

Malte Jansen/Claudia Neuendorf/Aleksander Kocaj

Welche Potenziale bieten Sekundäranalysen für die Erhöhung von Forschungsqualität und Replizierbarkeit?

Zur Rolle von Multiversumsanalysen und integrativen Datenanalysen für die Bestimmung der Robustheit und Generalisierbarkeit von Forschungsbefunden

Zusammenfassung: Die Bereitstellung von qualitativ hochwertigen und gut dokumentierten Forschungsdaten ist eine zentrale Forderung der Open Science Bewegung. Dadurch entstehen auch in der empirischen Bildungsforschung zunehmend Möglichkeiten für Forscher*innen, Daten sekundäranalytisch auszuwerten (Hopfenbeck et al., 2018; Lenkeit, Chan, Hopfenbeck & Baird, 2015). Bisher gab es in der Diskussion um Open Science und Replizierbarkeit kaum Beiträge dazu, wie im Rahmen solcher Sekundäranalysen eine hohe Forschungsqualität sichergestellt werden kann und welche besonderen Potenziale Sekundäranalysen für die Abschätzung der Robustheit und Generalisierbarkeit von Forschungsbefunden bieten (Weston, Ritchie, Rohrer & Przybylski, 2019). In diesem Beitrag loten wir diese Potenziale aus und beschreiben, wie im Rahmen von Sekundäranalysen multiple Analysemethoden und multiple Datensätze genutzt werden können, um Robustheits- und Generalisierbarkeitsanalysen durchzuführen. Dafür stellen wir die Konzepte der *Multiversumsanalyse* (Steege, Tuerlinckx, Gelman & Vanpaemel, 2016) und der *integrativen Datenanalyse* (Curran & Hussong, 2009) vor. Wir beschreiben jeweils die Zielsetzung und die Analyseschritte und nennen Beispielstudien. Wir schließen mit der Empfehlung, auch bei Sekundärdatenanalysen Überlegungen zur Replizierbarkeit anzustellen und ihre Potenziale zur Untersuchung der Robustheit von Forschungsbefunden zu nutzen.

Schlagnworte: Robustheit, Generalisierbarkeit, Integrative Datenanalyse, Sekundäranalyse, Multiversumsanalyse

1. Sekundäranalysen in der Bildungsforschung

Die vergangene Dekade zeichnete sich durch einen Kulturwandel hin zum vermehrten Teilen von Forschungsdatensätzen in den Sozial- und Verhaltenswissenschaften und, diese Entwicklung flankierend, den kontinuierlichen Ausbau der Forschungsdateninfrastruktur aus (DGfE, GEBF & GFD, 2020; Nosek et al., 2015; Schönbrodt, Gollwitzer & Abele-Brehm, 2017). Dadurch entstanden zunehmend Möglichkeiten für Forscher*innen, Sekundäranalysen durchzuführen (Weston et al., 2019). Auch in der empirischen Bildungsforschung steigt der Anteil an Studien auf Grundlage von Sekundäranalysen kontinuierlich (Hopfenbeck et al., 2018; Lenkeit et al., 2015). Auch wenn Studien in der empirischen Bildungsforschung Bildungsprozesse über die gesamte Le-

benzspanne betrachten, konzentrieren wir uns in diesem Beitrag auf Sekundäranalysen in der Schulforschung. Unter dem Begriff Sekundäranalysen versteht man die Auswertung bereits vorliegender Datensätze durch Forscher*innen, die an der Erhebung und ersten Auswertung der Daten nicht beteiligt waren (Smith, 2008). Mit Sekundäranalyse kann sowohl die Beantwortung der gleichen Fragestellung mit einer neuen Analysemethode, als auch die Untersuchung völlig neuer Fragestellungen, die nicht bereits im Rahmen der Primärstudie untersucht wurden, gemeint sein (Jansen, Kocaj & Stanat, im Druck).

Im Mittelpunkt dieses Artikels soll das Potenzial von Sekundäranalysen zur Förderung der Replizierbarkeit, Robustheit und Generalisierbarkeit von Forschungsbefunden stehen. Replikation, Robustheit und Generalisierbarkeit sind Konzepte, die eng miteinander zusammenhängen, aber gegeneinander abgrenzbar sind (siehe Abb. 1). *Replizierbarkeit* bedeutet, dass sich Forschungsergebnisse auf Basis einer neuen Datenerhebung mit gleichem Studiendesign in vergleichbarer Richtung und Stärke zeigen (Makel & Plucker, 2014). Im Gegensatz dazu ist eine Studie *reproduzierbar*, wenn unabhängige Forscher*innen auf Basis derselben Datengrundlage und unter Nutzung der beschriebenen Analysemethode zu den gleichen Ergebnissen kommen wie die Primärforscher*innen (Artner et al., 2020). Insbesondere die Reproduzierbarkeit,

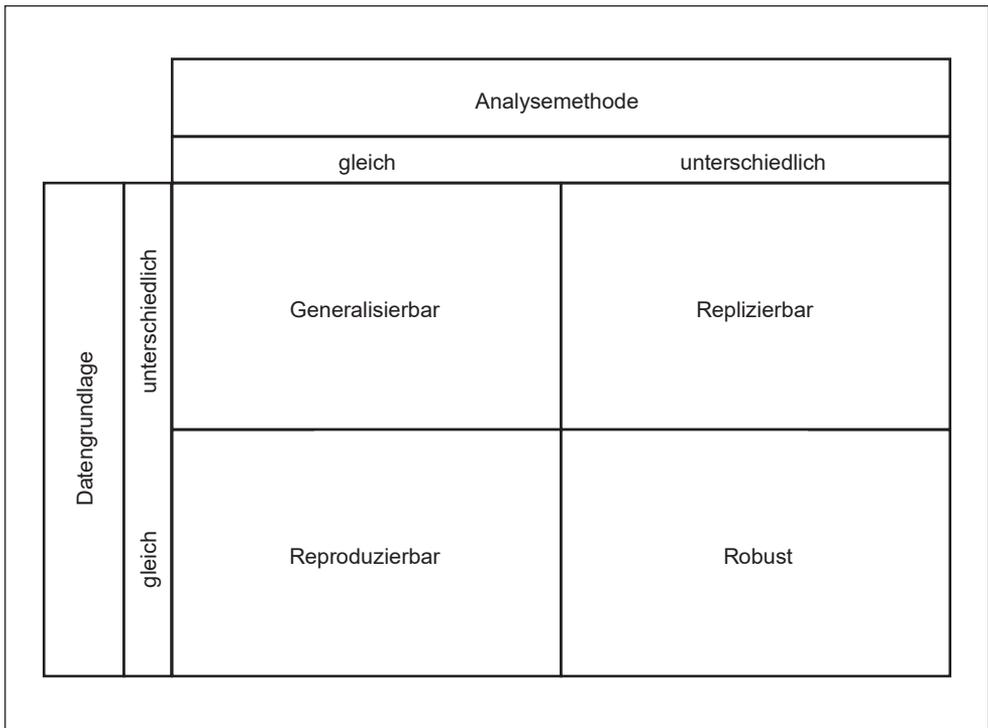


Abb. 1: Definition und Abgrenzung zentraler Begriffe

aber auch die Replizierbarkeit, können durch transparente Forschungsprozesse erhöht werden.

Die *Robustheit* von Forschungsergebnissen bezieht sich darauf, inwieweit unterschiedliche, plausible Analysestrategien bei der Auswertung eines Datensatzes zu ähnlichen Ergebnissen führen (Weston et al., 2019). Duncan, Engel, Claessens & Dowsett (2014) erweitern den Robustheitsbegriff, indem sie auch Analysen über mehrere Datensätze sowie Analysen über verschiedene Subgruppen einer Stichprobe als Robustheitschecks betrachten. Diese Sichtweise auf Robustheit ist eng verbunden mit Fragen der Generalisierbarkeit von Ergebnissen einer oder mehrerer Stichproben auf eine zugrundeliegende Population (Brennan, 2011; Duncan et al., 2014). Wir verwenden im Weiteren den Begriff Robustheit für die Konsistenz von Ergebnissen über verschiedene Analysestrategien innerhalb eines Datensatzes (und einer Stichprobe), aber den klassischen Begriff der *Generalisierbarkeit*, wenn die Konsistenz von Ergebnissen über mehrere Datensätze (und mehrere Stichproben) geprüft wird.

Die Robustheit und Generalisierbarkeit sowie Replizierbarkeit von Forschungsergebnissen eines Forschungsfeldes stehen in Zusammenhang. Sind die Ergebnisse einzelner Studien robust gegenüber unterschiedlichen Auswertungsmethoden und generalisierbar über verschiedene Stichproben, dann steigen damit auch die Chancen einer Replizierbarkeit der Ergebnisse. Im Gegensatz dazu steigt die Wahrscheinlichkeit falsch positiver und somit nicht replizierbarer Befunde, wenn sich ein Forschungsergebnis in einem Datensatz nur bei einer spezifischen Analysestrategie zeigt (*knife-edge specification*¹, siehe Abschnitt 2).

Studien, die sich mit der Replizierbarkeit von Forschungsergebnissen auseinandersetzen, beziehen sich bisher meist auf experimentelle Untersuchungen. Zu Analysen, die auf Daten großer Schulleistungstudien mit repräsentativen Stichproben basieren, werden Replikationen deutlich seltener durchgeführt, da zumindest der Anteil der Zufallsbefunde, der durch geringe Stichprobengrößen zustande kommt, vermutlich geringer ist als bei kleineren experimentellen Studien. Die Bedeutung von Robustheitschecks und Forschungstransparenz ist allerdings auch bei Analysen solcher Datensätze hoch (Weston et al., 2019). Sekundäranalysen ermöglichen es zwar, qualitativ hochwertige und verfügbare Datensätze mit repräsentativen Stichproben, komplexen Erhebungsdesigns und umfangreichen Variablen umfassend auszuwerten. Allerdings könnte die einfache Verfügbarkeit der Datensätze und die große Anzahl an Variablen und Schüler*innen zu einem stärker explorativ ausgerichteten Analyseverfahren verleiten, welches die Aussagekraft von Signifikanztests und damit die Chancen einer Replikation von Forschungsergebnissen reduziert (Scheel, Tiokhin, Isager & Lakens, 2020). Ins-

1 Den Begriff *knife-edge specification* könnte man wortwörtlich als Spezifikation eines Analysemodells übersetzen, bei dem die Analyseergebnisse ‚auf Messers Schneide‘ stehen. Darunter versteht man die Auswahl eines spezifischen Analysemodells mit einer bestimmten Konfiguration von Analyseentscheidungen, das zu signifikanten Ergebnissen führt, während bei der (überwiegenden) Mehrzahl anderer plausibler Auswertungsstrategien das angestrebte Signifikanzniveau verfehlt wird (Young & Holsteen, 2017).

besondere wenn viele Freiheitsgrade hinsichtlich der Datenaufbereitung und -auswertung vorliegen (was bei großen Schulleistungsstudien der Fall ist), können systematische und transparente Entscheidungen der Forscher*innen zu einer höheren Robustheit und damit zu einer besseren Replizierbarkeit zunächst auf explorative Weise generierter Ergebnisse beitragen (Weston et al., 2019). Hieraus erklären sich die besonderen Potenziale, die sich bei der Durchführung von Sekundäranalysen zur Absicherung der Robustheit und Generalisierbarkeit ihrer Ergebnisse bieten.

Wir beschäftigen uns daher in diesem Beitrag mit den zwei Dimensionen Robustheit und Generalisierbarkeit (siehe Abb. 1) – Robustheit von Ergebnissen über verschiedene Analysemethoden (siehe Abschnitt 3) und Generalisierbarkeit von Ergebnissen über verschiedene Datensätze hinweg (siehe Abschnitt 4). Der erste Aspekt scheint in der empirischen Bildungsforschung bisher weniger stark verbreitet zu sein als in anderen Fachdisziplinen wie z. B. der Bildungsökonomie; der zweite Aspekt wird bisher vor allem in Rahmen von systematischen Übersichtsarbeiten behandelt. Für jede Dimension wird eine Strategie vorgestellt, die die Robustheit bzw. Generalisierbarkeit von Sekundäranalysen verbessern und damit die Chancen für eine gelingende Replikation – zum Beispiel im Rahmen von neuen Primärerhebungen – erhöhen kann.

2. Rolle von Robustheitschecks in der empirischen Bildungsforschung

In nicht-experimentellen Studien beruhen Schlussfolgerungen zu Ergebnissen auf Annahmen über einen datengenerierenden Prozess, der in der Regel unbekannt ist. Dieser Prozess beschreibt die Verteilung und Zusammenhänge zwischen unabhängigen und abhängigen Variablen, die die beobachteten Daten generieren. In Analysemodellen nähert man sich diesem Prozess an, in dem man ein statistisches Modell wählt, das notwendigerweise mit bestimmten Annahmen über den datengenerierenden Prozess verknüpft ist. Ein Ziel von Robustheitschecks ist es, zu schätzen, wie sensitiv diese Ergebnisse gegenüber verschiedenen Annahmen über den datengenerierenden Prozess (z. B. über die Verteilung der analysierten Variablen, über relevante Kovariaten im Modell, über die Form der Zusammenhänge, also z. B. linear oder quadratisch, usw.) und wie verlässlich folglich die abgeleiteten Schlussfolgerungen sind. Diese Annahmen können im Rahmen von Robustheitschecks explizit gemacht und dadurch hinsichtlich ihrer Plausibilität diskutiert werden.

Es können drei Stufen von Robustheit unterschieden werden: die höchste Robustheit weisen Ergebnisse auf, die sich unabhängig von bestimmten Auswertungsstrategien zeigen (Young & Holsteen, 2017). Als weniger robust werden Ergebnisse klassifiziert, die sich nur zeigen, wenn eine bestimmte Entscheidung im Analyseprozess auf eine bestimmte Art getroffen wird (zum Beispiel für eine bestimmte Stichprobe, eine bestimmte Kodierung einer Variablen oder eine Art des Analysemodells). Die geringste Robustheit weisen schließlich Ergebnisse auf, die sich nur bei einer ganz bestimmten Kombination von mehreren Analyseentscheidungen zeigen, während sie bei der Mehrzahl alternativer Analyseentscheidungen nicht auftreten – sie sind also untypisch für den

Entscheidungsraum der möglichen Analysestrategien (*knife-edge specification*, Young & Holsteen, 2017; siehe auch Fußnote 1). Insbesondere das Aufspüren von kritischen Entscheidungen im Forschungsprozess, die a priori als gleichwertig (d. h. als jeweils gut begründet und geeignet, die Forschungsfrage zu beantworten) angenommen worden waren, hat das Potenzial, die Theorie- und Methodenentwicklung voranzutreiben, da sie den Grundstein für weitere Forschungsbemühungen legen können. Hierfür ist es allerdings wichtig, dass mit diesen Ergebnissen, die explorativ im Rahmen der Robustheitsprüfungen gewonnen werden, transparent umgegangen wird, also der volle Umfang der Robustheitsanalysen berichtet wird (Weston et al., 2019).

Robustheitschecks, oft auch als Sensitivitätsanalysen bezeichnet, sind in der Bildungsökonomie bereits weit verbreitet und werden auch in den Manuskriptrichtlinien führender Zeitschriften vorgeschrieben (Duncan et al., 2014). In der Ökonometrie geriet die Frage, wie sensitiv Ergebnisse gegenüber verschiedenen Modellspezifikationen sind, bereits in den 1980er Jahren in den Fokus der Aufmerksamkeit. Beispielsweise wurde die ‚Extreme Bounds Analyse‘ eingeführt, bei der Ergebnisse von Regressionsanalysen in Abhängigkeit von systematischen Variationen in der Auswahl von Kontrollvariablen überprüft werden (Leamer, 1983). In anderen Disziplinen im Feld der Bildungsforschung, wie zum Beispiel der Erziehungswissenschaft, pädagogischen Psychologie und Entwicklungspsychologie, sind Robustheitschecks unserem Eindruck nach allerdings in jüngerer Zeit weniger verbreitet. Meist wird lediglich eine geringe Anzahl unterschiedlicher Analysemodelle verwendet und nicht immer begründet, wie Forschende zu dieser Auswahl an Modellen gelangten (Duncan et al., 2014). Werden einzelne Analysestrategien variiert (zum Beispiel verschiedene Arten des Umgangs mit fehlenden Werten, Umgang mit Ausreißern) wird oft nur knapp berichtet, dass diese zu ähnlichen Ergebnissen führten – systematische und umfassende Prüfungen alternativer Analysestrategien fehlen zumeist.

In letzter Zeit wurden Systematiken für die Durchführung von Robustheitschecks innerhalb eines Datensatzes (Multiversumsanalyse und Spezifikationskurvenanalyse) und Generalisierbarkeitsprüfungen über verschiedene Datensätze hinweg (Integrative Datenanalyse) entwickelt, die das Potenzial besitzen, im Rahmen von Replikationsstudien oder bei Originalarbeiten mit Datensätzen aus großen Schulleistungsstudien zusätzliche Erkenntnisse zu generieren. Diese werden in den nächsten Abschnitten detaillierter vorgestellt.

3. Robustheit von Befunden über verschiedene Analysestrategien: Multiversumsanalyse und Spezifikationskurvenanalyse

3.1 Definition und Analyseschritte

Zwei eng verwandte kürzlich vorgeschlagene Konzepte für Robustheitsprüfungen stellen die Multiversumsanalyse (Multiverse Analysis; Steegen et al., 2016) und die Spezifikationskurvenanalyse (Specification Curve Analysis; Simonsohn, Simmons & Nelson, 2020) dar. Beiden Ansätzen gemeinsam ist, dass sie die Freiheitsgrade, die Forschende bei der Datenanalyse haben, sichtbar und deren Einfluss auf die Analyseergebnisse transparent machen sollen. Dies wird erreicht, indem zunächst Entscheidungen im Forschungsprozess identifiziert werden, die (a) das Potenzial haben, die Ergebnisse zu beeinflussen, und (b) gut begründbar auch anders hätten getroffen werden können („analytic decisions [...] which are both arbitrary and defensible“; Simonsohn et al., 2020, S. 1). Für diese einzelnen Entscheidungen wird dann eine Menge möglicher plausibler Vorgehensweisen bzw. Strategien definiert und schließlich werden alle möglichen und sinnvollen Kombinationen plausibler Alternativen für die unterschiedlichen Entscheidungen bezüglich ihres Einflusses auf die Analyseergebnisse geprüft. Dabei bedeutet plausibel und gut begründbar, dass entweder die Theorie, die der Forschungsfragestellung zugrunde liegt, keine Aussage zur Überlegenheit einer Entscheidungsalternative zulässt oder dass verschiedene state-of-the-art-Analysemethoden genutzt werden, die jeweils mit unterschiedlichen Stärken und Schwächen behaftet sind, indem sie zum Beispiel unterschiedliche, nicht überprüfbare Annahmen zum datengenerierenden Prozess voraussetzen. Typische solcher Entscheidungen im Rahmen von Sekundäranalysen betreffen zum Beispiel (a) die Bildung von Skalen, (b) die Auswahl einer Metrik für die Variablen, (c) den Fallausschluss, (d) die Auswahl von Kovariaten, (e) den Umgang mit fehlenden Werten und (f) die Auswahl eines Analysemodells mit unterschiedlichen zugrundeliegenden Annahmen (frequentistisch vs. bayesianisch; nicht-parametrisch vs. parametrisch; Regressionen vs. Strukturgleichungsmodelle etc.). Die Nutzung von state-of-the-art-Analysemethoden bedeutet auch, dass eindeutig unterlegene Analyseansätze (zum Beispiel keine Behandlung fehlender Werte, wenn substantielle Anteile fehlender Werte vorhanden und systematische Ausfallmuster wahrscheinlich sind) nicht als gleichwertig angesehen und somit auch explizit nicht berücksichtigt werden.

Aus diesem Vorgehen ergibt sich eine Vielzahl plausibler Ergebnisse, welche in ihrer Abhängigkeit von den einzelnen Entscheidungsparametern analysiert werden können. Der Multiversums- und der Spezifikationskurvenansatz verfolgen grundlegend die gleiche Idee der systematischen Kombination von Analysestrategien, sind nicht konzeptuell unterschiedlich und wurden parallel entwickelt. Es scheint gut möglich, dass zukünftig der Begriff der Multiversumsanalyse als Überbegriff, unter dem sich die Spezifikationskurvenanalyse einordnen lässt, verstanden wird (Rohrer, 2021). Trotzdem sollen im Folgenden die leicht unterschiedlichen Schwerpunkte dargestellt werden, die die Autor*innen bei der Entwicklung der beiden Ansätze gesetzt haben.

Das von Steegen et al. (2016) formulierte Ziel liegt darin, das Multiversum möglicher Datensätze für die Zukunft einzugrenzen, indem problematische Bereiche, in denen noch Lücken in der Theorie oder Methodik bestehen, die von Forschenden abverlangen, mehr oder weniger plausible Analyseentscheidungen zu treffen, identifiziert und deren Bedeutsamkeit für die Unsicherheit in den Ergebnissen abgeschätzt wird. Die Multiversumsanalyse kann also als Ausgangspunkt verstanden werden, den Möglichkeitsraum durch bessere und präzisere Theoriebildung und die Entwicklung dazu passender Messinstrumente einzugrenzen, sodass die Anzahl gleichwertiger, theoretisch vertretbarer Alternativen (die zu unterschiedlichen Resultaten führen) zugunsten einer theoretisch überlegenen Spezifikation verringert wird (Steegen et al., 2016).

In der Spezifikationskurvenanalyse hingegen besteht das vordergründige Ziel in der Beantwortung der Forschungshypothese und einer validen statistischen Inferenz. Entsprechend gibt es hier einen zusätzlichen Schritt, in dem die Spezifikationskurve, also das Multiversum plausibler Ergebnisse, noch einem statistischen Test unterzogen wird. Ähnlich wie bei einer Metaanalyse sollen dadurch ein mittlerer Effekt, eine Effektstreuung und Einflussfaktoren auf den Effekt (in diesem Fall Entscheidungen bei der Datenauswertung) identifiziert werden. Damit eignet sich die Spezifikationskurvenanalyse besonders für Re-analysen, in denen aufgedeckt werden soll, ob eine Originalstudie möglicherweise einen Zufallsbefund identifiziert hat, der durch einzelne, sehr spezifische Entscheidungen im sogenannten „Garden of Forking Paths“ (Gelman & Loken, 2013) zustande gekommen ist und stellt eine Möglichkeit dar, entsprechend der von Young und Holsteen (2017) vorgeschlagenen Klassifikation, die Robustheit eines Effekts abzuschätzen. Eine Auflistung der Schritte zur Durchführung einer Multiversums- bzw. Spezifikationskurvenanalyse findet sich in Abbildung 2, Bereich A.

Während Steegen und Kollegen (2016) in ihrem Artikel die Ergebnisse in einer tabellarischen Form darstellen, ist bei der Spezifikationskurvenanalyse (Simonsohn et al., 2020) die namensgebende Kurve die zentrale Ergebnisdarstellung (siehe Abb. 3). Die Spezifikationskurve eignet sich zur Darstellung einer großen Zahl an Entscheidungsalternativen. Im oberen Teil findet sich dabei die Kurve der Effekte in aufsteigender Größe sortiert. Jedem Punkt auf der Kurve ist eine Spezifikation im unteren Teil des Plots gegenübergestellt.

Simonsohn et al. (2011) schlagen über die visuelle Auswertung hinaus verschiedene Teststatistiken zur inferenzstatistischen Absicherung der Spezifikationskurve vor. Die erste beantwortet die Frage, ob sich der Median des Effekts signifikant von demjenigen unterscheidet, den man unter der Nullhypothese, dass es keinen Effekt gibt, erwarten würde. Der zweite Test bezieht sich auf die Frage, ob der Anteil signifikanter Ergebnisse an der Spezifikationskurve sich signifikant von dem unter der Nullhypothese erwarteten Anteil unterscheidet. Der dritte Test ist eine kontinuierliche Variante des zweiten Tests, in dem nicht der Anteil signifikanter Ergebnisse gezählt, sondern der mittlere Z-Wert ermittelt wird, der diesen Ergebnissen zugrunde liegt.

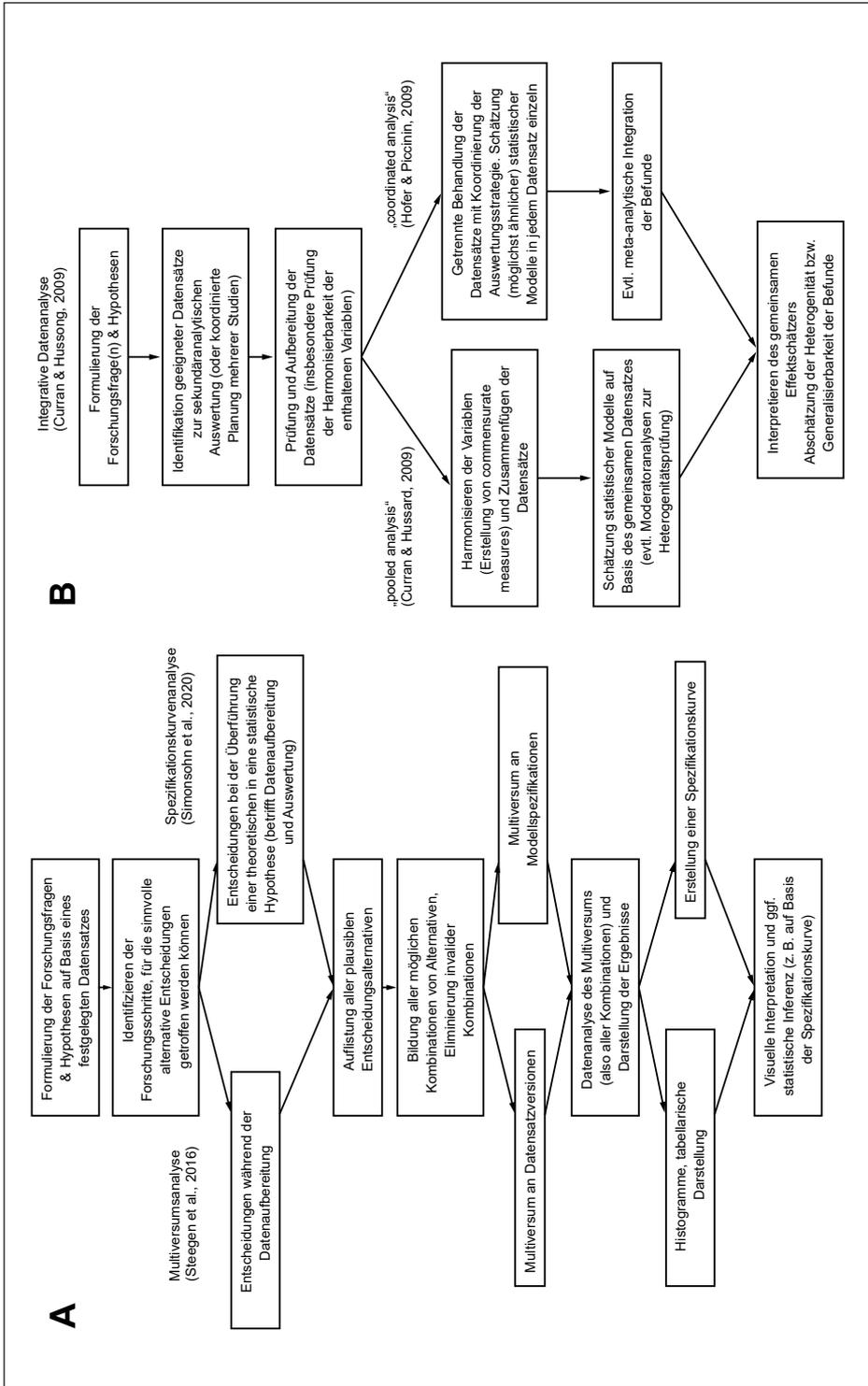


Abb. 2: Überblick über Schritte der Multiversums- bzw. Spezifikationskurvenanalyse (Bereich A) und der integrativen Datenanalyse (Bereich B)

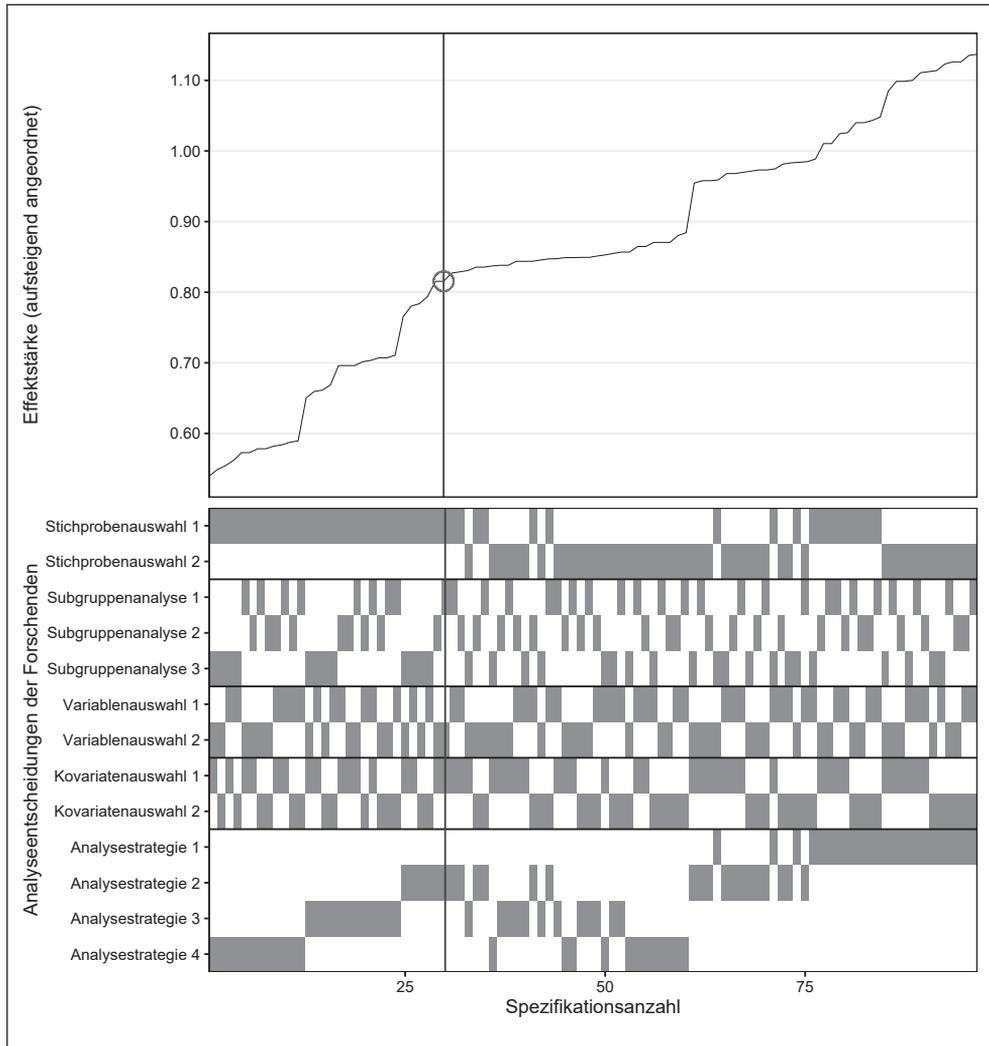


Abb. 3: Beispielhafte Darstellung einer Spezifikationskurve mit 96 verschiedenen Modellspezifikationen. Im unteren Bereich kann mithilfe der Einfärbungen die Kombination der Entscheidungsfaktoren für jeden Punkt der Kurve abgelesen werden. Auf der Kurve selbst sind die Effektstärker der jeweiligen Modelle abgetragen. Es könnte sich hierbei zum Beispiel um einen Gruppenunterschied (d), einen Regressionskoeffizienten, ein Odds-Ratio oder einen beliebigen anderen Schätzer handeln.

3.2 Beispiele

Es gibt bereits eine Reihe von Studien, welche die vorgeschlagene Methodik innerhalb der Bildungsforschung und in verwandten Disziplinen angewandt haben. So beschäftigte sich zum Beispiel eine Studie mit der Kompetenzentwicklung von Schüler*innen unterschiedlicher Ausgangsleistung in der Sekundarstufe (Neuendorf Jansen & Kuhl, 2020). Dabei wurden unter anderem die Operationalisierung der leistungsstarken Gruppe, die Skalierung der Leistungstests sowie der Umgang mit fehlenden Werten variiert.

Eine weitere Anwendung der Spezifikationskurvenanalyse im Bereich der Bildungsforschung demonstrierten Serra-Garcia, Hansen und Gneezy (2020), die eine Reanalyse einer Interventionsstudie von Miyake, Kost-Smith, Finkelstein, Pollock, Cohen und Ito (2010) durchführten. In dieser wurde der Effekt einer Selbstaffirmationsintervention auf den Leistungsunterschied zwischen weiblichen und männlichen Studierenden im Fach Physik untersucht. Serra-Garcia und Kollegen verorteten den Effekt der Primärstudie auf der Spezifikationskurve. Es war der fünfzehnthöchste der 1566 berechneten Effekte, während 77% der Spezifikationen kein signifikantes Ergebnis zeigten. Die Autor*innen folgerten, dass die Aussage der Primärstudie durch die vorliegenden Daten nicht gestützt werden könne, da die Ergebnisse der Originalstudie dem Robustheitstest nicht standgehalten hätten. Die Studie zeigt somit das Potenzial von Spezifikationskurvenanalysen, Spezifikationen zu identifizieren, die nicht typisch für den Entscheidungsraum plausibler Auswertungsstrategien sind.

Weitere Studien befassen sich zum Beispiel mit dem Effekt von Computerspielen auf die schulische Leistung, wobei u. a. verschiedene Operationalisierungen von Computerspielverhalten und von Leistung verglichen wurden (Gnams, Stasielowicz, Wolter & Appel, 2020), oder mit dem Zusammenhang der Qualitätsmerkmale des besuchten Kindergarten und der besuchten Grundschule innerhalb individueller Bildungsverläufe (Bailey, Jenkins & Alvarez-Vargas, 2020). In einer Studie wurde die Spezifikationskurve sogar im Rahmen einer Metaanalyse auf das analytische Vorgehen in der Metaanalyse selbst angewendet (Voracek, Kossmeier & Tran, 2019).

3.3 Zusammenfassung

Mit der Multiversums- und Spezifikationskurvenanalyse wurden zwei Methoden vorgestellt, die sich eignen, um zum einen die Robustheit von Effekten überprüfen zu können, aber auch um Faktoren zu identifizieren, die einen Einfluss auf die Effektgrößen haben. Dabei liegt der Fokus bei Steegen et al. (2016) stärker darauf, die Variabilität sichtbar zu machen und damit die Theoriebildung und Methodenweiterentwicklung anzuregen (wenn die Wahl der Entscheidungsalternativen theoriegeleitet erfolgt; siehe Rohrer, 2021), während der Fokus in der Arbeit von Simonsohn et al. (2020) zur Spezifikationskurvenanalyse stärker darauf liegt, über die verschiedenen möglichen Analyseentscheidungen hinweg valide Inferenz zu ermöglichen. Da beiden Ansätzen ver-

gleichbare methodische Vorgehensweisen zugrunde liegen, werden sie vielfach unter dem Begriff Multiversumsanalyse zusammengefasst (vgl. Rohrer, 2021). Eine Kritik, die zu Methoden der Multiversumsanalyse vorgebracht wurde, ist die vermeintliche Beliebbarkeit der unterschiedlichen Analysemethoden, die zur Anwendung kommen. Tatsächlich hängt der Nutzen einer Multiversumsanalyse von der Art und Weise ab, wie sie eingesetzt wird. Hier wird dafür plädiert, die Methodik im Rahmen eines kritischen Multiplismus (Patry, 2013; Shadish, 1986, 1993) zu sehen. Der kritische Multiplismus als Forschungsprogramm plädiert für den Einsatz einer Vielzahl an Operationalisierungen, Analysemethoden, Studien, Hypothesen und Theorien. Die Multiversumsanalyse ordnet sich in diesen historischen Ansatz ein, indem multiple Operationalisierungen und Auswertungsstrategien angewandt werden. Wittmann (1987) nimmt dabei einen typischen Vorbehalt gegenüber dessen Methoden vorweg: „Kritischer Multiplismus darf nicht mit Methodenanarchismus verwechselt werden“ (Wittmann, 1987, S. 216). Hetherington (1997, zit. nach Patry, 2013) betont, dass ein kritischer Multiplismus einem naiven (mindless) Multiplismus gegenüberzustellen ist, bei dem Ersterer theoriegeleitet, systematisch und rational sei, während Letzterer beliebig sei. Wie Rohrer (2021) argumentiert, sind Analyseentscheidungen bei genauerem Hinsehen selten wirklich austauschbar, arbiträr und gleichwertig. Multiversumsanalysen befreien also nicht von der Notwendigkeit, die einzelnen berücksichtigten Analyseentscheidungen konzeptuell und methodisch zu begründen und zu gewichten. Trotzdem können sie ein Werkzeug sein, um (a) die Forschungstransparenz zu erhöhen, (b) explorativ Entscheidungsfaktoren zu identifizieren, die sich auf die Analyseergebnisse auswirken, (c) *knife edge specifications* als solche kenntlich zu machen und insgesamt (d) die Idee der Robustheitsprüfung in der empirischen Bildungsforschung weiter zu verbreiten.

4. Generalisierbarkeit von Befunden über Datensätze: Die integrative Datenanalyse

4.1 Definition und Analyseschritte

Die Multiversumsanalyse und die Spezifikationskurvenanalyse sind mögliche Strategien, um den Einfluss ausgewählter Entscheidungen bei der Datenaufbereitung und Datenauswertung sichtbar zu machen. Diesen Entscheidungen ist bei sekundäranalytischen Forschungsprojekten aber noch ein wichtiger Schritt vorgelagert – die Suche nach geeigneten Datensätzen und die Entscheidung für einen oder mehrere Analysedatensätze. Während Multiversumsanalysen prüfen, inwiefern Ergebnisse über verschiedene Analysestrategien variieren, wird im Folgenden die Variation von Ergebnissen über verschiedene Datensätze thematisiert. Es handelt sich also bei der Betrachtung der Variation von Befunden über verschiedene Datensätze um eine qualitativ andere Aussage, für die typischerweise bisher nicht der Begriff der Robustheit, sondern der der Generalisierbarkeit bzw. der externen Validität herangezogen wurde (siehe aber Duncan et al., 2014, die explizit auf die Nutzung mehrerer Datensätze als Robustheitsprüfung hin-

weisen). Mit dem Datensatz variieren Studiencharakteristika wie die untersuchte Stichprobe, der Erhebungszeitpunkt und -kontext oder die eingesetzten Instrumente. Somit prüft eine Variation des Datensatzes eher Fragen der Generalisierbarkeit von Befunden einer Studie auf andere Populationen und Studienkontexte als Fragen der Robustheit und Qualität des konkreten analytischen Vorgehens einer (sekundäranalytischen) Studie. Fragen der Generalisierbarkeit von Einzelstudien werden in der empirischen Bildungsforschung schon lange berücksichtigt – etwa indem auf die Generalisierbarkeit und Repräsentativität der Stichprobe im Diskussionsteil empirischer Artikel eingegangen wird. Trotzdem wird zumeist pro Artikel nur ein Datensatz verwendet und die Aussagen zur Generalisierbarkeit werden nicht empirisch geprüft, obwohl dies aufgrund der zunehmenden Verfügbarkeit von Forschungsdaten zumindest in einigen Fällen direkt möglich wäre. Empirische Aussagen zur Generalisierbarkeit von Forschungsergebnissen ergeben sich stattdessen zumeist erst aus längerfristigen Forschungsprogrammen mit mehreren Studien bzw. durch systematische Übersichtsarbeiten und Metaanalysen (Curran, 2009).

Die integrative Datenanalyse (IDA) nimmt eine Perspektive ein, die zwischen der Beschränkung auf einen Datensatz bzw. auf eine Einzelstudie und der häufig notwendigerweise groben Aggregation von Effektstärken in systematischen Übersichtsarbeiten und Metaanalysen liegt (Curran & Hussong, 2009; Hussong, Curran & Bauer, 2013). Dazu werden für eine sekundäranalytische Studie zu einer bestimmten Fragestellung mehrere verfügbare Datensätze kombiniert, die aber, im Gegensatz zu einer typischen Metaanalyse, auf Rohdatenebene analysiert werden können (vgl. auch Riley, Lambert & Abo-Zaid, 2010; Roisman & IJzendoorn, 2018). Der Begriff der integrativen Datenanalyse wurde im Rahmen des Special Issues der Zeitschrift „Psychological Methods“ zum Thema ‚Multi-Study Methods for Building a Cumulative Psychological Science‘ von Curran und Hussong (2009) prominent gemacht, nachdem er kurz zuvor in einer Einzelarbeit bereits erwähnt wurde (Curran et al., 2008). Er wurde seitdem vor allem in der klinischen Psychologie und in der Entwicklungspsychologie verwendet (Bainter & Curran, 2015; Graham et al., 2017; Hussong et al., 2013) und bezeichnet bewusst kein einzelnes festgelegtes Analyseverfahren, sondern ist als Überbegriff für verschiedene, teilweise schon früher bereits vereinzelt angewandte Ansätze zu verstehen, bei denen Datensätze mehrerer Studien für sekundäranalytische Forschung kombiniert werden (Curran & Hussong, 2009).

Die Mehrzahl der Anwendungen kann nach Graham et al. (2017) in zwei Vorgehensweisen eingeteilt werden. Zum einen können einzelne Datensätze im ersten Schritt zu einem integrierten Datensatz zusammengefügt werden (*pooled analysis*). Dabei kann es sich zum Beispiel um Datensätze handeln, die von unterschiedlichen Forschungsgruppen zu einem ähnlichen Thema erhoben wurden oder um Teildatensätze einer multizentrischen Studie. Die Datensätze basieren also auf verschiedenen Stichproben und können nicht auf Personenebene verknüpft werden, sondern die Verknüpfung basiert auf einer Harmonisierung der erfassten Variablen, so dass im Idealfall die verschiedenen Stichproben, für die vergleichbare Variablen vorliegen, zu einer großen Stichprobe zusammengefasst werden können (Curran & Hussong, 2009). Dies ist analog zum Kon-

zept der Individual Participant Meta-Analysis, also einer Metaanalyse, die ebenfalls auf Rohdaten auf Personenebene basiert (Riley et al., 2010; Roisman & IJzendoorn, 2018; Verhage et al., 2020).

Für eine erfolgreiche Harmonisierung müssen Entscheidungen über die Vergleichbarkeit (und potenzielle ‚Gleichsetzbarkeit‘) von Variablen getroffen werden. So muss geprüft und entschieden werden, ob Instrumente hinreichend ähnlich (oder im Idealfall identisch) sind, um in eine gemeinsame Variable überführt werden zu können, und welche Transformationsschritte (zum Beispiel Zusammenfügen von Kategorien, Standardisierungen) dafür notwendig sind. Die mehrschrittige Erzeugung (zum Beispiel Itemauswahl, Wahl eines Messmodells) solcher „commensurate measures“ mit psychometrischen Techniken ist ein zentrales Forschungsthema der Arbeiten zur IDA (Bauer & Hussong, 2009; Curran, Georgeson, Bauer & Hussong, 2021; Davoudzadeh et al., 2020). Am Ende eines erfolgreichen Harmonisierungsprozesses steht ein Gesamtdatensatz mit größerer Stichprobe, mit dem man statistische Analysen für die Gesamtstichprobe durchführen und so gemeinsame Effektschätzer ableiten kann. Gleichzeitig können aber auch, ähnlich wie in einer Metaanalyse, Studiencharakteristika bei der Analyse einbezogen und kontrolliert werden oder die Effekte zunächst für die einzelnen Studien geschätzt und dann aggregiert werden (Curran & Hussong, 2009).

Neben der *pooled analysis*, die zunächst im Mittelpunkt der Überlegungen von Curran und Hussong (2009) stand, wurde die *coordinated analysis* als zweite Form der IDA vorgeschlagen. Dabei bleiben die Datensätze grundsätzlich separat und werden nicht zusammengefasst, aber mit dem gleichen Analyseprotokoll ausgewertet (Graham et al., 2017; Hofer & Piccinin, 2009). Dies ist vor allem dann sinnvoll, wenn die Variablen nicht direkt ineinander überführbar sind oder wenn die Datensätze aus datenschutzrechtlichen oder anderen forschungspraktischen Gründen nicht zwischen mehreren beteiligten Forschungsgruppen weitergegeben werden können. Im zweiten Fall beginnt die Koordination im Idealfall bereits in der Studienplanung. So wurde etwa die Möglichkeit diskutiert, große Längsschnittstudien so zu planen, dass diese koordiniert verwendbar sind (Hofer & Piccinin, 2009). Bei der *coordinated analysis* besteht also keine direkte Möglichkeit, Modelle an den Gesamtdatensatz anzupassen. Stattdessen wird zunächst eine Effektschätzung in jedem Datensatz separat vorgenommen. Diese können, je nach Forschungsfragen, Anzahl der Studien und Vergleichbarkeit der Studienkontexte, entweder qualitativ verglichen oder wiederum mit meta-analytischen Techniken aggregiert werden.

Wie lassen sich integrative Datenanalysen vor diesem Hintergrund von Metaanalysen abgrenzen? Die Metaanalyse bezeichnet im engeren Sinne statistische Aggregationstechniken zur quantitativen Zusammenfassung der Ergebnisse mehrerer Studien, die in der Regel aus einer systematischen Literaturrecherche hervorgehen. Integrative Datenanalysen greifen also teilweise auf meta-analytische Techniken zur Effektaggregation zurück (Becker, Kocaj, Dumont, Jansen & Lütke, im Druck; Jansen Lütke & Robitzsch, 2020), haben aber im Gegensatz zu Metaanalysen nicht den Anspruch, systematische Übersichtsarbeiten darzustellen und ein Forschungsfeld vollständig abzubilden. Anstatt einer systematischen Literaturrecherche übernehmen die Autor*innen da-

bei die Generierung der Evidenz im ersten Schritt selbst, indem sie mehrere verfügbare Datensätze auswerten und so Effektschätzer erhalten. Im zweiten Schritt kann dann eine Metaanalyse der Effektschätzer erfolgen. Auch wenn die Generalisierbarkeitsprüfung mit mehr Datensätzen umfassender wird, ist ein zentrales Kriterium für die Datensatzauswahl auch die Verfügbarkeit und es ist kein zwingendes Erfordernis der IDA, dass eine systematische Suche nach (ggf. auch nicht publizierten) Datensätzen zu einer Fragestellung erfolgt. Eine Zusammenfassung der Schritte der IDA findet sich in Abbildung 2, Bereich B.

4.2 Beispiele

Nachdem der Begriff der IDA zunächst eher in den Bereichen der klinischen Psychologie, Entwicklungspsychologie und Persönlichkeitspsychologie diskutiert wurde, sind in den letzten Jahren einige Studien im Bereich der Bildungsforschung erschienen, die sich explizit auf den Begriff beziehen. So befasste sich eine IDA mit der längsschnittlichen Stabilität des akademischen Selbstkonzepts (Jansen et al., 2020), eine weitere Anwendung befasste sich mit dem potenziell nichtlinearen Zusammenhang von Selbstkonzept und Leistung (Keller, Preckel & Brunner, 2020) und eine längsschnittliche Studie untersuchte schließlich Leistungskompositionseffekte auf Basis von fünf Datensätzen aus Deutschland (Becker et al., in Begutachtung). Die Generalisierbarkeit der Befunde war bei den Studien unterschiedlich stark ausgeprägt.

Bei diesen drei Anwendungen handelte es sich jeweils um koordinierte Analysen – die Datensätze der verschiedenen Studien wurden also nicht zusammengefasst, sondern es wurden „nur“ möglichst harmonisierte Analysemodelle in allen Datensätzen gerechnet und die Koeffizienten entweder qualitativ verglichen oder mit meta-analytischen Techniken zusammengefasst. Schließlich liegt eine Studie aus der Schulqualitätsforschung vor, die sich ebenfalls auf den IDA Begriff bezieht und zum Ziel hat, im Rahmen einer *pooled analysis* auf Basis von Schüler*innen- und Lehrkräfteeinschätzungen von Schulmerkmalen gemeinsame Indikatoren zu bilden (McGrath, Leighton, Ene, DiStefano & Monrad, 2020).

4.3 Zusammenfassung

Die IDA macht die Heterogenität von Effekten über verschiedene Datensätze sichtbar, die bei der Analyse eines einzelnen Datensatzes – auch wenn dieser groß und gut zur Beantwortung einer Forschungsfrage geeignet ist – verborgen bleiben würde. Darüber hinaus hat die IDA noch weitere Vorteile. So kann durch die IDA, wenn Datensätze im Sinne einer *pooled analysis* zusammengefasst werden, eine höhere Stichprobengröße und damit höhere Teststärke erreicht werden. Dies ist insbesondere dann nützlich, wenn bei der untersuchten Fragestellung eine Gruppe im Mittelpunkt steht, die innerhalb der Population klein ist, da durch das Zusammenfügen von Datensätzen eine ausrei-

chende Stichprobengröße erreicht werden kann (Curran & Hussong, 2009). Auch könnten durch die IDA Forschungsfragen beantwortet werden, die mit den jeweils einzelnen Datensätzen nicht zu beantworten wären, etwa wenn durch die Kombination mehrerer Längsschnitt- oder Kohortenstudien die insgesamt abdeckte Altersspannbreite erhöht wird (Graham et al., 2017). Schließlich können koordinierte IDAs auch zusätzliche Erkenntnispotenziale gegenüber klassischen Metaanalysen, die nicht auf Rohdaten basieren, bieten. Metaanalysen, in die häufig eine Vielzahl von Studien eingehen, müssen mit der Herausforderung einer großen Heterogenität der Primärstudien sowohl in Bezug auf ihre Qualität als auch auf das methodische Vorgehen umgehen, was die durchführenden Forscher*innen vor komplexe Entscheidungsprobleme in Bezug auf die Einschlusskriterien und die Harmonisierung von Effektstärkenschätzern stellt. Im Rahmen einer selbst durchgeführten IDA hat man die Auswertungsstrategie selbst „in der Hand“, kann also ein standardisiertes Protokoll einsetzen und so Variabilität, die durch die Auswertungsmethodik erzeugt wird, entweder vermeiden oder, durch eine systematische Variation im Sinne einer Multiversumsanalyse (s. o.), gezielt beeinflussen.

5. Gesamtfazit, Implikationen und Empfehlungen

Wie andere sozial- und verhaltenswissenschaftliche Disziplinen setzt sich auch die empirische Bildungsforschung zunehmend mit Open Science Praktiken zur Erhöhung der Transparenz von Forschungsprozessen und der Replizierbarkeit von Forschungsergebnissen auseinander. Die damit verbundene Verfügbarmachung von Forschungsdaten hat daher ebenfalls an Bedeutung gewonnen und wird in der Bildungsforschung umfassend diskutiert (DGfE, GEBF & GFD, 2020; Radisch, Stanat, Gräsel & Maaz, 2020). Ziel unseres Beitrags war es, die Potenziale von Sekundäranalysen bereits vorhandener Datensätze zur Prüfung von Robustheit und Generalisierbarkeit und damit einerseits zur potenziellen Erhöhung der Replizierbarkeit von Forschungsbefunden, andererseits aber auch zur Weiterentwicklung von Theorien (ausgehend von unerwarteten Befunden bei systematischen Robustheits- und Generalisierbarkeitsprüfungen) herauszuarbeiten.

Hierfür haben wir dargestellt, wie durch die systematische Nutzung (a) mehrerer Analysestrategien (mittels einer Multiversumsanalyse) und (b) mehrerer Datensätze (mittels einer integrativen Datenanalyse) abgeschätzt werden kann, wie robust und generalisierbar Forschungsbefunde sind.

Die hier vorgestellten Strategien sind konzeptuell nicht neu. Die Multiversumsanalyse und verwandte Ansätze stehen in der Tradition des kritischen Multiplismus (Shadish, 1986; Wittmann, 1987) und die IDA kann auch als Form der Rohdaten-Metaanalyse verstanden werden. Trotzdem ist es unseres Erachtens positiv zu bewerten, dass diese Techniken zur Robustheitsprüfung und Generalisierbarkeitsprüfung nun, nachdem die Sekundäranalyse seit ca. zwei Jahrzehnten durch die Verfügbarkeit der Datensätze der internationalen Schulleistungsstudien fester Teil der bildungswissenschaftlichen Forschung ist (Hopfenbeck et al., 2018; Lenkeit et al., 2015), auch in der empirischen Bildungsforschung vermehrt angewandt werden.

Trotz der Vorteile der genannten Verfahren sei nochmals darauf hingewiesen, dass die Nutzung mehrerer Datensätze und Analysestrategien nicht dazu führen sollte, diese als austauschbar zu betrachten oder die Auswahl der Strategien und Datensätze weniger gut zu begründen als im Fall einer Einzelstudie. Wie Rohrer (2021) argumentiert, sind Analysestrategien fast nie wirklich austauschbar und beliebig – man kann sie aber insofern als „arbitrary and defensible“ (Simonsohn et al., 2020, S. 1) sehen als dass in vielen Fällen vermutlich viele denkbare Analysestrategien zu einer Fragestellung existieren, die erfolgreich publiziert werden könnten. Es wird, auch aufgrund theoretischer Unklarheiten, nicht immer möglich sein, eine theoriebasiert eindeutige Entscheidung für eine Analysestrategie zu finden. Daher halten wir die vorgestellten Verfahren in jedem Fall für nützlich, um *knife edge specifications* zu identifizieren und den Entscheidungsraum aufzuzeigen (Young & Holsteen, 2017). Darüber hinaus können Faktoren identifiziert werden, die die Effektschätzung beeinflussen und damit vielleicht auch theoretisch und inhaltlich bedeutsame Mechanismen abbilden, die im Weiteren untersucht werden können. Auf diese Weise können Prüfungen der Generalisierbarkeit und Robustheit auch zu einer Theorieentwicklung führen, indem die Zutreffensbedingungen theoretischer Annahmen eingegrenzt werden (vgl. auch Gigerenzer, 1991; Greenwald, Pratkanis, Leippe & Baumgardner, 1986).

Durch die einfachere Verfügbarkeit von Forschungsdaten für Sekundäranalysen sowie die erhöhte Rechenkapazität moderner Hardware ist es heute leichter möglich, die Ansätze zur systematischen Variation von Datensätzen und Analysemodellen anzuwenden, auch wenn dies bedeutet, eine weit höhere Zahl statistischer Analysen durchzuführen als früher üblich (im Falle von Multiversums- oder Spezifikationskurvenanalysen wird die Anzahl der einzelnen gerechneten Analysemodelle oft im dreistelligen, manchmal auch im vierstelligen Bereich oder sogar noch höher liegen, je nachdem wie viele Entscheidungsschritte berücksichtigt werden; Muñoz & Young, 2018). Vor diesem Hintergrund möchten wir Bildungsforscher*innen ermuntern, diese Analysepotenziale noch stärker zu nutzen. Schließlich bleibt zu beobachten, ob die vorgestellten Methoden ihr Versprechen einhalten und das Forschungsfeld der empirischen Bildungsforschung voranbringen können. Dies wird davon abhängen, ob wir mit den Ergebnissen aus den Untersuchungen transparent umgehen und insbesondere auch nach Gründen für mangelnde Robustheit und Generalisierbarkeit suchen, um damit die Theorieentwicklung voranzutreiben.

Literatur

- Artner, R., Verliefe, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2020). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*. <https://doi.org/10.1037/met0000365>.
- Bailey, D. H., Jenkins, J. M., & Alvarez-Vargas, D. (2020). Complementarities between early educational intervention and later educational quality? A systematic review of the sustaining environments hypothesis. *Developmental Review*, *56*, 100910. <https://doi.org/10.1016/j.dr.2020.100910>.

- Bainter, S. A., & Curran, P. J. (2015). Advantages of integrative data analysis for developmental research. *Journal of Cognition and Development, 16*(1), 1–10. <https://doi.org/10.1080/15248372.2013.871721>.
- Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14*(2), 101–125. <https://doi.org/10.1037/a0015583>.
- Becker, M., Kocaj, A., Dumont, H., Jansen, M., & Lüdtke, O. (im Druck). Class-average achievement and individual achievement: Testing achievement composition and peer spillover effects using five German longitudinal studies. *Journal of Educational Psychology*.
- Curran, P. J. (2009). The seemingly quixotic pursuit of a cumulative psychological science: Introduction to the special issue. *Psychological Methods, 14*(2), 77–80. <https://doi.org/10.1037/a0015972>.
- Curran, P. J., Georgeson, A. R., Bauer, D. J., & Hussong, A. M. (2021). Psychometric models for scoring multiple reporter assessments: Applications to integrative data analysis in prevention science and beyond. *International Journal of Behavioral Development, 45*(1), 40–50. <https://doi.org/10.1177/0165025419896620>.
- Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*(2), 81–100. <https://doi.org/10.1037/a0015914>.
- Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology, 44*(2), 365–380. <https://doi.org/10.1037/0012-1649.44.2.365>.
- Davoudzadeh, P., Grimm, K. J., Widaman, K. F., Desmarais, S. L., Tueller, S., Rodgers, D., & Van Dorn, R. A. (2020). Estimation of latent variable scores with multiple group item response models: Implications for integrative data analysis. *Structural Equation Modeling, 27*(6), 931–941. <https://doi.org/10.1080/10705511.2020.1724113>.
- DGfE (= Deutschen Gesellschaft für Erziehungswissenschaft), GEBF (= Gesellschaft für Empirische Bildungsforschung) & GfD (= Gesellschaft für Fachdidaktik) (2020). *Empfehlungen zur Archivierung, Bereitstellung und Nachnutzung von Forschungsdaten im Kontext erziehungs- und bildungswissenschaftlicher sowie fachdidaktischer Forschung*. <https://www.fachdidaktik.org/wp-content/uploads/2020/03/PP-22-Gemeinsame-Stellungnahme-DGFE-GEBF-und-GfD-zum-Forschungsdatenmanagement-11-03-2020.pdf> [06.07.2021].
- Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50*(11), 2417–2425. <https://doi.org/10.1037/a0037996>.
- Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no „fishing expedition“ or „p-hacking“ and the research hypothesis was posited ahead of time*. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf [06.07.2021].
- Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review, 98*(2), 254–267. <https://doi.org/10.1037/0033-295X.98.2.254>.
- Gnamb, T., Stasielowicz, L., Wolter, I., & Appel, M. (2020). Do computer games jeopardize educational outcomes? A prospective study on gaming times and academic achievement. *Psychology of Popular Media, 9*(1), 69–82. <https://doi.org/10.1037/ppm0000204>.
- Graham, E. K., Gerstorff, D., Yoneda, T., Piccinin, A. M., Booth, T., Beam, C., Petkus, A. J., Rutsohn, J. P., Estabrook, R., Katz, M., Turiano, N., Lindenberger, U., Smith, J., Drewelies, J., Wagner, G., Pedersen, N., Allemand, M., Spiro, A., Deeg, D., ... Mroczek, D. (2017). *A coordinated analysis of big-five trait change across 16 longitudinal samples*. <https://doi.org/10.31234/osf.io/ryjpc>.

- Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review*, 93(2), 216–229. <https://doi.org/10.1037/0033-295X.93.2.216>.
- Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychological Methods*, 14(2), 150–164. <https://doi.org/10.1037/a0015566>.
- Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, 62(3), 333–353. <https://doi.org/10.1080/00313831.2016.1258726>.
- Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology*, 9(1), 61–89. <https://doi.org/10.1146/annurev-clinpsy-050212-185522>.
- Jansen, M., Kocaj, A., & Stanat, P. (im Druck). Sekundäranalysen. In T. Hascher, W. Helsper & T. Idel (Hrsg.), *Handbuch Schulforschung*. Springer VS.
- Jansen, M., Lüdtke, O., & Robitzsch, A. (2020). Disentangling different sources of stability and change in students' academic self-concepts: An integrative data analysis using the STARTS model. *Journal of Educational Psychology*, 112(8), 1614–1631. <https://doi.org/10.1037/edu0000448>.
- Keller, L., Preckel, F., & Brunner, M. (2020). Nonlinear relations between achievement and academic self-concepts in elementary and secondary school: An integrative data analysis across 13 countries. *Journal of Educational Psychology*. <https://doi.org/10.1037/edu0000533>.
- Leamer, E. E. (1983). Let's Take the Con Out of Econometrics. *The American Economic Review*, 73(1), 31–43. JSTOR.
- Lenkeit, J., Chan, J., Hopfenbeck, T. N., & Baird, J.-A. (2015). A review of the representation of PIRLS related research in scientific journals. *Educational Research Review*, 16, 102–115. <https://doi.org/10.1016/j.edurev.2015.10.002>.
- McGrath, K. V., Leighton, E. A., Ene, M., DiStefano, C., & Monrad, D. M. (2020). Using integrative data analysis to investigate school climate across multiple informants. *Educational and Psychological Measurement*, 80(4), 617–637. <https://doi.org/10.1177/0013164419885999>.
- Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, 330(6008), 1234–1237. <https://doi.org/10.1126/science.1195996>.
- Muñoz, J., & Young, C. (2018). We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology*, 48(1), 1–33. <https://doi.org/10.1177/0081175018777988>.
- Neuendorf, C., Jansen, M., & Kuhl, P. (2020). Competence development of high achievers within the highest track in German secondary school: Evidence for Matthew effects or compensation? *Learning and Individual Differences*, 77, 101816. <https://doi.org/10.1016/j.lindif.2019.101816>.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafeo, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>.
- Patry, J.-L. (2013). Beyond multiple methods: Critical multiplism on all levels. *International Journal of Multiple Research Approaches*, 7(1), 50–65. <https://doi.org/10.5172/mra.2013.7.1.50>.
- Radisch, F., Stanat, P., Gräsel, C., & Maaz, K. (2020). Kommentierung der gemeinsamen Stellungnahme zum Forschungsdatenmanagement von DGfE, GEBF und FGD aus Sicht der Kommission „Arbeitsgruppe Empirische Pädagogische Forschung“. *Erziehungswissenschaft*, 61(2), 29–38. <https://doi.org/10.3224/ezw.v31i2.04>.

- Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ*, *340*, c221. <https://doi.org/10.1136/bmj.c221>.
- Rohrer, J. (2021). Mülltiverse analysis. *The 100% CI (Blogbeitrag vom 07. 03. 2021)*. <http://www.the100.ci/2021/03/07/mulltiverse-analysis/> [15.06.2021].
- Roisman, G. I., & IJzendoorn, M. H. van. (2018). Meta-analysis and individual participant data synthesis in child development: Introduction to the special section. *Child Development*, *89*(6), 1939–1942. <https://doi.org/10.1111/cdev.13127>.
- Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *17*, 174569162096679. <https://doi.org/10.1177/1745691620966795>.
- Schönbrodt, F., Gollwitzer, M., & Abele-Brehm, A. (2017). Der Umgang mit Forschungsdaten im Fach Psychologie: Konkretisierung der DFG-Leitlinien. *Psychologische Rundschau*, *68*(1), 20–35. <https://doi.org/10.1026/0033-3042/a000341>.
- Serra-Garcia, M., Hansen, K. T., & Gneezy, U. (2020). Can short psychological interventions affect educational performance? Revisiting the effect of self-affirmation interventions. *Psychological Science*, *31*(7), 865–872. <https://doi.org/10.1177/0956797620923587>.
- Shadish, W. R. (1986). Planned critical multiplism: Some elaborations. *Behavioral Assessment*, *8*(1), 75–103.
- Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. *New Directions for Program Evaluation*, *1993*(60), 13–57. <https://doi.org/10.1002/ev.1660>.
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*, 1208–1214. <https://doi.org/10.1038/s41562-020-0912-z>.
- Smith, E. (2008). Pitfalls and promises: The use of secondary data analysis in educational research. *British Journal of Educational Studies*, *56*(3), 323–339. <https://doi.org/10.1111/j.1467-8527.2008.00405.x>.
- Stegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(5), 702–712. <https://doi.org/10.1177/1745691616658637>.
- Verhage, M. L., Schuengel, C., Duschinsky, R., van IJzendoorn, M. H., Fearon, R. M. P., Madigan, S., Roisman, G. I., Bakermans-Kranenburg, M. J., & Oosterman, M. (2020). The collaboration on attachment transmission synthesis (CATS): A move to the level of individual-participant-data meta-analysis. *Current Directions in Psychological Science*, *29*(2), 199–206. <https://doi.org/10.1177/0963721420904967>.
- Voracek, M., Kossmeier, M., & Tran, U. S. (2019). Which data to meta-analyze, and how?: A specification-curve and multiverse-analysis approach to meta-analysis. *Zeitschrift Für Psychologie*, *227*(1), 64–82. <https://doi.org/10.1027/2151-2604/a000357>.
- Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*, *2*(3), 214–227. <https://doi.org/10.1177/2515245919848684>.
- Wittmann, W. W. (1987). Grundlagen erfolgreicher Forschung in der Psychologie: Multimodale Diagnostik, Multiplismus, multivariate. *Diagnostica*, *33*(3), 209–226.
- Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, *46*(1), 3–40. <https://doi.org/10.1177/0049124115610347>.

Abstract: The sharing of well-documented research data is a central demand made by the Open Science movement to promote transparent and reproducible research. More and more research data is therefore made available, leading to increasing opportunities for researchers in educational research to conduct secondary data analyses (Hopfenbeck et al., 2018; Lenkeit et al., 2015). To date, it has rarely been discussed how high quality research can be ensured in the context of secondary analyses (Weston et al., 2019). Furthermore, the specific opportunities provided by secondary analyses to study the robustness and generalizability of research findings have not been a focus of the field. In this paper, we explore these potentials and describe how the usage of multiple analysis methods and multiple data sets allows for new strategies for robustness analysis. In particular, we introduce the concepts of *multiverse analysis* (Steegeen et al., 2016) and *integrative data analysis* (Curran & Hussong, 2009). For both we describe the central goals and steps of analysis and present example studies. We conclude with a recommendation to consider questions of replicability more carefully when conducting secondary data analyses and to use their potential to investigate the robustness of research findings.

Keywords: Robustness, Generalizability, Integrative Data Analysis, Secondary Analysis, Multiverse Analysis

Anschrift der Autor:innen

Dr. Malte Jansen, Institut zur Qualitätsentwicklung im Bildungswesen –
Wissenschaftliche Einrichtung der Länder an der Humboldt-Universität zu Berlin e. V.,
Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: Malte.Jansen@IQB.HU-Berlin.de

Dr. Aleksander Kocaj, Institut zur Qualitätsentwicklung im Bildungswesen –
Wissenschaftliche Einrichtung der Länder an der Humboldt-Universität zu Berlin e. V.,
Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: Aleksander.Kocaj@IQB.HU-Berlin.de

Dipl.-Psych. Claudia Neuendorf, Institut zur Qualitätsentwicklung im Bildungswesen –
Wissenschaftliche Einrichtung der Länder an der Humboldt-Universität zu Berlin e. V.,
Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: Claudia.Neuendorf@IQB.HU-Berlin.de