**Forschen mit archivierten Datensätzen: Wie ein Datenarchiv die
Bildungsforschung vorantreibt**

Mario Eich, Fitore Morina, Christian Fischer

University of Tübingen

Abstract: Das Forschungsdatenzentrum (FDZ) am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) archiviert die Datensätze von großen nationalen und internationalen Schulleistungsstudien sowie kleinerer lokaler Studien und stellt sie der Wissenschaft zur Verfügung. Für die Bildungsforschung liefert die Datenbank damit wahre Datens(ch)ätze, um neue Erkenntnisse zu generieren. Auch für Studierende und Promovierende werden so ambitionierte Forschungsprojekte greifbar.

Sollte es eine Liste aller Forschungsprojekte geben, die eigentlich undenkbar sind, würden die folgenden Studien auf jeden Fall dazu gehören: Eine Bachelorarbeit, in der Daten von über 900 Schülerinnen und Schülern analysiert wurden, eine Masterarbeit, die eine Längsschnittstudie mit drei Messzeitpunkten und 4.000 Schülerinnen und Schüler unter die Lupe nimmt und die Untersuchung der Lernausgangslage von über 4.000 Schülerinnen und Schülern mit Förderbedarf. Was diese Forschungsprojekte gemein haben, ist die Tatsache, dass sie die Menge an Daten in ihrer kurzen Entstehungszeit nicht selbst erheben können. Um solche Projekte dennoch zu realisieren, kann die Bildungsforschung auf das Forschungsdatenzentrum (FDZ) am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) zurückgreifen.

Seit der Gründung 2007 als Abteilung des IQB baut sich das Forschungsdatenzentrum stetig weiter aus. Seine Kernaufgabe ist die Aufbereitung und Archivierung von Daten aus nationalen und internationalen Schulleistungsstudien. Wissenschaftlerinnen und Wissenschaftler können diese Daten jedoch nicht nur für die eigene Forschung nutzen und auswerten, sondern gleichzeitig auch eigene Daten am FDZ archivieren, um weitere

Forschungsprojekte zu ermöglichen. Die Mitarbeiterinnen und Mitarbeiter am FDZ verfassen in der Regel ihre jeweiligen und thematisch verschiedenen Qualifikationsarbeiten ebenfalls mit FDZ-Daten, sind also selbst wissenschaftlich tätig. Daher haben sie nicht nur Expertise in der Aufbereitung und Archivierung der Daten, sondern auch in deren Nutzung für konkrete Forschungsprojekte.

Welche Daten bietet das Forschungsdatenzentrum?

Insgesamt verfügt das Forschungsdatenzentrum aktuell über 64 Studien. Neben den groß angelegten und bekannten Schulleistungstudien wie PISA, TIMSS oder dem IQB-Bildungstrend verfügt das FDZ auch über kleinere Studien. Diese konzentrieren sich auf thematisch spezifische Inhalte, wie beispielsweise die Sprachförderung im Sachunterricht von Grundschulen oder nehmen eine lokale Region Deutschlands genauer in den Blick. Der Datensatz zum Forschungsprojekt „Kompetenzerwerb und Lernvoraussetzungen (KuL)“ beschäftigt sich zum Beispiel mit der Interaktion zwischen Lehrpersonen und frisch eingeschulten Grundschulkindern und umfasst Angaben von etwa 1.000 Schülerinnen und Schülern, 840 Eltern und 69 Lehrkräften aus 39 Schulen in Nordrhein-Westfalen.

Alle Daten – sowohl aus den großen Erhebungen des Bundes und der Länder sowie entsprechende Datenspenden – werden nach Erhalt am Forschungsdatenzentrum geprüft, aufbereitet und je nach Sensibilität der Daten einer von drei Zugangskategorien zugeteilt. Je stärker die Daten anonymisiert werden, desto stärker wird zumeist auch das Analysepotenzial der Datensätze eingeschränkt. Der Kompromiss, der hierbei getroffen wird, um einerseits die Teilnehmerinnen und Teilnehmer der Studien zu schützen und gleichzeitig das Potential für weitere Analysen nicht einzuschränken, wird daher meist über eine Reglementierung für den Zugriff auf die Daten oder ein ausführlicheres Antragsverfahren umgesetzt, wie Abbildung 1 systematisch zeigt:

Abbildung 1: Grafik in Anlehnung an Statistisches Bundesamt
(www.forschungsdatenzentrum.de)

Wie ist ein Zugang zu den verschiedenen Datensätzen möglich?

Die Frage, wie interessierte Forscherinnen und Forscher Zugang zu den Datens(ch)ätzen erhalten, lässt sich anhand von drei Kategorien beantworten, denen die Daten zugeordnet sind und die sich je in der Art der Antragsstellung unterscheiden: (1) Campus Files, (2) Scientific Use Files und (3) Daten mit Fernrechenzugang.

Campus Files (CFs) sind teilsynthetisch generierte Daten, die zwar die Struktur von ‚echten‘ Daten abbilden, jedoch künstlich erzeugt wurden. Dafür wird die Struktur und Verteilung eines empirischen Datensatzes verwendet und mit Zufallszahlen repliziert. Auf diese Weise enthalten Campus Files keine ‚echten‘ Werte von Personen mehr, die Analyse ist jedoch weiterhin wie mit ‚echten‘ Daten möglich. Für den Einsatz in der Lehre bieten Campus Files daher einiges an Potenzial und kaum Einschränkungen. Insbesondere der Vorteil der datenschutzrechtlichen Unbedenklichkeit liegt auf der Hand. Da die Werte künstlich generiert wurden, eignen sich diese Datensätze jedoch nicht für die Beantwortung von Forschungsfragen. Um mit Campus Files zu arbeiten, genügt ein verkürztes Antragsverfahren,

das aus einer Online-Registrierung mit einer gültigen universitären E-Mail-Adresse besteht, die zum Download der Daten berechtigt.

Scientific Use Files (SUFs) bilden den Hauptbestandteil der Datensammlung. Um Zugriff auf diese Datensätze zu erhalten, ist ein regulärer Antrag einzureichen, der die Daten aller am (Forschungs-)Projekt beteiligten Personen sowie die Absichten für die Verwendung der Daten erfasst. Auch ein Datennutzungsvertrag muss abgeschlossen werden, um die sachgemäße Verwendung der Daten sicherzustellen. Anschließend erhalten die Nutzerinnen und Nutzer den Datensatz (in typischen Formaten wie z.B. SPSS) und weitere Dokumentationsmaterialien.

Zu den Daten mit Fernrechenzugang (IZAs) gehören alle sensiblen Datensätze, die lediglich über einen Fernrechenzugang analysiert werden können. Dieses Angebot ermöglicht die vertiefende Analyse von Daten, die aus Gründen des Datenschutzes nicht weitergegeben werden dürfen und oft detaillierte Inhalte enthalten. Wie der Namensbestandteil *„Fernrechenzugang“* bereits aussagt, werden die Datensätze in dieser Kategorie nicht zum Download zur Verfügung gestellt, sondern können lediglich syntaxbasiert analysiert werden. In der Praxis bedeutet dies, dass der Code zur Analyse und Aggregation der Daten über einen Zugang verschickt wird. Dadurch sind für die Arbeit mit Daten aus dieser Kategorie auch Erfahrungen im Schreiben von Analysecodes erforderlich und Programme wie „R“ oder „Stata“ notwendig. Anschließend wird geprüft, ob der Output – insbesondere welcher veröffentlicht werden soll – der beantragten und vertraglich festgelegten Fragestellung entspricht, um eine Verbreitung der sensiblen Daten zu verhindern.

Wie profitiert die Bildungsforschung von der Archivierung der Forschungsdaten?

Der Ruf nach „OpenScience“ seitens der Wissenschaftsgemeinschaft wird immer lauter. Viele Institute schließen sich zusammen, um Daten zugänglich, Forschung reproduzierbar und für die Scientific Community nutzbar zu machen. Diesem Gedanken folgt auch das Forschungsdatenzentrum seit mittlerweile über 20 Jahren. Die Datenbasis ermöglicht

breite Sekundärdatenanalysen und Replikationsstudien und trägt dabei zur Nutzung des bestehenden Datenpotenzials und zur Qualitätssicherung von Forschung bei.

Im Bildungsbereich profitieren auch Schülerinnen und Schüler und Lehrkräfte von der Verwendung bereits bestehender Daten. Schulen können entlastet werden, indem die Beanspruchung von Unterrichtszeit und der Organisationsaufwand für Studien möglichst geringgehalten und gleichzeitig der größtmögliche Nutzen aus den Daten durch breite Zugangsmöglichkeiten erzielt wird. Diese Entlastung kann dann wiederum den Feldzugang für diejenigen Studien erleichtern, die Datensätze zu Themenfeldern erheben, zu denen noch keine Daten vorliegen.

Einen weiteren Vorteil, den das Angebot des Forschungsdaten zentrums bietet, ist die Kombination von mehreren Datensätzen. Dadurch lassen sich robuste Studien durchführen, die durch eine größere Stichprobe weniger sensibel für Messfehler sind und dadurch die Generalisierbarkeit der Ergebnisse erhöhen. Wenn mehrere Datensätze zur Verfügung stehen, lassen sich häufig auch mehr Hypothesen erforschen als es mit lediglich einem Datensatz möglich ist. Beispielsweise konnte durch die Verwendung der am Forschungsdatenzentrum archivierten PISA-Daten ein Vergleich der fächer- und länderspezifischen Schulleistungen im Zuge der G8-Reform erfolgen, ohne dass hierfür eigene Daten erhoben werden mussten.

Durch sein großes Datenarchiv und den meist leichten Zugang öffnet das Forschungsdatenzentrum am IQB der Bildungsforschung Türen zu Forschungsmaterial, die einen wichtigen Beitrag zu „OpenScience“ darstellen. Das breite Angebot der Studien ermöglicht es unter anderem auch Studierenden und Promovierenden mit echten und großen Datensätzen zu arbeiten und auf diese Weise neue Erkenntnisse für die Bildungsforschung zu generieren – wie wäre es zum Beispiel mit einer Masterarbeit, die eine Längsschnittstudie zu Grundschülerinnen und Grundschülern mit Förderbedarf genauer untersucht? Dadurch könnte die Fragestellung in jedem Fall auch aus der Liste mit Forschungsprojekten gestrichen werden, die eigentlich undenkbar sind.

Weiterführende Literatur

Institut zur Qualitätsentwicklung im Bildungswesen (IQB, 2021). *Forschungsdatenzentrum*

am IQB, Retrieved from <https://www.iqb.hu-berlin.de/fdz>

Verbund Forschungsdaten Bildung (VerbundFDB, 2021), *Internet* Retrieved

from <https://www.forschungsdaten-bildung.de/> (abgerufen am 24.06.2021)

Forschungsdatenzentrum (2021), *Internet*. Retrieved

from www.forschungsdatenzentrum.de (abgerufen am 24.06.2021)

Rüdiger, C., Jansen, M., & Rjosk, C. (2021). Paul ist nicht so gut in Deutsch “.

Geschlechtsdifferenzielle Benotung im Fach Deutsch – eine Sekundäranalyse der Daten des IQB-Bildungstrends 2015. *Psychologie in Erziehung und Unterricht*, 68,.

<https://doi.org/10.2378/peu2021.art08d>

Kocaj, A., Jansen, M., Kuhl, P., & Stanat, P. (2020). Zusammenhänge der

Klassenkomposition an Förderschulen und allgemeinen Schulen mit schulischen

Kompetenzen, akademischem Selbstkonzept und Interesse. In C. Gresch, P. Kuhl, M.

Grosche, C. Sälzer, & P. Stanat (Hrsg.), *Schüler* innen mit sonderpädagogischem*

Förderbedarf in Schulleistungserhebungen (pp. 213-262). Springer VS, Wiesbaden.

https://doi.org/10.1007/978-3-658-27608-9_8

Mario Eich, B.A., studiert Empirische Bildungsforschung und Pädagogische Psychologie an der Eberhard-Karls-Universität Tübingen.

Fitore Morina, M.Sc., ist wissenschaftliche Mitarbeiterin am Hector-Institut für Empirische Bildungsforschung an der Eberhard-Karls-Universität Tübingen.

Christian Fischer, Ph.D., ist Tenure-Track-Professor für Educational Effectiveness am Hector-Institut für Empirische Bildungsforschung der Eberhard-Karls-Universität Tübingen.