Andreas Diekmann*

Experimentelle Studien und Repräsentativität

Zur Klärung einiger Irrtümer und Missverständnisse

Alexander Lenger und Stephan Wolf (2018; im Folgenden LW) führen in ihrem Beitrag „Experimente in der Soziologie?" eine kritische Diskussion über Vorteile, Risiken, und Nebenwirkungen experimenteller Methoden in der Soziologie. Über drei Jahrzehnte hinweg hat die Verhaltensökonomie mit experimentellen Methoden in den Wirtschaftswissenschaften an Terrain gewonnen, während Experimente in der Soziologie eher selten waren. Die Soziologie ist, wenn sie standardisiert empirisch arbeitet, bis heute vorwiegend eine ‚Surveywissenschaft'. Die Autoren sprechen sich durchaus für Experimente aus, benennen aber auch einige Schattenseiten. Mit qualitativen Methoden ergründen sie die Motive von Probanden sozialwissenschaftlicher Experimente. So finden sie mithilfe qualitativer Interviews drei Typen von Versuchsteilnehmern: Den strategisch-nutzenmaximierenden Typ, den ‚Interessierten Forscher' und den ‚Kritischen Experten'. Die Ergänzung hochstandardisierter Laborexperimente durch qualitative Nachbefragungen kann durchaus zu neuen Einsichten führen und der Mix von verschiedenen Methoden wird kreativ eingesetzt. Bezüglich Forschungsstrategie und einiger Schlussfolgerungen liegen allerdings auch Irrtümer und Missverständnisse vor. So berechtigt die Kritik an Laborexperimenten ist, so muss man doch die Probleme an einigen Stellen präzisieren, um angemessene Lösungen zu finden.

1 Unterschiedliche Ziele von Experimenten

Zunächst einmal sollten die Ziele einer experimentellen Studie unterschieden werden. Dabei kann man drei Arten unterscheiden: (1) Explorative Experimente (LW sprechen von „Erkundungsexperimenten"), (2) Deskriptive Experimente, mit denen Parameter einer Population geschätzt werden, z.B. die Prozentanteile bestimmter Persönlichkeitstypen und (3) Experimente zur Überprüfung von Kausalhypothesen. Diese Unterscheidung wird in dem Beitrag von LW zu wenig beachtet. Insbesondere werden (2) und (3) nicht unterschieden. Das ist aber wichtig, denn nur für den Typ (2) von Experimenten sind überhaupt repräsentative Stichproben erforderlich. Typ (3) ist das klassische Experiment und besonders hier kommen auch die Vorteile der experimentellen Methode zum Tragen. Grund ist, dass die Zufallsaufteilung auf verschiedene Versuchsgruppen die *unbeobachtete Heterogenität* neutralisiert (dazu weiter unten) und somit (3) das ideale Design, den ‚Goldstandard', zum Test von Kausalhypothesen darstellt. Feldexperimente, Quasi-Experimente und multivariate

* Andreas Diekmann, Wissenschaftskolleg Berlin und ETH Zürich, Weinbergstrasse 109, 8092 Zürich, Schweiz, E-mail: andreas.diekmann@soz.gess.ethz.ch

SozW, 69 (1) 2018, 95 – 100 DOI: 10.5771/0038-6073-2018-1-95

https://doi.org/10.5771/0038-6073-2018-1-95 - Generiert durch IP 216.73.216.143, am 01.02.2026, 12:59:23. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Analysen von Surveystudien orientieren sich an der Logik des Experiments, können die ideale Situation aber nur annähern. Dafür bieten sie allerdings oft andere Vorteile im Vergleich mit experimentellen Designs. Von explorativen Experimenten wird erwartet, dass mit ihrer Hilfe neue Hypothesen generiert werden können. Sie sind also eher induktiv, während für den Test von Kausalhypothesen das hypothetisch-deduktive Modell der Theorieprüfung als Leitlinie fungiert. Oft gibt es auch Mischformen. Man prüft eine Hypothese; vermutet nach dem Test, dass diese nur unter bestimmten Zusatzbedingungen zutrifft und modifiziert diese anhand der Daten. Test und Modifikation nach der Datenanalyse sollten dabei separat erfolgen und explizit beschrieben werden, um HARKING (= „Hypothesizing after the results are known", Kerr 1998) zu vermeiden. Explorativ entwickelte Hypothesen sollten dann erneuten, unabhängigen Tests unterzogen werden.

2 Repräsentativität ist wichtig für deskriptive Studien

‚Repräsentative Studien' sind wichtig zur Schätzung deskriptiver Parameter einer definierten Population, aus der eine Stichprobe gezogen wurde. Typische Beispiele sind Mikrozensus oder Meinungs- und Wahlumfragen. Es interessiert nicht der Anteil der Wähler in der speziellen Stichprobe, die ihre Stimme Partei X geben werden, sondern der geschätzte Anteil in der Population, z.B. die wahlberechtigte Bevölkerung eines Bundeslandes. ‚Repräsentativ' ist zudem – aber das ist hier nicht zentral – ein umgangssprachlicher Begriff; im Prinzip gibt es keine Stichprobe, die in allen Merkmalen ein Abbild der Population darstellt (das wäre die Population selbst). In der Statistik wird entsprechend ‚Repräsentativität' durch das Verfahren der Ziehung von Zufallsstichproben aus einer wohlbestimmten Population definiert.

In Experimenten ist Repräsentativität, wir benutzen diesen Begriff der Einfachheit halber, entsprechend erforderlich, wenn von im Experiment gefundenen Parametern auf Populationen geschlossen wird. Ein Beispiel ist die Studie von Hermann, Thöni und Gächter (2008) über antisoziale Bestrafung. Die Experimente zur Kooperation wurden in zwölf Ländern bzw. Städten durchgeführt, u. a. in Athen, Riad, Minsk, Istanbul, Zürich, Boston u.a. Teilnehmer der Experimente waren aus Gründen der Vergleichbarkeit Studierende (‚undergraduates'). Nun wurden Zusammenhänge mit der Kultur eines Landes hergestellt, wobei sich insbesondere zeigte, dass ‚antisoziale Bestrafung' stärker in Ländern vorkommt, die ein schwaches Rechtssystem haben. Hier wären in der Tat repräsentative Stichproben aus der Bevölkerung aussagekräftiger. Zwar gibt die Studie Hinweise auf einen möglichen Zusammenhang; doch wäre es voreilig zu sagen, dass z.B. in Athen oder Griechenland antisoziale Bestrafung stärker ausgeprägt sei als in der Schweiz. Dagegen haben Bigoni et al. (2016) Stichproben aus der ‚Normalbevölkerung' in je zwei Städten im Norden und Süden Italiens erhoben, um die Hypothese über die Nord-Süd-Spaltung von Vertrauen und Kooperation mit dem *public good game* und dem Vertrau-

https://doi.org/10.5771/0038-6073-2018-1-95 - Generiert durch IP 216.73.216.143, am 01.02.2026, 12:59:23. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

ensspiel zu überprüfen. Sie finden die erwarteten Unterschiede, wobei in ihrer Studie auch der Schluss auf die zugrundeliegende Population gerechtfertigt ist. Wenn das Ziel eines Experiments ist, Aussagen über die ‚Kultur' eines Landes zu machen oder die Verteilung von Persönlichkeitsmerkmalen, wie z.B. prosoziales Verhalten, Eigennutz oder Wettbewerbsorientierung, die Befunde aber nur auf ‚convenience samples' mit Studenten basieren, ist dagegen die Kritik von LW ohne Zweifel berechtigt.

3 Repräsentativität ist keine Voraussetzung zum Test von Kausalhypothesen

Für viele andere Zwecke sozialwissenschaftlicher Forschung ist dagegen Repräsentativität von Stichproben keine notwendige Voraussetzung für die Gültigkeit von Forschungsresultaten. Es ist ein verbreitetes Missverständnis, dass Repräsentativität eine Voraussetzung standardisierter Sozialforschung sei. Das ist aber überhaupt nicht der Fall! Vor allem ist Repräsentativität keine Voraussetzung für den Test von Kausalhypothesen. Streng genommen ist es auch gar nicht möglich, Hypothesen mit räumlich oder zeitlich nicht eingeschränktem Geltungsbereich an einer repräsentativen Stichprobe zu überprüfen. Wenn von der Geltung der Hypothese, ‚X hat einen kausalen Effekt auf Y' ausgegangen wird, dann wird die Gültigkeit der Hypothese auch am darauf folgenden Tag, Monat oder Jahr erwartet. Eine repräsentative oder Zufallsstichprobe kann sich aber nur auf eine gegenwärtige Population beziehen. Eine Zufallsstichprobe ist dadurch definiert, dass jedes Element der Population eine angebbare Wahrscheinlichkeit größer als Null hat, in die Stichprobe aufgenommen zu werden. Elemente, die erst in der Zukunft auftreten, haben aber eine Wahrscheinlichkeit von Null, in die Stichprobe zu gelangen. Die Logik der Untersuchung von Kausalhypothesen im Experiment ist daher auch nicht an der Repräsentativität von Stichproben orientiert. Vielmehr stellen Experimente, wenn sie nicht nur explorativ angelegt sind, Testsituationen für Hypothesen dar. Sie sind gewissermaßen Filter, um falsche Hypothesen auszusondern. Galilei brauchte keine repräsentative Auswahl von Kugeln, um das Fallgesetz zu untersuchen. Bei Kugeln geht es allerdings nicht um ‚sinnhaftes' Handeln, Bedeutungen oder Situationsdefinitionen. Aber auch wenn letzteres im Bereich menschlichen Verhaltens gegeben ist, lassen sich sehr robuste, kausale Zusammenhänge auffinden. Ein Beispiel ist die bekannte Hypothese von Darley und Latane (1968) über Verantwortungsdiffusion, die zunächst im Labor mit studentischen Probanden überprüft werden konnte. Es zeigte sich, dass die individuelle Bereitschaft zur Hilfeleistung (Y) mit der Gruppengröße (X) zurückging. Zahlreiche weitere Studien außerhalb des Labors in unterschiedlichen Situationen und Gesellschaften konnten den Nachweis für die Gültigkeit der Hypothese erbringen und dabei weitere relevante Faktoren aufzeigen.

https://doi.org/10.5771/0038-6073-2018-1-95 - Generiert durch IP 216.73.216.143, am 01.02.2026, 12:59:23. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

4 Zufallsaufteilung versus Zufallsstichproben

Bei deduktiv-hypothesentestenden Experimenten ist nicht die Ziehung von Zufallsstichproben wichtig, sondern die *Zufallsaufteilung (Randomisierung)*. Im Unterschied zu Surveystudien wird im Experiment Varianz durch Design hergestellt, durch die Aufteilung in Versuchsgruppen oder Versuchsgruppe(n) und Kontrollgruppe. Erst die Zufallsaufteilung auf verschiedene Versuchsgruppen neutralisiert alle störenden, bekannten und unbekannten Drittfaktoren. Mit anderen Worten werden in korrekt mit Zufallsaufteilung durchgeführten Experimenten beobachtete und unbeobachtete Heterogenität kontrolliert, so dass nur noch die unterschiedlichen Versuchsgruppen als kausal erklärende Varianzquelle verbleiben. (Multivariate Analysen mit Surveydaten kontrollieren nur für beobachtete Heterogenität; Fixed-Effects-Regressionen von Paneldaten können unter gewissen Annahmen auch einen Teil der unbeobachteten Heterogenität kontrollieren.) Experimente sorgen dagegen quasi automatisch dafür, dass die unabhängige Variable (X) mit sämtlichen gemessenen und unbeobachteten Variablen (bis auf Zufallsfehler) unkorreliert ist, so dass diese ‚Drittvariablen' den ermittelten Effekt von X auf Y nicht verzerren können. Wenn also LW in ihrer qualitativen Studie drei Typen von Versuchspersonen identifizieren, nämlich den Homo oeconomicus, den ‚Interessierten Forscher' und den ‚Kritischen Experten', dann folgt daraus nicht zwangsläufig eine mangelnde Aussagekraft des Hypothesentests. Denn bei randomisierter Zuweisung verteilen sich die drei Verhaltenstypen gleichmäßig auf die Versuchsgruppen. Selbst wenn der Homo oeconomicus anders in Bezug auf Y agiert, die unterschiedlichen Verhaltenstypen also mit der abhängigen Variable korreliert sind, bliebe die Schätzung des Effekts von X auf Y unverzerrt. Allerdings gilt das nicht, wenn Interaktionseffekte mit dem Verhaltenstypus vorliegen. Das führt uns zu dem Problem externer Validität.

5 Das Problem externer Validität

Auch hier muss man etwas genauer hinsehen. In der Sprache von Hypothesentests gibt es eigentlich gar nicht die Unterscheidung von interner und externer Validität. Hypothesen sollen möglichst streng in verschiedenen Kontexten geprüft werden. Im Kern trifft die Kritik von LW aber durchaus einen wunden Punkt der Laborforschung; nicht nur in der Verhaltensökonomie, sondern generell in Psychologie und Sozialpsychologie und sogar in der Medizin. Die weiße Ratte, die ‚mus norwegicus albinos', ist heute der Student des entsprechenden Fachs, der in der Psychologie oft noch Kreditpunkte für die Teilnahme an Experimenten bekommt. Auch in der Medizin wird kritisiert, dass Medikamentenstudien vorwiegend mit jungen, gesunden Erwachsenen durchgeführt werden. Den Mangel an externer Validität kann man auch so interpretieren. Die ‚convenience samples', die WL kritisieren, untersuchen Hypothesen in sehr engen und immer wieder gleichen Kontexten. Wenn nun *Interaktionseffekte* vorliegen, d.h. der experimentelle Faktor X andere Wirkun-

https://doi.org/10.5771/0038-6073-2018-1-95 - Generiert durch IP 216.73.216.143, am 01.02.2026, 12:59:23. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

gen hervorruft oder keinen Effekt zeitigt, wenn zugleich weitere Merkmale U, V, W usw. vorliegen, kann man sich über Existenz und Stärke eines Effekts täuschen (Diekmann 2017: Kap. VIII). Weiterhin ist die Zeitspanne von Laborexperimenten kurz; oft werden überhöhte Effekte gefunden, die aber über kurz oder lang wieder nachlassen. (Ganz davon abgesehen, dass überhöhte Effekte größere Publikationschancen haben, so dass Replikationen oft schwächere Effekte berichten, Young/Ioannidis/Al-Ubaydli 2008). Aus all dem folgt, dass es sehr wichtig ist, dass Experimente in möglichst unterschiedlichen Kontexten und Kulturen repliziert werden; dass Triangulationen mit anderen Methoden (Quasi-Experimenten, Feldexperimenten, Surveystudien) stattfinden. Der erwähnte Effekt der Diffusion von Verantwortung wurde in einer Vielzahl von Experimenten im Labor und in natürlichen Umgebungen beobachtet (Überblick Fischer et al. 2011). Henrich et al. (2001) haben das Ultimatumspiel in westlichen und mehreren indigenen Kulturen verwendet und dadurch aufschlussreiche Einsichten über Fairnessverhalten gewonnen. Laborexperimente sind nur erste Stationen; ähnlich wie vor-klinische und klinische oder ‚in vitro' und ‚in vivo' Studien in der Medizin. In Laborexperimenten hat man die Kausalhypothese von Verhaltenseffekten einer Voreinstellung (‚Default-Effekt') gefunden. Ländervergleichende Studien von Organspenden zeigen, dass der Default ‚Spendebereitschaft', wenn nicht explizit Widerspruch eingelegt wird, die Bereitschaft zu Organspenden deutlich erhöht. Feldversuche mit Energieversorgern, die wir derzeit in der Schweiz durchführen, zeigen diese Effekte ganz eindeutig bei der Bestellung von ‚grünem' Strom selbst durch Geschäftskunden. Vom Labor, über ökonometrische Analysen mit Länder vergleichenden Daten bis hin zu Feldexperimenten wurde die Default-Hypothese in unterschiedlichen Kontexten mit unterschiedlichen Methoden geprüft. Kumulative Forschung, Replikationen und Triangulation können auch die soziologische Forschung voranbringen, sofern sie mehr als Pseudowissenschaft sein möchte und auf Erkenntnis abzielt. Dem Experiment kommt hierbei ein wichtiger, aber zweifellos nicht der einzige Stellenwert zu. Mit Sicherheit aber lässt sich sagen, dass das Potential experimenteller Methoden, insbesondere auch von Experimenten im sozialen Feld, in der Soziologie heute bei weitem nicht ausgeschöpft wird.

Literatur

Bigoni, Maria, Bortolotti, Stefania, Casari, Marco, Gambetta, Diego, Pancotto, Francesca, 2016. Amoral Familism, Social Capital, or Trust? The Behavioural Foundations of the Italian North-South Divide. The Economic Journal.

Darley, John M., Latané, Bibb, 1968. Bystander Intervention in Emergencies. Diffusion of Responsibility. Journal of Personality and Social Psychology 8: 377-383.

Diekmann, Andreas, 2017. Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen. 28. Aufl. Reinbek: Rowohlt.

https://doi.org/10.5771/0038-6073-2018-1-95 - Generiert durch IP 216.73.216.143, am 01.02.2026, 12:59:23. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Fischer, Peter, Krueger, Joachim I., Greitemeyer, Tobias, Vogrincic, Claudia, Kastenmüller, Andreas, Frey, Dieter, Heene, Moritz, Wicher, Magdalena, Kainbacher, Martina, 2011. The Bystander-Effect: A Meta-Analytic Review on Bystander Intervention in Dangerous and Non-Dangerous Emergencies. Psychological Bulletin 137: 517-537.

Henrich, Joseph, Boyd, Robert, Bowles, Samuel, Camerer, Colin, Fehr, Ernst, Gintis, Herbert, McElreath, Richard, 2001. In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. American Economic Review 91: 73-78.

Herrmann, Benedikt, Thöni, Christian, Gächter, Simon, 2008. Antisocial Punishment Across Societies. Science 319: 1362-1367.

Kerr, Norbert L., 1998. HARKing: Hypothesizing After the Results are Known. Personality and Social Psychology Review 2: 196-217.

Lenger, Alexander und Stephan Wolf (2018): Experimente in der Soziologie? Über die systematische Verzerrung von Experimentergebnissen aufgrund strategisch agierender studentischer Teilnehmertypen. Soziale Welt 1/18.

Young, Neal S., Ioannidis, John P. A., Al-Ubaydli, Omar, 2008. Why Current Publication Practices May Distort Science. PLoS Medicine 5: 1418-1422.

https://doi.org/10.5771/0038-6073-2018-1-95 - Generiert durch IP 216.73.216.143, am 01.02.2026, 12:59:23. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.