gerade die Sozialpsychologie) in Zukunft keine selektiv publizierten und "underpowered studies" mehr akzeptieren sollte. Das hätte den Kern des Problems etwas besser getroffen.

03.09.2015

Prof. Dr. Detlef H. Rost

In der Stellungnahme des Vorstandes steht: "Die Untersuchung zeigt, dass die Psychologie der Stabilität ihrer Befunde einen hohen Stellenwert beimisst und damit ein Beispiel für andere Wissenschaften gibt. Insgesamt ist der ermittelte Replikationsanteil von 68% ein akzeptabler Wert."

Das ist ein Euphemismus. Die zusammenfassende Bewertung der beteiligten Forscher/Forscherinnen bezüglich der gelungenen Replizierbarkeit lag unter 40%. Die Realität sieht zudem anders aus: Replikationen finden sich in einschlägigen Journalen so gut wie nie: bei 100 Fachzeitschriften lag die Anzahl von Replikationen nur knapp über 1% (vgl. z. B. Makel, Plucker, & Hegarty, 2012; Plucker, 2014). Ist das ein Beleg für einen "hohen Stellenwert"?

Mögliche Gründe:

- von Psychologen/Psychologinnen werden kaum Replikationsstudien angelegt, weil der Karriere weniger dienlich;
- für Replikationsstudien werden in der Regel keine Drittmittel bewilligt;
- Replikationsstudien werden zwar durchgeführt, aber aus den anderen hier angeführten Gründen nicht zur Veröffentlichung eingereicht;
- Zeitschriftenherausgeber bzw. -herausgeberinnen und/oder Gutachter/Gutachterinnen akzeptieren Replikationstudien nur ungern oder gar nicht;
- es wird fälschlich angenommen, ein "statistisch signifikanter" oder "statistisch 'hoch' (sic!) signifikanter" oder "statistisch 'höchst' (sic!) signifikanter" Befund würde gleichzeitig eine Information über die Replizierbarkeit von Effekten bereit stellen.

Was in der Psychologie in weiten Teilen eingerissen ist, hatte Gustav A. Lienert vor rund 25 Jahren einmal beim Kaffeegespräch süffisant kommentiert: "publizieren, nicht replizieren". Hat sich seitdem etwas deutlich verändert?

Makel, M. C. Plucker, J., & Hegarty, C. B. (2012). Replications in psychology research: How often do they really occur? Perspectives on Psychological Science, 7, 537–542 (DOI: 10.1177/1745691612460688).

Plucker, J. A. (Ed). (2014). Replications in psychology [Special Section]. Psychology of Aesthetics, Creativity, and the Arts, 8, 2–29 (DOI: 10.1177/1745691612460688).

Prof. Dr. Detlef H. Rost

---

nach oben

Über diese Website

Copyright

Hosting und Betreuung:
Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID)

Die DGPs

Die Deutsche Gesellschaft für Psychologie e.V. (DGPs) ist eine Vereinigung der in Forschung und Lehre tätigen Psychologinnen und Psychologen. Die über 3000 Mitglieder der DGPs erforschen das Erleben und Verhalten des Menschen. Sie publizieren, lehren und beziehen Stellung in der Welt der Universitäten, in der Forschung, der Politik und im Alltag.
mehr

Service-Menü

- Startseite
- Kontakt
- Sitemap

- Impressum
- Rechtliche Hinweise