

## Ist eine kleine Schulklassengrösse wirklich wichtig? Eine empirische Replikation für die Schweiz zum Zusammenhang von Klassengrösse und Schulleistung in der Sekundarstufe I mit ÜGK-Daten

**Rolf Becker**, Universität Bern

*Es wird für die Schweiz untersucht, ob die Grösse der Schulklasse einen kausalen Einfluss auf die Mathematikleistungen hat. Hierfür werden Daten von ÜGK aus dem Jahre 2016, einer nationalen Schulleistungsstudie zu den Grundkompetenzen von Schülerinnen und Schülern in der letzten Klassenstufe während der Pflichtschulzeit im Fach Mathematik, herangezogen. Zunächst legen bivariate Analysen nahe, dass grössere statt kleinerer Schulklassen mit höheren Testwerten in Mathematik einhergehen. Anhand multivariater Schätzverfahren, die geeignet sind, kausale Effekte der Schulklassengrösse aufzudecken, wird bei Kontrolle konkurrierender Einflüsse dieser Befund validiert. Es ist für die Sekundarstufe I nicht notwendig, die Klassengrösse mit dem Ziel zu reduzieren, die schulischen Leistungen anzuheben.*

### 1. Einleitung

Moderne Gesellschaften versuchen über bildungspolitische Massnahmen – etwa über die Festlegung von maximalen Schulklassengrösse – die schulischen Leistungen nachwachsender Generationen zu optimieren. So werden etwa leistungsschwächere Schülerinnen und Schüler im Unterschied zu leistungsstärkeren Schulkindern in kleineren Klassen unterrichtet. Zugleich befassen sich seit über 100 Jahren Studien aus der Bildungsforschung mit der Frage, ob – unabhängig davon, wie die Schulklassengrösse kategorisiert werden – kleinere Schulklassen tatsächlich zu besseren Bildungsergebnissen führen als grössere Schulklassen (Chingos, 2013; Cho et al., 2012; Finn et al., 2003; Glass & Smith, 1979; Hanushek, 1986, 1999; Krueger, 2003; Li & Konstantopoulos, 2016; Slavin, 1990; Smith & Glass, 1980; Sohn, 2016). Implizit wird stets die Annahme vertreten, dass (insbesondere sozial benachteiligte) Schulkinder in kleinen Klassen bessere schulische Leistungen erbringen als in grossen Klassen (z. B. Akerhielm, 1995; Konstantopoulos, 2008; Krueger, 1999). Die empirische Forschung, welche diese Annahme über den Zusammenhang von Klassengrösse und Schulleistung überprüft hat, kam bislang zu inkonsistenten Befunden (z. B. Dobbelen et al., 2002; Hallinan & Sørensen, 1985; Hanushek, 1986; Hoxby, 2000; Shin & Raudenbush, 2011; Wößmann, 2005a; Wößmann & West, 2006). Hierfür werden verschiedene Gründe angeführt. Zum einen liegen Messprobleme vor, wenn statt der eigentlichen Schulklassengrösse – d.h. Anzahl von Schülerinnen und Schüler in einer Schulklasse – das aggregierte Schüler-Lehrerverhältnis verwendet wird (z. B. Hanushek, 1986). Zum anderen werden mehrheitlich Analysen im quasi-experimentellen Design durchgeführt, wobei systematische Allokationen und Selektionen der Schülerinnen und Schüler in die Klassen mit unterschiedlicher Anzahl von Mitschüler\*innen durch Schulen und Eltern unberücksichtigt bleiben (Wößmann & West, 2006). Wenn jedoch die Klassengrösse eine systematische statt zufällige Wahlgrösse ist, weil die Schulkinder über Eltern oder Lehrpersonen in überzufälliger Weise in Klassen mit einer bestimmten Grösse gelangen, dann führen einfache Regressionsanalysen mit Schulleistungen als abhängige und Klassengrösse als unabhängige Variable zu verzerrten Schätzungen (Akerhielm, 1995). Randomisierte und kontrollierte Feldexperimente wie etwa das STAR-Projekt in Tennessee, bei denen die Schulkinder zufällig auf kleine Klassen (13-17 Kinder) und grosse Klassen (22-25 Kinder) verteilt wurden, haben ergeben, dass Lerneffekte in kleinen Klassen am grössten sind (Blatchford & Mortimore, 1994; Finn & Achilles, 1990, 1999; Krueger & Whitmore, 2001; Mosteller, 1995; Schanzenbach, 2006). Jedoch ist die externe Validität solcher Experimente gering (Ehrenberg et al., 2001; Zanger & R. Becker, 2019). So werden eben diese Selektionsprozesse in die Klassen unterschiedlicher Grössen als soziales Faktum im Schulwesen ausser Acht gelassen (Shin & Raudenbush, 2011). Zudem beschränken sich die äusserst seltenen Feldexperimente auf die Elementar- und Primarschulstufe in den USA (Hanushek, 1999) und ferner werden – abgesehen von den Ausgangszuständen in Leistungen (Zanger & R. Becker, 2019) – erklärende Mechanismen für den Zusammenhang von Klassengrösse und Schulleistungen nicht gemessen (Hallinan & Sørensen, 1985).

Denn offensichtlich ist es der Fall, dass die Klassengrösse an sich *keinen* bedeutsamen Effekt auf schulische Leistungen der Schüler\*innen hat (Hanushek, 2003; Hattie, 2005; Wößmann, 2005a). Beobachtete Effekte erfolgen über komplexe Prozesse, die unter anderem mit der Sozialstruktur der Schülerschaft (z. B. soziale Selektion in die Schulen und Schulklassen), den Unterrichtsmerkmalen (z. B. Qualität und Quantität des Unterrichts), der Schulstufe (z. B. kumulative Lerneffekte im Bildungsverlauf), dem Schultyp (z. B. Curriculum, soziale oder ethnische Herkunft der Schülerschaft) oder dem sozialen Kontext der Schulen (z. B. Grad der Urbanität oder ökonomische Wohlfahrt der Quartiere) einhergeht (Blatchford & Mortimore, 1994; Blatchford et al., 2003; Coleman et al., 1966; Ehrenberg et al., 2001; Finn et al., 2003; Hattie, 2005; Rivkin et al., 2005). Bei Kontrolle sozialer Kontexte und Prozesse sind die Klasseneffekte auf Schulleistungen oftmals äusserst klein oder gar statistisch insignifikant (Wößmann, 2005a). Zudem weisen die Vorzeichen der Klassengrösseeffekte in verschiedene Richtungen (z. B. Chingos, 2013; Hallinan & Sørensen, 1985; Hanushek, 1986; Hanushek & Woessmann, 2017; Hoxby, 2000; Leuven et al., 2008; Mitchell et al., 1989; Shen & Konstantopoulos, 2017; Slavin, 1990; Wößmann & West, 2006).

Inzwischen wurden methodische Probleme quasi-experimenteller Studien zur Identifikation eines *Kausaleffekts* von Schulklassengrösse dadurch weitgehend behoben, indem die systematische Selektion und Allokation der Schulkinder in unterschiedliche Klassengrösse sowie die elterliche Wohnort- und Schulwahl (Endogenität) nach Schulleistung (Simultanität) mit angemessenen Schätzverfahren wie etwa die *Instrument-Variable-Regression* in Rechnung gestellt werden (Wößmann, 2005a). So wird der zweistufige Kleinst-Quadrat-Schätzansatz (2SLS) verwendet, um das Endogenitätsproblem zu umgehen, indem exogene Schulvariablen (z. B. durchschnittliche Klassengrösse in einer Schule) für die Identifikation des kausalen Effekts der Klassengrösse auf die Schulleistungen verwendet wird (Akerhielm, 1995; Altinok & Kingdon, 2012; Muller et al., 2015; Wößmann & West, 2006). Ein alternativer Ansatz ist das *RDD (Regression Discontinuity Design)*, welches geeignet ist, mit Querschnittsdaten im quasi-experimentellen Design den kausalen Effekt der Klassengrösse auf schulische Leistungen zu schätzen. Mit diesem Verfahren und den Querschnittsdaten von TIMSS 1995 hat Wößmann (2005b) für die Schweiz im Unterschied zu anderen Teilnehmerstaaten einen positiven statistisch signifikanten Effekt der Schulklassengrösse auf die Mathematikleistung aufgedeckt. Dieser Kausaleffekt wurde für die Schweiz erst dann insignifikant, als weitere Informationen zu den Mitschüler\*innen und zum familiären Hintergrund kontrolliert wurden. So gelangt Wößmann (2005b) zum Schluss, dass es zweifelhaft sei, dass die Reduktion der Schulklassengrösse eine aussichtsreiche Strategie sei, um die Schulleistungen in der unteren Sekundarstufe anzuheben.

Weil dieses abweichende Ergebnis bislang ein Einzelbefund ist, welches bereits längere Zeit zurückliegt, soll in der vorliegenden Studie eine erweiterte *Replikation* – d.h. eine Reanalyse einer Studie mit derselben Analysestrategie, aber mit aktuellen Daten – für die Schweiz, einem Land mit einem föderal und hierarchisch ausgestalteten Schulsystem sowie mit einem ausgeprägten System des frühzeitigen leistungsbezogenen und sozial selektiven Übergangs von der Primar- in die untere Sekundarstufe, vorgenommen werden (Angelone, 2019; R. Becker, 2024; Buchmann et al., 2016). In Bezug auf den Übergang in die Schultypen ist es eine institutionelle Besonderheit, dass leistungsstarke Schüler\*innen in grösseren Gymnasialklassen unterrichtet werden, während die leistungsschwächeren Schulkinder eher in die kleineren Klassen in den Sekundarschulen mit Grundanforderungen oder erweiterten Anforderungen gelangen. Zudem liegen für dieses Land immer noch keine detaillierten mechanismenbasierte Befunde für den Zusammenhang von Klassengrösse und Schulleistung vor.

Diese Replikation wird in einem quasi-experimentellen Design mit Querschnittsdaten der *Überprüfung des Erreichens der Grundkompetenzen (ÜGK) 2016* durchgeführt (Konsortium ÜGK, 2019). Im Frühjahr 2016 wurde landesweit für alle Sprachregionen die Mathematikleistung am Ende der Pflichtschulzeit – in der letzten Klassenstufe der Sekundarstufe I (9. Klasse oder in der HarmoS-Zählweise: 11. Klassenstufe) – getestet. Ziel von ÜGK 2016 war, mittels standardisierter Leistungstests zu überprüfen, inwieweit die Schüler\*innen die curricular festgelegten Bildungsziele im Fach Mathematik erreicht haben. Weil neben den Testwerten auch Informationen zur sozialen Herkunft der Probanden und ihre individuelle Lernvoraussetzungen (z. B. Motivation, Interesse etc.), zum Mathematik-Unterricht (z. B. Lehrmethoden, Wahrnehmung mathematischer Inhalte, Unterrichtsgestaltung), zu den Schulklassen (z. B. Klassengrösse, Erwartungen der Eltern, Wertschätzung von Mathematik in der Klasse) und Schulen (z. B. Anzahl der Klassen, Sozialstruktur der Schülerschaft) einschliesslich ihrer kantonalen Einbettung (z. B. lokaler Wohlstand, durchschnittliche Klassengrösse im Kanton) vorliegen, sind differenzierte Untersuchungen zu den Auswirkungen unterschiedlicher Klassengrösse auf die Mathematikleistung möglich. Hierbei kommen diejenigen statistischen Verfahren wie etwa die IV-Regression oder das RDD zum Zuge, mit denen der kausale Einfluss der Klassengrösse auf die Mathematikleistungen unverzerrt geschätzt werden kann.

## 2. Theoretischer Hintergrund

Für eine empirische Analyse des kausalen Effektes der Schulklassengrößen auf die schulischen Leistungen von Schulpflichtigen ist es essenziell, dass ihr eine *mechanismenbasierte Erklärung* zugrunde liegt. Solch eine konsistente Theorie erlaubt eine präzise empirische Analyse des vermuteten Kausaleffekts variierender Größen von Schulklassen (Raub et al., 2022). Betrachtet man den Forschungsstand, so mangelt es an empirisch bewährten und konsistenten Theorien, welche erklären, *warum* es einen Effekt der Klassengröße auf die Schulleistung geben soll (z. B. Finn et al., 2003; Pedder, 2006; Pong & Pallas, 2001; Slavin, 1990). Eher dominieren Ad-hoc-Annahmen über Wirkmechanismen reduzierter Klassengrößen. Zum einen stellen sie auf qualitativ verbesserten Unterricht und intensivere Lehrperson-Schulkinder-Interaktionen ab. Zum anderen werden das effektivere Klassenmanagement und die Kontrolle des Klassenkontextes als leistungsförderliche Prozesse angeführt (Bourke, 1986; Hattie, 2005). Allerdings zeigt eine Vielzahl empirischer Studien, dass nicht nur der Effekt der Klassengröße auf die schulischen Leistungen vergleichsweise gering sei, sondern dass es keinen bedeutsamen Konnex zwischen der Klassengröße als Ursache und den sozialen Prozessen im Klassenraum gibt (Hanushek, 1999; Hattie, 2005). Oder die Korrelation zwischen variierender Klassengröße und dem Instruktionsverhalten der Lehrpersonen entspricht nicht den theoretischen Erwartungen (Finn et al., 2003; Slavin, 1990). Insgesamt kommen etliche Studien und Meta-Analysen zum Schluss, dass der gegenwärtige Wissensstand gering sei und es an empirischer Evidenz mangelt, Effekte unterschiedlicher Klassengrößen zu verstehen (Blatchford, 2003).

Somit ist nicht ausgeschlossen, dass gänzlich *andere* soziale Mechanismen und Prozesse als die Klassengröße die Unterschiede in den Schulleistungen hervorbringen (Dobbelsteen et al., 2002). So zeigten bereits Coleman et al. (1982) für die USA, dass die Schülerschaft in katholischen Privatschulen deswegen höhere Schulleistungen als diejenige in öffentlichen Schulen erbringen, weil sie eher leistungsstarke Jugendliche aus sozial privilegierten Elternhäusern rekrutieren und auch höhere Leistungsanforderungen an ihre sozial privilegierte «Klientel» stellen. Für die Analyse des Zusammenhangs von Klassengröße und schulischer Leistung (in Mathematik) am Ende der Pflichtschulzeit ist anhand dieses Beispiels zu bedenken, dass sowohl der persönlichen und leistungsbezogenen Entwicklung als auch dem Bildungsverlauf eine Vielzahl an schulischen und ausserschulischen Einflüssen vorangegangen ist (Blossfeld et al., 2019; Coleman et al., 1966; Ehrenberg et al., 2001; Hoxby, 2000). So belegt eine kaum überschaubare Zahl an empirischen Studien mit Daten aus Schulleistungsuntersuchungen wie etwa PISA, PIRLS oder TIMSS einen positiven Zusammenhang zwischen sozialer Herkunft (Klassenlage, sozioökonomischer Status oder Bildungsniveau des Elternhauses) und schulischer Leistung sowie zwischen der sozialstrukturellen Zusammensetzung der Schülerschaft in der Schulkasse oder auf Schulebene und der Schulleistung (z. B. Baumert & Schümer, 2001; Baumert et al., 2003; D. Becker, 2019; Jungbauer-Gans, 2004; Schubert & R. Becker, 2010; Weiß, 2011; Zahner et al., 2002). Ebenso unüberschaubar ist die Vielzahl an Studien, welche den Einfluss der sozialen Herkunft auf den Übergang in die Schullaufbahnen in der Sekundarstufe I nachweisen (z. B. R. Becker, 2022, 2024; Erikson, 2019; Jackson, 2013). Diese Herkunfts- und Selektionsprozesse seitens des Elternhauses tragen – abgesehen von Jahrgangsstärken – zu einem wesentlichen Teil dazu bei, wie in Schulklassen mit unterschiedlicher Größe die Schulleistungen letzten Endes ausfallen. Zudem gibt es keine gesicherten Erkenntnisse dafür, wie lange Schulkinder bestimmten Klassengrößen ausgesetzt sein müssen, bis sich signifikante Folgen reduzierter Klassengrößen zeigen.

So wird wahrscheinlich aufgrund der beschriebenen sozialen Selektivität der Rekrutierung für höhere Schullaufbahnen und sozialen Homogenität der Schülerschaft in diesen Schulklassen ein positiver Zusammenhang zwischen Klassengröße und Schulleistung beschrieben (Wößmann, 2005b). Zudem können Schulen bei paralleler Klassenführung durch Selektion und Allokation der Schulkinder in die Schulklassen festlegen, ob leistungsschwächere Schülerinnen und Schüler in kleine Klassen (um sie besser durch ein effizientes Klassen- und Unterrichtsmanagement fördern zu können) und leistungsstärkere Schülerinnen und Schüler in grosse Klassen (um sie zu fordern) gelangen (Konstantopoulos, 2008; Nye et al., 2002). Somit wird der Zusammenhang zwischen Klassengröße und Schulleistung systematisch hergestellt (Altinok & Kingdon, 2012). Diese Korrelation kann anhand der *sozial-kognitiven Lerntheorie* von Bandura (1986) theoretisch plausibilisiert werden. Demnach lernen Schüler\*innen nicht nur von den Lehrpersonen, sondern auch von Klassenkamerad\*innen. So würde die schulische Leistung von einer größeren Anzahl von Mitschüler\*innen positiv beeinflusst, die ein ähnliches Niveau von kognitiven Fähigkeiten aufweisen (Dobbelsteen et al., 2002; Pedder, 2006). Zudem gibt es seit dem ersten Coleman-Report (Coleman et al., 1966) immer wieder Hinweise dafür, dass in größeren Schulklassen sozial benachteiligte Gruppen in ihrer Leistungsentwicklung von einer höheren Zahl sozial privilegierter und leistungsstarker Mitschüler\*innen profitieren (Coleman et al., 1975; D. Becker, 2019). Diesen theoretischen Argumenten folgend kann geschlossen werden, dass reduzierte Klassengrößen das Gegenteil bewirken kann, als dies von den Akteuren intendiert wird.

Allerdings werden den Wahlentscheidungen von Eltern und Schulen institutionelle Grenzen gesetzt (Angrist & Lavy, 1999; Hoxby, 2000). So legen etwa in der Schweiz die Kantone Richtgrössen für die festen Klassenverbände in der Sekundarstufe I fest. Die Normwerte für die maximale Klassengrösse liegen in der Regel zwischen 23 und 25 Schüler\*innen pro Klasse. Jedoch variieren sie zwischen den Kantonen und für einige der Kantone auch zwischen den Schultypen. Im letzteren Fall wird für das Gymnasium eine höhere Maximalgrösse definiert als für die anderen Schultypen. Somit gibt es eine Korrelation zwischen vorheriger schulischer Performanz (etwa in der Primarstufe), der Allokation in die Schultypen und in die Schulklassengrösse und der aktuell beobachteten Schulleistung (Pong & Pallas, 2001). Die kantonale Festlegung der maximalen Schulklassengrösse kann – wie etwa von Angrist & Lavy (1999) für die Regel von Maimonides in Israel demonstriert – genutzt werden, um Analysen im quasi-experimentellen Design vorzunehmen und hierbei – etwa mit dem *Regression Discontinuity Design* (Angrist & Pischke, 2009; Lee & Lemieux, 2015) – den *kausalen* Effekt von Schulklassengrösse auf die Schulleistungen zu schätzen.

Im Folgenden wird – ausgehend von diesen theoretischen und methodischen Darlegungen – für die Schweiz anhand der *ÜGK-Daten 2016* untersucht, ob in der letzten Klassenstufe am Ende der Pflichtschulzeit die Klassengrösse eine bedeutsame Rolle für die Mathematikleistungen spielt. Weil bislang *keine* aktuelle Untersuchung für die Schweiz vorliegt, wird versucht, diese Forschungslücke zumindest teilweise für eine bestimmte Population unter den Schüler\*innen zu schliessen. Zudem ist der Messzeitpunkt am Ende der Pflichtschulzeit interessant, weil sich die Schülerinnen und Schüler entscheiden müssen, welchen weiteren Bildungsweg – Wechsel in die Berufsausbildung oder Fortfahren auf dem Gymnasium – sie einschlagen wollen. Gemäss vorliegender Forschung haben die schulischen Leistungen einen erheblichen Einfluss auf diese Bildungsentscheidung (R. Becker, 2024).

### 3. Daten, Variablen, Analysestrategie und statistische Verfahren

#### 3.1 Daten

Die empirischen Analysen basieren auf Querschnittsdaten (Nidegger, 2019), die im Rahmen der Erhebung zur *Überprüfung des Erreichens der Grundkompetenzen (ÜGK)* im Frühjahr 2016 in der Schweiz erhoben wurden (Konsortium ÜGK, 2019). Mit ÜGK 2016 sollte auf nationaler und kantonaler Ebene untersucht werden, inwieweit Schülerinnen und Schüler am Ende der obligatorischen Schulzeit die nationalen Bildungsziele – d.h. die im Rahmen des HarmoS-Konkordats definierten Grundkompetenzen – in Mathematik erreicht haben. Die standardisierten Leistungstests in Mathematik wurden durch einen Fragebogen zu individuellen, schulischen und familiären Aspekten ergänzt, die zentrale Indikatoren zur Erklärung von Leistungsunterschieden darstellen (Hupka-Brunner et al., 2016).

Die schweizweite Grundgesamtheit umfasste 85.000 Schüler\*innen, die in der 9. Klassenstufe (11. Klasse in der HarmoS-Zählweise) in rund 1.500 Schulen unterrichtet wurden (Verner & Helbling, 2019). In den Kantonen Appenzell-Innerrhoden, Bern (französischer Teil), Glarus, Jura, Nidwalden, Obwalden, Schaffhausen, Uri, Wallis (deutscher Teil) sowie Zug wurden keine Stichproben gezogen, sondern Vollerhebungen durchgeführt. In den Kantonen Basel (Stadt und Land), Freiburg, Genf, Graubünden, Neuenburg, Solothurn, Schwyz, Thurgau, Tessin und Wallis (französischer Teil) wurde mittels eines einstufigen Stichprobenverfahrens ein bestimmter Anteil der Schülerinnen und Schüler in den zur Teilnahme aufgebotenen Schulen zufällig gezogen. In den Kantonen Aargau, Bern (deutscher Teil), Luzern, Sankt Gallen, Waadt und Zürich wurde ein zweistufiges Stichprobenverfahren eingesetzt. Zuerst sind Schulen und dann innerhalb der teilnehmenden Schulen Schüler\*innen zufällig ausgewählt worden. Insgesamt beteiligten sich 22.423 Schüler\*innen an dieser Studie (Konsortium ÜGK, 2019). Sämtliche Analysen erfolgen unter Berücksichtigung des Survey-Designs anhand der im Datensatz bereitgestellten Ausfallgewichte für die Schulen und die BRR-Gewichte («*balanced repeated replication*») (Jann & Hupka-Brunner, 2019).

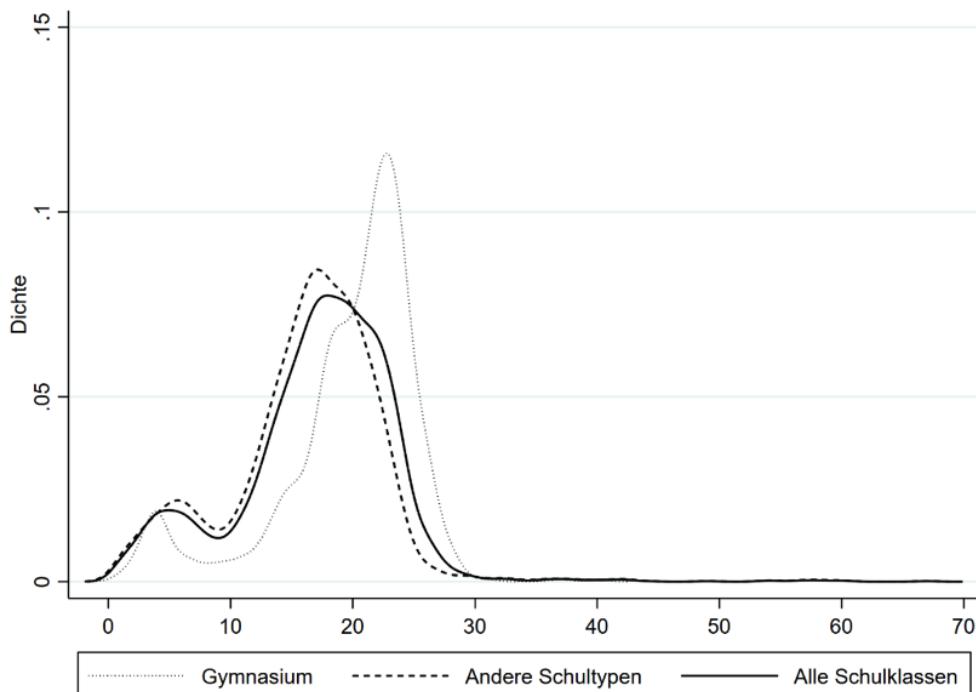
#### 3.2 Abhängige und unabhängige Variablen

Die abhängige Variable sind die *Fachkompetenzen in Mathematik*. Hierbei wird anstelle der «*Plausible Values*» der WLE-Score («*weighted likelihood estimate*») verwendet (Angelone & Keller, 2019). Der Mittelwert des

WLE-Scores (Wertebereich: -5,7; 5,2) liegt bei 0,01 mit einer Standardabweichung von 1,4 (Kurtosis: 3).<sup>1</sup> Die zentrale erklärende Variable für die Variation in den Mathematikleistungen ist die *Schulklassengröße*. Sie wird anhand der Anzahl der Schüler\*innen in einer Schulkasse gemessen. Weil davon ausgegangen wird, dass der Einfluss der Schulklassengröße auf die Mathematikleistung nicht linear, sondern kurvilinear verläuft, wird auch die *quadrierte Schulklassengröße* als erklärende Variable berücksichtigt (Borland et al., 2005).<sup>2</sup>

### Abbildung 1

Kerndichteschätzung für Schulklassengröße



Daten: ÜGK 2016 – eigene Darstellung

In *Abbildung 1* ist auf Schulklassenebene die Verteilung der Klassengrößen dargestellt. Von den 3.520 Schulklassen liegen für 3.337 Klassen vollständige Angaben zur Anzahl der Schüler\*innen im Wertebereich von 1 bis 67 sowie dem Schultyp vor. Der Mittelwert liegt bei 17 Schüler\*innen (Standardabweichung: 6,4). Die Verteilung ist linkssteil (Schiefe: 0,3) und hat bei einer Kurtosis von 7,8 mehr extreme Ausreisser am rechten Rand als eine normale Verteilung. Daher werden die Kennwerte für die Verteilungen nur bis zu einer Klassengröße von 29 Schülerinnen und Schülern betrachtet. Denn gerade 1 Prozent der Schulklassen sind grösser als 29 und allenfalls 1,5 Prozent der Schüler\*innen befinden sich in Klassen mit mehr als 29 schulpflichtigen Jugendlichen. Die Fallzahl reduziert sich auf 3.299 Schulklassen. Hervorzuheben ist, dass sich dabei die Mittelwerte und Standardabweichungen kaum verändern. Für diese Schulklassen ist die Verteilung linksschief und weicht

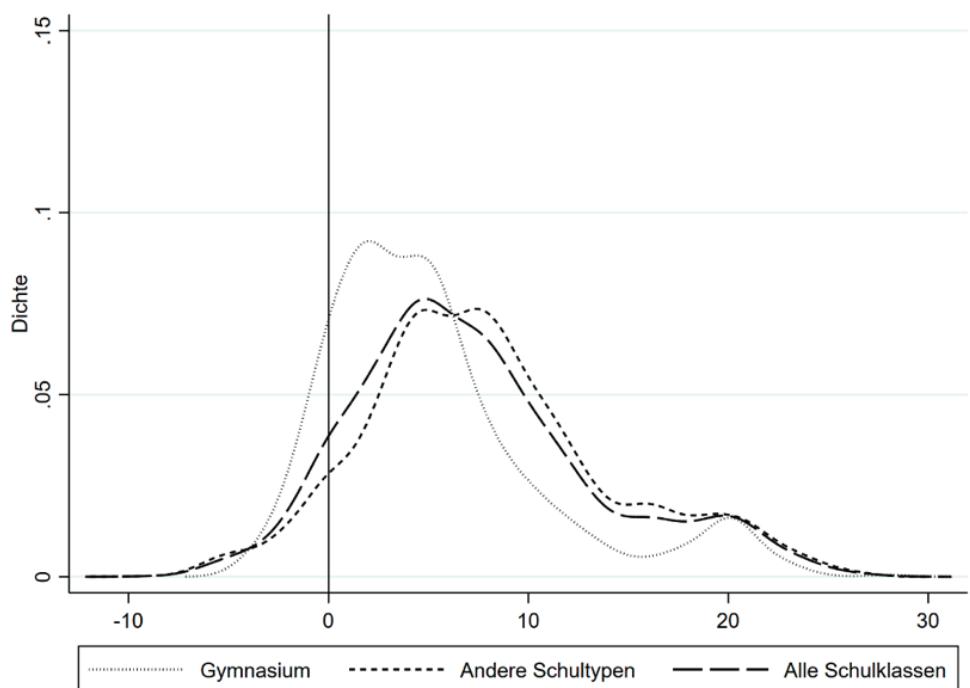
<sup>1</sup> Die Verteilung der Mathematikleistungen variiert mit der Anzahl Jugendlicher in einer Schulkasse (*Abbildung A.1* im Anhang). Werden die Klassengrößen nach dem STAR-Projekt in Tennessee gewählt, so liegt die mittlere Mathematikleistung in kleinen Schulklassen (13-17 Schülerinnen und Schüler) bei -0,46 und diejenige für grosse Schulklassen (22-25 Schüler\*innen) bei 0,41 mit jeweils der gleichen Standardabweichung von 1,3 (Effektgrösse  $d$  nach Cohen: 0,66). Die Mathematikleistung ist desto höher, je grösser die Schulkasse ist.

<sup>2</sup> Bei Kontrolle des sozioökonomischen Status, der auf allen analytischen Ebenen (Individuum, Klasse, Schule und Kanton) positiv mit der Mathematikleistung korreliert, zeigt sich, dass der WLE-Score mit steigender Klassengröße (IV) zunimmt, aber bei sehr grossen Schulklassen wieder abnimmt (*Tabelle A.1* im Anhang). Der Wendepunkt für den kurvilinearen Zusammenhang – und damit die optimale Klassengrösse – liegt für alle Schulklassen bei  $(0,4759661/(2 \cdot 0,0124895)) = 19$  Schüler\*innen.

deutlich weniger von einer Normalverteilung ab. Für die 808 Gymnasialklassen liegt der Mittelwert bei 19,6 mit einer Standardabweichung von 5,7. Deren Verteilung ist ebenso linksschief wie für die Schulklassen in den anderen Schultypen. Für diese 2.491 Schulklassen liegt der Mittelwert bei 15,8 und die Standardabweichung beträgt 5,6. In Bezug auf die Schulklassengrösse sind die Unterschiede zwischen den Schultypen bedeutsam, denn der für ungleiche Gruppen korrigierte Wert für Cohen's  $d$  liegt bei  $-0,7$  und die Effektgrösse  $r$  bei  $-0,2$ .

**Abbildung 2**

*Kerndichteschätzung für Differenz der vorgegebenen und beobachteten Klassengrösse («cutoff»  $c=0$ )*



Daten: ÜGK 2016 – eigene Darstellung

Der *Differenz zwischen der kantonal vorgegebenen Schulklassengrösse und der faktisch gemessenen Schulklassengrösse* wird deswegen ein besonderes Augenmerk geschenkt, weil – wie noch ausgeführt wird – der Nullwert als Grenzwert («cutoff»  $c=0$ ) für die statistische Analyse im *Regression Discontinuity Design* (RDD) verwendet wird. Sie beträgt für alle Schulklassen zusammen bei einem Wertebereich von  $-9$  bis  $28$  im Mittel rund  $7,3$  und die Standardabweichung  $6,2$  (Abbildung 2). Diese Verteilung ist rechtsschief (Schiefe:  $0,79$  und weicht am rechten Rand von der Normalverteilung ab (Kurtosis:  $3,2$ ). Bei den Gymnasien mit den durchschnittlichen grösseren Schulklassen liegt in einem Wertebereich von  $-4$  bis  $28$  die mittlere Differenz bei  $5,3$  mit einer Standardabweichung von  $5,7$  (Schiefe:  $1,3$ ; Kurtosis:  $4,5$ ). Für die anderen Schultypen ist bei einem grösseren Wertebereich ( $-9$ ;  $27$ ) die mittlere Differenz der vorgegebenen und tatsächlichen Klassengrösse von  $8$  bei einer Streuung von  $6,2$  grösser als für die Gymnasien. Die Unterschiede zwischen den Schultypen sind bedeutsam. Der für ungleiche Gruppen korrigierte Wert für Cohen's  $d$  beträgt  $0,4$  und die Effektgrösse  $r$  ist  $0,2$ .

Weitere unabhängige Variablen sind auf folgenden analytischen Ebenen – wie etwa der Individualebene, der Schulklassenebene, der Schulebene und der Kantonalebene – angesiedelt. Auf der Individualebene wird das *Geschlecht* der Probanden (Referenz: männlich), der *Migrationsstatus nach Generationen* (Ref.: Schweizer Staatsbürgerschaft), die *zu Hause gesprochene Sprache* (Ref.: nicht Testsprache) sowie der *sozioökonomische Status des Elternhauses* (ISEI) und das *elterliche Bildungsniveau* (mittleres und hohes Niveau vs. tiefes Niveau) (Pham et al., 2019) berücksichtigt. Des Weiteren werden die *Schulnoten in Mathematik und Naturwissenschaften* als Ersatz für die fehlende Messung kognitiver Leistungen (IQ) sowie die vorherige Klassenwiederholung (Ref.: keine Klassenwiederholung) und die *idealistische Bildungsaspiration für tertiäre Bildung* (Referenz: keine Tertiärbildung) als Indikator für die Motivation zur schulischen Anstrengung in Rechnung gestellt. Als weitere

Korrelate der Fachkompetenzen gehen das *mathematische Selbstkonzept*, die *wahrgenommene Wertschätzung von Mathematik* in der Klasse, die *elterliche Leistungserwartung in diesem Schulfach* (Hascher et al., 2015) und die *Inanspruchnahme von Privatunterricht in Mathematik* (Ref.: keine Nutzung) in die Analysen ein.

Ferner wird auf der Schulklassenebene für den *durchschnittlichen sozioökonomischen Status* der Klasse, den *prozentualen Anteil von Schülerinnen* sowie die *kognitive Aktivierung durch die Lehrperson* (ja vs. nein) sowie deren *Klassenmanagement* (Störung des Unterrichts versus ruhigen Unterricht) kontrolliert (Ehrenberg et al., 2001). Als Schulmerkmale werden der *durchschnittliche sozioökonomische Status der Schülerschaft*, der *prozentuale Anteil der Migranten*, die *Anzahl der Schulklassen* (9. Klassenstufe) und der *Schultyp* (Gymnasium versus andere Schultypen) sowie der *regionale Kontext* der Schule (urbaner oder ländlicher Kontext versus gemischten Kontext) in Rechnung gestellt.

### 3.3 Analysestrategie und statistische Verfahren

Um einen *kausalen Einfluss* der Grösse von Schulklassen auf die Leistungen schulpflichtiger Jugendlicher eindeutig identifizieren zu können, bedarf es entweder eines randomisierten Feldexperiments (Zanger & R. Becker, 2019) oder zeitkontinuierlicher Längsschnittdaten (Pötter & Blossfeld, 2001). Weil im vorliegenden Fall weder ein experimentelles Design noch eine Längsschnittanalyse möglich sind (Angrist & Pischke, 2009; Murnane & Willett, 2011), wird trotz aller methodischer Vorbehalte versucht, mit Querschnittsdaten im quasi-experimentellen Design einen Ursache-Wirkungszusammenhang von institutionell variablen Klassengrössen und den Bildungsergebnissen aufzudecken (Hanushek & Woessmann, 2017). Wegen des im theoretischen Teil beschriebenen Problems der Endogenität von Schulklassengrössen und Simultanität von Leistungsgrössen (Wößmann, 2005a) führt der Versuch, den Einfluss von Schulklassengrössen auf die Mathematikleistung mit einer einfachen OLS-Regression zu schätzen, zu erheblich verzerrten Ergebnissen (Angrist & Lavy, 1999; Hoxby, 2000; Murnane & Willett, 2011; Wößmann & West, 2006). Daher werden im Folgenden mögliche Strategien diskutiert, welche in der vorliegenden Studie zur Anwendung kommen. Die Auswahl dieser Strategien orientiert sich an ihrer statistischen Logik – der Lieferung eines unverzerrten kausalen Schätzers für den Einfluss der Schulklassengröße – und ihrer erfolgreichen Umsetzung in früheren Studien (z. B. Hanushek & Woessmann, 2017; Wößmann, 2005a, 2005b). Es werden deswegen mehrere verschiedene Verfahren verwendet, um sicherzugehen, keine weitreichenden Schlussfolgerungen auf Basis eines zufälligen Ergebnisses zu ziehen. Über Reanalysen – d.h. Analysen mit den gleichen Daten, aber unterschiedlichen Designs und statistischen Verfahren – soll im Sinne von Robustheits- und Reliabilitätstest das Risiko für fehlerhafte Ergebnisse minimiert werden.

Eine erste Möglichkeit, dieses Problem zu lösen, ist die *Instrumental Variable Regression* (IV-Schätzung) mit der zweistufigen Kleinstquadrate-Schätzung (Muller et al., 2015). Im vorliegenden Fall wird in einem ersten Schritt durch Kontrolle von Schulmerkmalen die Variation zwischen Schulen berücksichtigt, so dass der IV-Schätzer auf der Variation innerhalb von Schulen basiert. Die IV-Schätzung vergleicht Schülerleistungen in Klassen, deren beobachtete Grösse aus exogenen Gründen variiert. Sie zeigt an, ob unterschiedliche Klassengrössen zu systematischen Leistungsunterschieden führen. Weil die Verteilung eines Jahrgangs auf verschiedene Klassen innerhalb einer Schule ebenfalls systematisch erfolgen kann, wird im zweiten Schritt die empirisch beobachtete Grösse einer Schulkasse im ÜGK-Datensatz durch die durchschnittliche faktische Grösse der Klassen in der ein- und derselben Jahrgangsstufe im jeweiligen Kanton instrumentiert. Diese durchschnittliche kantonale Klassengröße für Neunklässler als Instrumentalvariable korreliert hochsignifikant mit der Grösse der Schulkasse ( $r = 0,3$ ;  $p \leq 0,001$ ), in der die Probanden unterrichtet werden, aber nicht mit ihren Mathematikleistungen als abhängiger Variable ( $r = -0,0124$ ), und ist zudem keine Linearkombination anderer erklärender Variablen (Wößmann & West, 2006). Auch die quadrierte Klassengröße auf Kantonalebene korreliert signifikant mit der faktisch gemessenen Klassengröße im Quadrat ( $r = 0,3$ ), aber nicht mit der getesteten Mathematikleistung ( $r = -0,0124$ ). Der Wu-Hausman-F-Test ergibt in der ersten Stufe der IV-Schätzung (ohne Survey-Gewichtung mit BRR) einen hochsignifikanten Wert von 25,8 und weist bei einem Wert deutlich grösser als 10 darauf hin, dass ein «starkes» Instrument vorliegt. Dieser Befund wird durch den Anderson-Rubin-Chi<sup>2</sup>-Test mit einem statistisch signifikanten Wert von 64,3 bestätigt. Ein abschliessender Test auf Endogenität des gewählten IV-Regressors war der statistisch signifikante Hausman-Test für die IV-Regression und OLS-Regression. Weil jetzt die Variation von Schulklassengrössen exogen in Bezug auf die Mathematikleistung in der Schulkasse ist, können systematische Leistungsunterschiede nach Klassengrössen als kausale Folge unterschiedlich grosser Schulklassen interpretiert werden.

Der Vorzug der IV-Schätzung liegt darin, dass das Instrument einen ergebnisdefinierten Behandlungseffekt darstellt und somit ein lokaler durchschnittlicher Behandlungseffekt (LATE: *local average treatment effect*)

geschätzt werden kann (Murnane & Willett, 2011). Mit anderen Worten: Das Instrument – die durchschnittliche Klassengrösse der ein- und derselben Jahrgangsstufe im jeweiligen Kanton – identifiziert den LATE, d.h. den durchschnittlichen kausalen Effekt der Behandlung für die Teilmenge der Schülerschaft, deren Behandlungsauswahl durch das Instrument induziert wird.

Um die Robustheit der IV-Schätzungen beurteilen zu können, nicht zuletzt wegen ihrer restriktiven Voraussetzungen (Murnane & Willett, 2011), werden auch alternative Schätzverfahren herangezogen. Weil sowohl die tatsächliche Schulklassengrösse (bemessen an der Zahl der Schülerinnen und Schüler in der 9. Klasse) erhoben wurde, als auch die maximale Schulklassengrösse (ohne Klein- oder Sonderklassen), die von den Kantonen festgelegt wird und zwischen den einzelnen Kanton variiert, bekannt ist (Schweizerische Konferenz der kantonalen Erziehungsdirektoren, 2024), kann die *Differenz zwischen der vorgeschriebenen Maximalgrösse und beobachteten Klassengrösse* mit einem *Regression Discontinuity Design* (RDD) als ein «natürliches» Experiment analysiert werden (Angrist & Pischke, 2009; Lee & Lemieux, 2015). Die zufällige Abweichung der tatsächlichen von der vorgegebenen Klassengrösse kann sich demnach kausal auf die Mathematikleistungen auswirken. Diese Diskontinuität in der Klassengrösse erlaubt die Festlegung einer Experimental- und Kontrollgruppe (Angrist & Lavy, 1999). Die Schulkasse, welche die festgelegte Maximalgrösse überschreitet, ist die Experimentalgruppe, während die Schulkasse mit einem darunter liegenden Wert die Kontrollgruppe darstellt. Gäbe es eine einzige Richtgrösse für alle Kantone, dann kann die Auswirkung der Unstetigkeit in der Schulklassengrösse auf die Mathematikleistung mittels einer «scharfen» RD-Analyse analysiert werden. Weil aber die festgelegte Maximalgrösse einer Schulkasse zwischen den Kantonen variiert, liegt auch die Verwendung der «unscharfen» Regressions-Diskontinuitätsanalyse (*fuzzy RD*) nahe (Lee & Lemieux, 2015). Der Nullpunkt der Differenz beider Grössen fungiert hierbei als der «cutoff», und so können links und rechts vom Nullpunkt zwei Kleinsten-Quadrat-Schätzungen vorgenommen werden. Ob die Schüler\*innen in einer «grossen» oder «kleinen» Schulkasse sind, liegt nicht in deren Hand. Somit liegt eine exogene Variation vor, die konsistente Schätzungen des Effekts beobachteter Klassengrössen auf die Mathematikleistung der Schülerinnen und Schüler zulässt. Wie von Lee und Lemieux (2015) hervorgehoben, kann mit diesem Verfahren mittels lokaler Randomisierung wie in einem randomisierten Experiment in transparenter und zuverlässiger Weise der LATE von Schulklassen unterschiedlicher Grösse auf die Mathematikleistung kausal geschätzt werden.

Ein weiteres Schätzverfahren ist das von Heckman (1979) vorgeschlagene *zweistufige Verfahren für die Aufdeckung von Stichprobenverzerrung* infolge einer Selbstselektion in eine Situation. Die hier interessierende Situation ist die Schulklassengrösse. Hierbei soll dem Umstand Rechnung getragen werden, dass Jugendliche entweder durch ihre Eltern in eine bestimmte Klassengrösse oder durch die Schulen in eine Schulkasse selektiert werden, deren Grösse gesetzlich vorgegeben ist, aber in der Schulpraxis von der Norm abweichen kann. In einem ersten Schritt wird die Selektion in eine grosse Schulkasse mit mindestens 19 Schülerinnen und Schülern gegenüber der Allokation in eine kleine Schulkasse mit bis zu 18 Jugendlichen geschätzt und als inverser Mills-Quotient gespeichert. Dieser geschätzte *inverse Mills-Quotient* wird dann im zweiten Schritt in die eigentliche Schätzgleichung mit der Mathematikleistung als abhängiger Variable eingesetzt, um für die Selbstselektion in eine Schulkasse mit einer bestimmten Grösse zu korrigieren. Hierbei kann dieser Quotient auch als kontrafaktisches Mass («non-selection hazard») dafür interpretiert werden: Was würde daraus für die schulische Leistung einer schulpflichtigen Person resultieren, wenn sie statt in einer kleinen in einer grossen Schulkasse unterrichtet würde?

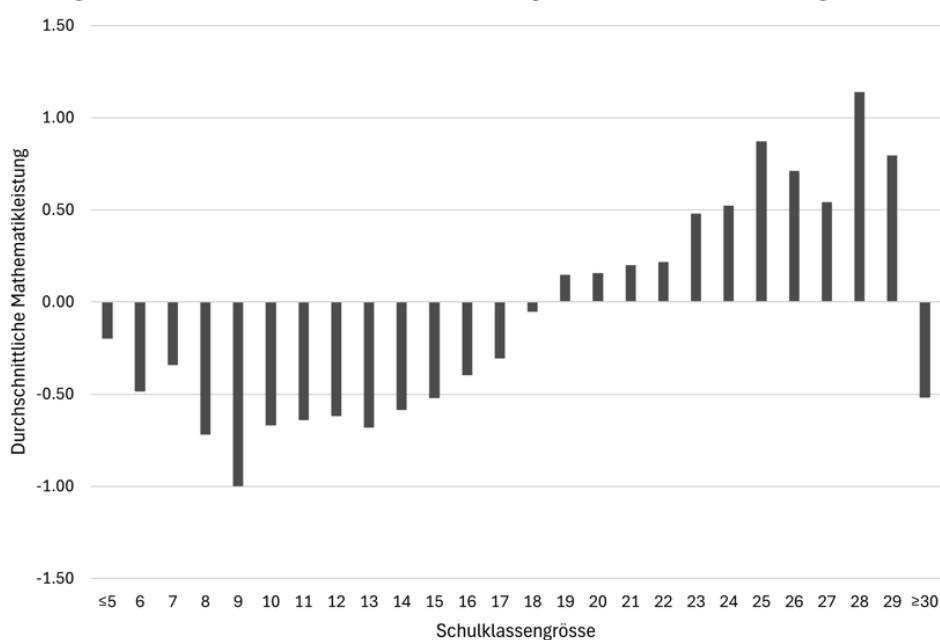
## 4. Empirische Ergebnisse

### 4.1 Verteilung der Mathematikleistung nach Grösse der Schulklassen

Wie verteilt sich die durchschnittliche Mathematikleistung über die Klassengrössen? Anhand der *Abbildung 3* wird über die negativen und positiven Mittelwerte für die Mathematikleistung bei einer jeweiligen Klassengrösse der Zusammenhang zwischen Mathematikleistung und Klassengrösse augenfällig. In Schulklassen mit weniger als 19 Jugendlichen ist die mittlere Leistung in Mathematik geringer als in den Schulklassen mit mehr als 18 Jugendlichen. Die Zunahme erfolgt bis zur Klassengrösse von 29 Jugendlichen und danach sinken die Mathematikleistungen wieder. Bei einer Zunahme der mittleren Testwerte über die Klassengrösse ( $= 1,7$  für  $< 6$ ) nimmt die Streuung der Mathematikleistung bis zur Klassengrösse von 28 Personen ( $= 1,2$ ) ab und nimmt danach zu ( $\sigma = 1,4$ ).

**Abbildung 3**

Verteilung der durchschnittlichen Mathematikleistung (WLE) nach Schulklassengrösse

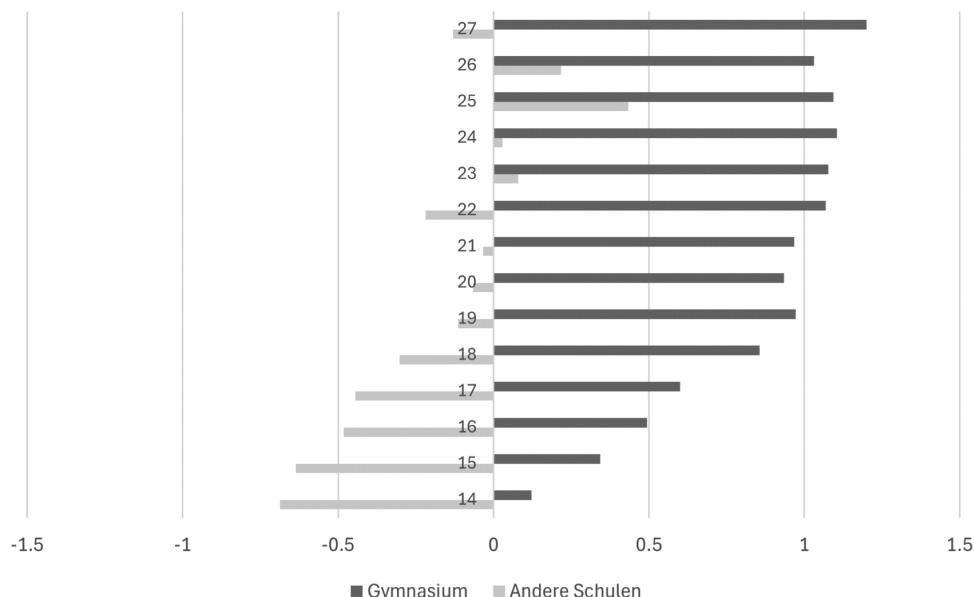


Daten: ÜGK 2016 – eigene Darstellung

Zudem variiert der zuvor aufgezeigte Zusammenhang zwischen den Schultypen (*Abbildung 4*). Die mittleren Mathematikleistungen sind im Gymnasium für alle Klassengrößen generell höher als für Schülerinnen und Schüler in den anderen Schultypen (Realschule oder Sekundarstufe I mit erweiterten Anforderungen). Unabhängig vom Schultyp nimmt die mittlere Mathematikleistung mit steigender Klassengrösse zu. Die anhand von Cohen's  $d$  bemessene Effektstärke (Cohen's  $d$ ) für «kleine» (1-18 Neunklässler) versus «grosse» Klassen (19-29 Neunklässler) beträgt für das Gymnasium 0,5 (Effektgrösse  $r = -0,2$ ) und für die anderen Schultypen -0,4 ( $r = -0,2$ ).

**Abbildung 4**

Verteilung der mittleren Mathematikleistungen (WLE) nach Schulklassengrösse und Schultypen



Daten: ÜGK 2016 – eigene Darstellung

#### 4.2 IV-Schätzungen für den Zusammenhang von Klassengröße und Mathematikleistung

Unter Kontrolle von Einflüssen auf der individuellen und schulischen Ebene auf die Mathematikleistung der Jugendlichen wird offensichtlich, dass es am Ende der Pflichtschulzeit einen systematischen kurvilinearen Zusammenhang zwischen Klassengröße und Performanz gibt (Modell 1 in *Tabelle 1*). Mit zunehmender Klassengröße steigen – bei Kontrolle allseits bekannter Korrelate der Mathematikleistung auf Ebene der Individuen, Schulklassen und Schulen – die durchschnittlichen Mathematikleistungen zunächst an, nehmen aber dann nach einem Grenzwert wieder ab, wenn die Schulklassen besonders gross sind.

Werden für die Substichproben, für die weitere Leistungsmerkmale in Mathematik erhoben wurden (Modell 2), nämlich die Wertschätzung der Mitschülerinnen und Mitschüler für Mathematik sowie die Leistungserwartungen der Eltern in diesem Fach, kontrolliert, dann zeigt sich, dass hohe elterliche Leistungserwartungen negativ mit den Testwerten korrelieren. Des Weiteren gibt es signifikante Effekte der Schulklassenmerkmale auf die Mathematikleistungen. Je höher der Anteil der Mädchen in der Schulkasse ist, desto besser fallen die Testwerte aus. Zudem gibt es Effekte des Verhaltens der Lehrpersonen auf die Mathematikleistung. Wenn die Lehrperson es vermag, die Schülerinnen und Schüler kognitiv zu aktivieren, so wirkt sich dies positiv auf ihre Mathematikleistungen aus. Treten disziplinäre Probleme in der Schulkasse auf, so hat das defizitäre Schulklassenmanagement der Lehrperson hingegen einen negativen Effekt auf die Testergebnisse.

**Tabelle 1**

*Mathematikleistung nach Schulklassengröße*

	Modelle	1	2
<i>Schulklassengröße</i>			
Grösse der Schulkasse		0,315 (0,140)*	0,224 (0,164)
Grösse der Schulkasse <sup>2</sup>		-0,010 (0,004)*	-0,007 (0,005)
<i>Individuelle Merkmale</i>			
Sozialer Status		0,006 (0,001)***	0,005 (0,001)***
Migration: 2. Generation (vs. Einh.)		-0,390 (0,032)***	-0,344 (0,045)***
Migration: 1. Generation (vs. Einh.)		-0,268 (0,044)***	-0,170 (0,056)**
Testsprache zu Hause gesprochen		0,165 (0,050)**	0,155 (0,064)*
Geschlecht (Referenz: männlich)		-0,129 (0,022)***	-0,151 (0,031)***
<i>Schulnoten</i>			
Mathematik		0,105 (0,025)***	0,134 (0,032)***
Naturwissenschaften		0,167 (0,022)***	0,150 (0,032)***
<i>Leistungsmerkmale für Mathematik</i>			
Selbstkonzept		0,333 (0,019)***	0,336 (0,022)***
Wertschätzung in Schulkasse			-0,021 (0,018)
Leistungserwartungen der Eltern			-0,112 (0,017)***
Private Unterstützung in Mathe		-0,296 (0,030)***	-0,261 (0,037)***

Klassenwiederholung	-0,370 (0,030)***	-0,419 (0,048)***
Idealistische Bildungsaspiration	0,510 (0,035)***	0,469 (0,042)***
<i>Schulklassenmerkmale</i>		
Anteil der Frauen	0,472 (0,139)***	0,439 (0,174)*
Lehrperson: Kognitive Aktivierung		0,056 (0,018)**
Lehrperson: Klassenmanagement		-0,154 (0,018)***
<i>Schulmerkmale</i>		
Sozialer Status der Schülerschaft	0,016 (0,003)***	0,011 (0,003)***
Anteil der Migranten	-0,612 (0,129)***	-0,548 (0,125)***
Anzahl der Schulklassen (9. Klasse)	0,045 (0,008)***	0,036 (0,009)***
Gymnasium (vs. andere Schultypen)	0,857 (0,073)***	0,827 (0,074)***
Konstante	-5,083 (1,089)***	-4,307 (1,238)***
Fallzahl	17.773	8.492
Population	62.954	27.566
Korrigiertes R <sup>2</sup>	0,3192	0,4136
F-Wert (#Koeffizienten, Freiheitsgrad)	268,8 (18, 102)	171,1 (22, 98)

\* p < 0,05; \*\* p < 0,01; \*\*\* p < 0,001; 2SLS-Regression (Instrumental Variable Regression, in Klammern: Standardfehler der Koeffizienten).

Daten: ÜGK 2016 – eigene Berechnungen

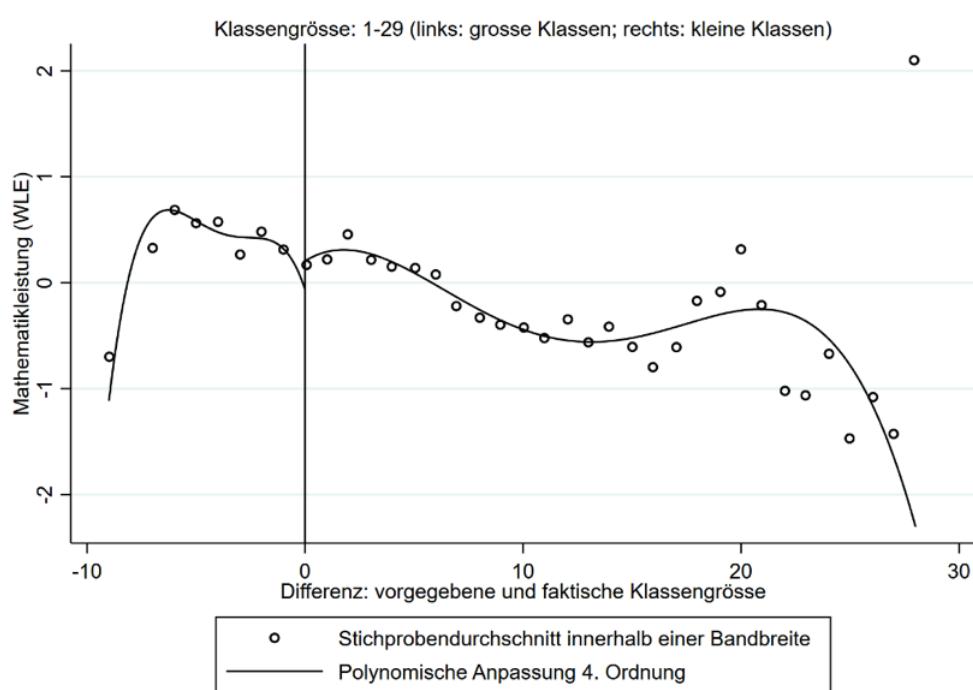
Werden diese Prozesse innerhalb des Elternhauses und der Schulkasse kontrolliert, dann ist der positive Effekt der Schulklassengrösse statistisch *insignifikant*. Somit wird auch der Befund von Wößmann (2005b) mit älteren TIMSS-Daten aus dem Jahre 1995 erfolgreich repliziert. Mit anderen Worten: Es wäre anzunehmen, dass der positive und statistisch signifikante Kausaleinfluss grösserer Schulklassen auf die Mathematikleistung der Neuntklässler über diese sozialen Interaktionen vermittelt wird. Allerdings korrelieren insbesondere das Klassenmanagement und die kognitive Aktivierung durch die Lehrpersonen *nicht* mit der Schulklassengrösse. Mit den Worten von Hattie (2005): Wenn es um den Einfluss der Schule – d.h. um die Optimierung der Mathematikleistungen geht –, dann kommt es – unabhängig von der Grösse der Schulkasse – wohl eher auf das pädagogische Geschick der Lehrpersonen an.

#### 4.3 Überprüfung der Robustheit vorheriger Schätzergebnisse

Die graphische Darstellung der Befunde im *Regression Discontinuity Design (RDD)* bestätigen die vorherigen Schätzergebnisse (zur Güte des RDD: *Tabelle A.2* im Anhang). Wird – unabhängig vom Schultyp – die polynomische Anpassung der vierten Ordnung für *alle* Schulklassen bis maximal 29 Schüler\*innen betrachtet (*Abbildung 5*), dann zeigt sich zum einen Folgendes: Je mehr sich die tatsächlich beobachtete Klassengrösse der kantonal empfohlenen Klassengrösse («cutoff» c=0) annähert (vgl. *Abbildung 2*), desto besser fällt die Mathematikleistung (WLE) aus. Generell zeigen sich für Schulklassen, die faktisch grösser sind als vorgeschrieben, höhere Testwerte in Mathematik als für unter dem Richtwert liegende Klassengrösse.

Abbildung 5

Verteilung der Mathematikleistungen nach Schulklassengröße (RDD) über alle Schultypen



Daten: ÜGK 2016 – eigene Darstellung

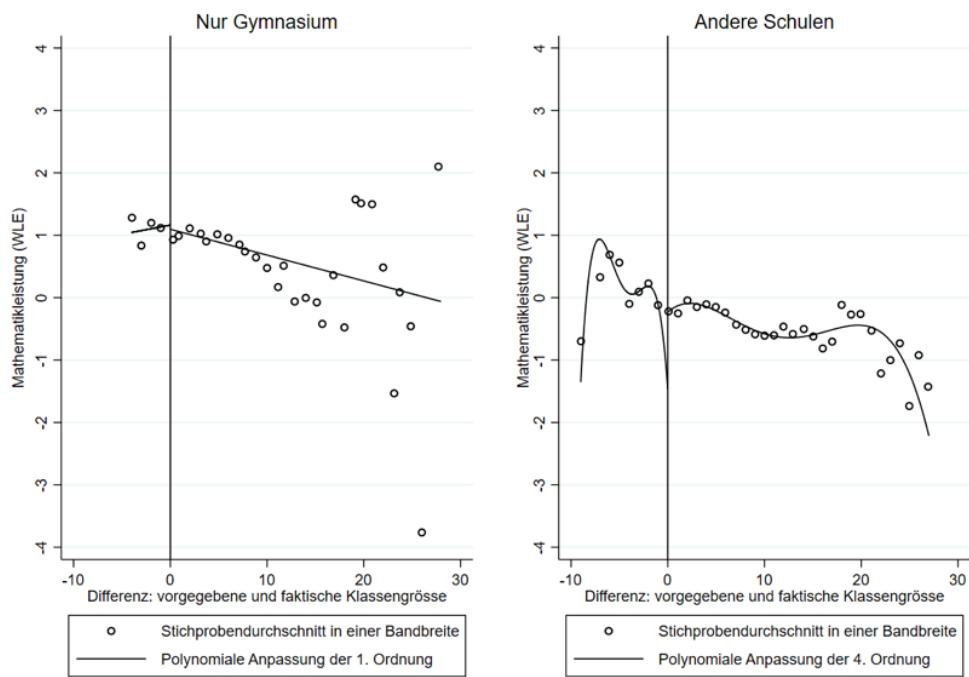
Nach dem Nullpunkt (links vom «cutoff») gibt es einen Sprung in der mittleren Mathematikleistung. Er besagt, dass in noch grösseren Schulklassen die Testwerte noch besser ausfallen. Allerdings ist dieser günstige Wertebereich äusserst klein, in dem sich eine grössere Schulkasse noch positiver auf die Mathematikleistung auswirkt als eine kleinere Schulkasse. So nehmen die Mathematikleistungen ersichtlich ab, wenn die Klassengröße den Richtwert (rechts vom «cutoff») deutlich überschreitet.<sup>3</sup>

Diese Interpretation wird durch die polynomische Anpassung der ersten und vierten Ordnung gestützt, die getrennt nach den Schultypen vorgenommen wird (Abbildung 6). Während die anderen Schultypen in der Sekundarstufe I (Grundanforderung oder erweiterte Anforderung; vierte Ordnung) das zuvor für alle Schultypen beschriebene Muster aufweisen (rechte Grafik), zeigt sich für das Gymnasium (linke Grafik), dargestellt in der ersten Ordnung, dass die Testwerte mit der Klassengröße steigen, dann nach dem «cutoff» einen Sprung nach oben machen, und danach wieder abnehmen. Die Anpassung der Werte an die Regressionsgerade ist am besten für die grösseren Schulklassen.

<sup>3</sup> Die hier geschätzte Bandbreite («mimicking variance evenly-spaced method») ist das „Fenster“ um den Grenzwert, in dem die Ergebnisse der Behandlungs- und Kontrollgruppen verglichen werden. Bei der Auswahl der optimalen Bandbreite in RDD wurde die gängige und robuste Calonico-, Cattaneo- und Titiunik-Methode (CCT) verwendet.

**Abbildung 6**

Verteilung der Mathematikleistungen nach Schulklassengrösse (RDD), getrennt für Schultypen



Daten: ÜGK 2016 – eigene Darstellung

Diese deskriptiven Befunde werden des Weiteren durch multivariate Analysen unterstützt: je grösser die Schulkasse ist, desto besser fallen die Testwerte aus. Je eher die tatsächlich beobachtete Klassengrösse unter der vorgeschriebenen Maximalgrösse liegt, desto geringer sind die Mathematikleistungen (Spalte 3 in *Tabelle A.3* im Anhang). Unter Kontrolle der die Verteilung der Testwerte erklärenden Variablen (wie etwa soziale Herkunft, individuelle Motivation und Leistungsfähigkeit oder Schulklassenmerkmale) zeigt sich, dass die Klassengrösse in Abhängigkeit vom «cutoff» jeweils unterschiedliche Auswirkungen auf die Mathematikleistung haben. Bei Schulklassen, welche unter der vorgegebenen Maximalgrösse für eine Schulkasse liegen (d.h. kleine Schulklassen), sind die Mathematikleistungen am höchsten, wenn sie sich der zulässigen Maximalgrösse annähern (Spalte 1 in *Tabelle A.3*). Für Schulklassen oberhalb der Maximalgrösse steigen die Mathematikleistungen mit zunehmender Klassengrösse – also mit steigender Abweichung von der Maximalgrösse – an, und sinken dann bei grosser Abweichung zusehends wieder. Damit werden die zuvor bivariat beschriebenen und mittels der IV-Regression multivariat geschätzten Befunde gestützt.

#### Rolle der Selektion und Allokation in Schulklassen unterschiedlicher Grösse

Spielt also die Selektion oder Allokation in die unterschiedlich grossen Schulklassen eine Rolle für die Mathematikleistungen der Schüler\*innen? Um diese Frage zu klären, wird als ein weiterer Robustheits- und Reliabilitätstest das *zweistufige Verfahren* von Heckman (1979) verwendet. Die Schätzung der Mathematikleistung als zu erklärendes Phänomen (zweite Stufe) basiert auf dem ersten Modell in *Tabelle 1* (siehe oben) und für das Selektionsmodell (als erste Stufe) wird zunächst die binäre Klassengrösse (1 für 19-29 Schüler\*innen und 0 für kleinere Klassen) als abhängige Variable verwendet (*Tabelle A.4* im Anhang). Bei Kontrolle der sozialen Herkunft anhand des Bildungsniveaus der Eltern, der durchschnittlichen kantonalen Klassengrösse und des lokalen Kontextes der Schule (urban oder rural versus gemischt) ist das *inverse Mills Ratio* positiv signifikant. Die Selektion in die grösseren Schulklassen wirkt sich offenkundig positiv auf die Mathematikleistung aus. Mit kontrafaktischen Worten: die Schülerinnen und Schüler, welche in kleinen Schulklassen sind, würden bessere Mathematikleistungen erzielen, wären sie in grossen Schulklassen.

Dieses Ergebnis wird repliziert, wenn im eigentlichen Schätzmodell nicht die nominale Schulklassengrösse, sondern die *Differenzen zwischen den kantonal vorgegebenen und faktisch in diesen Kantonen beobachteten Klassengrösse* kontrolliert wird. Zum einen zeigt sich ein negativer Zusammenhang zwischen abnehmender

Klassengrösse und der Mathematikleistung (zweite Stufe). Zum anderen indiziert der positive und statistisch signifikante Mills-Quotient, dass die Allokation in grosse Schulklassen der Mathematikleistung zugutekommt. Allerdings gilt auch, dass das sozial selektive «tracking» am Ende der Primarschulzeit in die Schultypen mit jeweils unterschiedlichen Leistungsanforderungen in der unteren Sekundarstufe zur sozialen Ungleichheit von Mathematikleistungen beiträgt (Angelone, 2019).

## 5. Zusammenfassung und Schlussfolgerungen

Ziel der vorliegenden Replikation war, an eine Analyse von Wößmann (2005b) anknüpfend, einen empirischen Beitrag zu der Frage zu leisten, ob – wie in vielen anderen Ländern – auch in der *Schweiz* die schulische Leistung in einem Schulfach wie etwa Mathematik von der Grösse der unterrichteten Schulklasse abhängt. Hintergrund dieser Fragestellung ist die häufig geteilte Annahme, dass «kleine» Schulklassen zu besseren Schulleistungen beitragen (Finn & Achilles, 1990; Hallinan & Sørensen, 1985), während eine Vielzahl von Studien eher zu inkonsistenten Befunden hierzu kommt (Chingos, 2013; Li & Konstantopoulos, 2016; Wößmann, 2005a; Wößmann & West, 2006). So gibt es Belege dafür, dass eher «grosse» Schulklassen – etwa in der Schweiz (Wößmann, 2005b) – zu besseren Leistungen führen; andere Belege legen nahe, dass es *keinen* kausalen Zusammenhang zwischen Klassengrösse und Schulleistung gibt. Vielmehr seien Einflüsse unabhängig von der Grösse einer Schulklasse für die optimale Leistungsentwicklung von Schüler\*innen verantwortlich (Hattie, 2005). Eine nicht unerhebliche Zahl solcher Studien kommt daher zum Ergebnis, dass eine Reduzierung der Klassengrösse ineffektiv ist (z. B. Dobbelen et al., 2002; Hanushek, 2003).

Dieses für andere Länder gültige Ergebnis sollte vor dem Hintergrund einer Studie von Wößmann (2005b) für die *Schweiz* mit aktuellen Daten repliziert werden. Für die empirischen Analysen wurden Daten von *ÜGK* (Überprüfung der Grundkompetenzen in Mathematik) herangezogen, die im Rahmen der nationalen und kantonalen Überprüfung erreichter Bildungsstandards im Jahre 2016 erhoben wurden. Zielpopulation waren Schüler\*innen in der letzten Klassenstufe der Pflichtschulzeit. Gemessen wurde die Mathematikleistung in den 9. Klassen der Regelschule, um beurteilen zu können, ob die geforderten Kompetenzen in diesem Schulfach erreicht wurden. Um das Endogenitäts- und Simultanitätsproblem in Bezug auf den Zusammenhang von Klassengrösse und Schulleistung in einem quasi-experimentellen Design zu überwinden, bot es sich an, bewährte Schätzverfahren wie etwa die *Instrumental Variable Regression* (2SLS) (Shin & Raudenbush, 2011), das *Regression Discontinuity Design* (RDD) (Angrist & Pischke, 2009) oder das *zweistufige Verfahren* von Heckman (1979) zur Korrektur von Stichprobenverzerrungen zu verwenden.

Die bi- und multivariaten Analysen belegen konsistent, dass in Schulklassen mit mehr als 18 und weniger als 30 Schülerinnen und Schülern die besten Leistungen in Mathematik erzielt werden. Während in Gymnasien die Schulklassengrösse eine untergeordnete Rolle spielt, ist es für Sekundarschulen mit Grundanforderungen oder erweiterten Anforderungen vorteilhaft, mindestens 19 Jugendliche gemeinsam zu unterrichten. Werden neben den Sozialstrukturen in den Schulklassen und Schulen auch die individuellen und elterlichen Motivationen sowie Prozesse des Unterrichts in Rechnung gestellt, dann ist der Einfluss der Klassengrösse statistisch insignifikant. Eine bildungspolitische Massnahme, die Klassengrösse zwecks Leistungsoptimierung zu reduzieren, scheint daher ineffektiv. Eher weisen die Analysen darauf hin, dass weniger in die Menge der Lehrpersonen als in deren pädagogische Leistungsfähigkeit zu investieren wäre (Hattie, 2005). Zudem scheint es sinnvoll zu sein, die soziale Segregation in den Schulen und Klassenräumen zu reduzieren (Coleman et al., 1975).

Dieser aktuelle Befund zur Klassengrösse ist nicht einmalig für die Schweiz, sondern wurde bereits zuvor mit älteren Daten aufgedeckt (Wößmann, 2005b). Allerdings ist für die Schweiz hervorzuheben, dass diese Schlussfolgerung vorerst nur für die Sekundarstufe I, wahrscheinlich auch nur für die letzte Klassenstufe in der Pflichtschulzeit und im Schulfach Mathematik, gültig ist. Die hier durchgeführte Analyse müsste für die Klassenstufen im Primarschulbereich und für verschiedene Schulfächer wiederholt repliziert werden.

Die vorliegende Studie weist weitere *Limitationen* auf. Erstens erfolgte die Analyse im komparativ-statistischen Querschnitt. Allerdings ist in Rechnung zu stellen, dass die am Ende der Pflichtschulzeit gemessenen Leistungen das kumulierte Ergebnis mehrjähriger Beschulung sind (Schubert & R. Becker, 2010). Somit fehlen Informationen über die zeitabhängigen Lernraten in unterschiedlichen Schul- und Klassenkontexten. Idealerweise wäre eine Längsschnittanalyse des Bildungs- und Kompetenzverlaufs zu erwägen (Blossfeld et al., 2019). Zeitabhängige Aussetzung in einer Schulkasse mit einer bestimmten Grösse sowie Wechsel zwischen verschiedenen Klassengrössen könnten weitere Varianten für die Prüfung der Hypothese darstellen, dass schulische Leistungen auch von der Grösse der besuchten Schulklassen abhängen. Zweitens liegen hier keine zeitabhängigen Infor-

mationen über den vorherigen Bildungsverlauf der Probanden vor. Zum einen wäre es interessant, zu wissen, welche langfristigen Effekte unterschiedliche Klassengrößen bei einer anderen sozialen Zusammensetzung der Schülerschaft in der Primarstufe auf die schulischen Leistungen haben (Coleman et al., 1966; Fredriksson et al., 2012; Nye et al., 2001). Zum anderen fehlen Informationen über die Lernvoraussetzungen bei der Einschulung sowie bei der Selektion und Allokation in die Schultypen nach dem Ende der Primarstufe. Drittens wurde mit Mathematik nur ein einziges Schulfach unter vielen berücksichtigt. Vorliegende Daten von ÜGK 2017, bei welcher das Erreichen der Grundkompetenzen in der Schulsprache (Lesen und Orthografie) und in der ersten Fremdsprache (Lese- und Hörverstehen) in der 6. Klasse (8. Schuljahr nach HarmoS) überprüft wurde, eignen sich in der vorliegenden Form als «scientific use file» nicht für weiterführende Analysen in der hier behandelten Problemstellung.

### Literatur

- Akerhielm, K. (1995). Does class size matter? *Economics of Education Review*, 14(3), 229–241. [https://doi.org/10.1016/0272-7757\(95\)00004-4](https://doi.org/10.1016/0272-7757(95)00004-4)
- Altinok, N., & Kingdon, G. (2012). New Evidence on Class Size Effects: A Pupil Fixed Effects Approach. *Oxford Bulletin of Economics and Statistics*, 74(2), 203–234. <https://doi.org/10.1111/j.1468-0084.2011.00648.x>
- Angelone, D. (2019). Schereneffekte auf der Sekundarstufe I? Zum Einfluss des Schultyps auf den Leistungszuwachs in Deutsch und Mathematik. *Schweizerische Zeitschrift für Bildungswissenschaften*, 41(2), 446–466. <https://doi.org/10.24452/sjer.41.2.11>
- Angelone, D., & Keller, F. (2019). *Überprüfung des Erreichens der Grundkompetenzen (ÜGK) im Fach Mathematik im 11. Schuljahr. Technische Dokumentation zur Testentwicklung und Skalierung*. Geschäftsstelle der Aufgabendatenbank.
- Angrist, J.D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Angrist, J.D., & Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics*, 114(2), 533–575. <https://doi.org/10.1162/003355399556061>
- Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Prentice Hall.
- Baumert, J., Trautwein, U., & Artelt, C. (2003). Schulumwelten – institutionelle Bedingungen des Lehrens und Lernens. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (S. 259–330). Leske + Budrich.
- Baumert, J., & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In Deutsches PISA-Konsortium (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 323–407). Leske + Budrich.
- Becker, D. (2019). The Social Structure of Schools and Classrooms. In R. Becker (Hrsg.), *Research Handbook on the Sociology of Education* (S. 347–364). EEP. <https://doi.org/10.4337/9781788110426.00029>
- Becker, R. (2022). 'Explaining educational differentials' revisited: an evaluation of rigorous theoretical foundations and empirical findings. In K. Gérhanian, N.-D. De Graaf, & W. Raub (Hrsg.), *Handbook of Sociological Science. Contributions to Rigorous Sociology* (S. 356–371). Edward Elgar Publishing. <https://doi.org/10.4337/9781789909432>
- Becker, R. (2024). Social classes, school performance, subjective expected utilities, and educational decisions: A rigorous application of a Rational Choice Theory for explaining educational differentials in German-speaking parts of Switzerland. *Soziale Welt*, 75(2), 165–197. <https://doi.org/10.5771/0038-6073-2024-2-165>
- Blatchford, P. (2003). *The Class Size Debate: Is Small Better?* Open University Press.
- Blatchford, P., & Mortimore, P. (1994). The Issue of Class Size for Young Children in Schools: What Can We Learn from Research? *Oxford Review of Education*, 20(4), 411–428. <https://doi.org/10.1080/0305498940200402>
- Blatchford, P., Bassett, P., Goldstein, H., & Martin, C. (2003). Are Class Size Differences Related to Pupils' Educational Progress and Classroom Processes? Findings from the Institute of Education Class Size Study of Children aged 5–7 years. *British Educational Research Journal*, 29(5), 709–730. <https://doi.org/10.1080/0141192032000133668>
- Blossfeld, G.J., Blossfeld, P.N., & Blossfeld, H.-P. (2019). A sociological perspective on education as a lifelong process. In R. Becker (Hrsg.), *Research Handbook on the Sociology of Education* (S. 18–34). EEP. <https://doi.org/10.4337/9781788110426.00009>
- Borland, M.V., Howsen, R.M., & Trawick, M.W. (2005). An investigation of the effect of class size on student academic achievement. *Education Economics*, 13(1), 73–83. <https://doi.org/10.1080/0964529042000325216>
- Bourke, S. (1986). How Smaller Is Better: Some Relationships Between Class Size, Teaching Practices, and Student Achievement. *American Educational Research Journal*, 23(4), 558–571.
- Buchmann, M., Kriesi, I., Koomen, M., Imdorf, C., & Basler, A. (2016). Differentiation in Secondary Education and Inequality in Educational Opportunities: The case of Switzerland. In H.-P. Blossfeld, S. Buchholz, J. Skopek, & M. Triventi (Hrsg.), *Models of Secondary Education and Social Inequality: An International Comparison* (S. 111–128). Edward Elgar Publishing.
- Chingos, M.M. (2013). Class Size and Student Outcomes: Research and Policy Implications. *Journal of Policy Analysis and Management*, 32(2), 411–438. <https://doi.org/10.1002/pam.21677>
- Cho, H., Glewwe, P., & Whitler, M. (2012). Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools. *Economics of Education Review*, 31(3), 77–95. <https://doi.org/10.1016/j.econedurev.2012.01.004>
- Coleman, J.S., Campbell, E.Q., Hobson, C.J., Mcpartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). *Equality of Educational Opportunity*. U.S. Government Printing Office.
- Coleman, J.S., Hoffer, T., & Kilgore, S. (1982). *High school achievement: Public, Catholic, and Private Schools Compared*. Basic Books.
- Coleman, J.S., Kelly, S.D., & Moore, J.A. (1975). *Trends in school segregation, 1968–73*. Urban Institute.

- Dobbelsteen, S., Levin, J., & Oosterbeek, H. (2002). The Causal Effect of Class Size on Scholastic Achievement: Distinguishing the pure Class Size Effect from the Effect of Changes in Class Composition. *Oxford Bulletin of Economics and Statistics*, 64(1), 17–38. <https://doi.org/10.1111/1468-0084.00003>
- Ehrenberg, R.G., Brewer, D.J., Gomoran, A., & Willms, J.D. (2001). Class Size and Student Achievement. *Psychological Science in the Public Interest*, 2(1), 1–30. <https://doi.org/10.1111/1529-1006.0003>
- Erikson, R. (2019). How does education depend on social origin? In R. Becker (Hrsg.), *Research Handbook on the Sociology of Education* (S. 35–56). EEP. <https://doi.org/10.4337/9781788110426.00010>
- Finn, J.D., & Achilles, C.M. (1990). Answers and Questions About Class Size: A Statewide Experiment. *American Educational Research Journal*, 27(3), 557–577. <https://doi.org/10.3102/00028312027003557>
- Finn, J.D., & Achilles, C.M. (1999). Tennessee's Class Size Study: Findings, Implications, Misconceptions. *Educational Evaluation and Policy Analysis*, 21(2), 97–109. <https://doi.org/10.3102/01623737021002097>
- Finn, J.D., Pannozzo, G.M., & Achilles, C.M. (2003). The “Why’s” of Class Size: Student Behavior in Small Classes. *Review of Educational Research*, 73(3), 321–368. <https://doi.org/10.3102/00346543073003321>
- Fredriksson, P., Öckert, B., & Oosterbeek, H. (2012). Long-term Effects of Class Size. *The Quarterly Journal of Economics*, 128(1), 249–285. <https://doi.org/10.1093/qje/qjs048>
- Glass, G.V., & Smith, M.L. (1979). Meta-Analysis of Research on Class Size and Achievement. *Educational Evaluation and Policy Analysis*, 1(1), 2–19. <https://doi.org/10.3102/01623737001001002>
- Hallinan, M.T., & Sørensen, A.B. (1985). Class Size, Ability Group Size, and Student Achievement. *American Journal of Education*, 94(1), 71–89.
- Hanushek, E.A. (1986). The Economics of Schooling: Production and Efficiency in Public Schools. *Journal of Economic Literature*, 24(3), 1141–1177. <https://doi.org/10.1086/443832>
- Hanushek, E.A. (1999). Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis*, 21(1), 143–163. <https://doi.org/10.3102/01623737021002143>
- Hanushek, E.A. (2003). Evidence, Politics, and the Class Size Debate. In L. Mishel & R. Rothstein (Hrsg.), *The Class Size Debate* (S. 37–66). Economic Policy Institute.
- Hanushek, E.A., & Woessmann, L. (2017). School Resources and Student Achievement: A Review of Cross-Country Economic Research. In M. Rosén, M., K. Yang Hansen, & U. Wolff (Hrsg.), *Cognitive Abilities and Educational Outcomes. Methodology of Educational Measurement and Assessment*. Springer. [https://doi.org/10.1007/978-3-319-43473-5\\_8](https://doi.org/10.1007/978-3-319-43473-5_8)
- Hascher, T., Brühwiler, C., Erzinger, A., Girnat, B., & Hagenauer, G. (2015). *Erläuterungen zu den Skalen des Kontextfragebogens Mathematikteil: Theoretischer Hintergrund und Forschungsinteressen. Überarbeitung der Pilotierung*. o.O.
- Hattie, J. (2005). The paradox of reducing class size and improving learning outcomes. *International Journal of Educational Research*, 43(6), 387–425. <https://doi.org/10.1016/j.ijer.2006.07.002>
- Heckman, J.J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1), 153–161. <https://doi.org/10.2307/1912352>
- Hoxby, C.M. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *The Quarterly Journal of Economics*, 115(4), 1239–1285. <https://doi.org/10.1162/003355300555060>
- Hupka-Brunner, S., Jann, B., Meyer, T., Imdorf, C., Sacchi, S., Müller, B., Scharenberg, K., Rotz, C.V., Koomen, M., & Becker, R. (2016). *Erläuterungen zum Kontextfragebogen der ÜGK 2016: Allgemeiner Teil*. Universität Bern.
- Jackson, M. (Hrsg.) (2013). *Determined to Succeed? Performance versus Choice in Educational Attainment*. Stanford University Press. <https://doi.org/10.11126/stanford/9780804783026.001.0001>
- Jann, B., & Hupka-Brunner, S. (2019). Warum werden Frauen so selten MINT-Fachkräfte? Zur Bedeutung der Differenz zwischen mathematischen Kompetenzen und Selbstkonzept. TREE.
- Jungbauer-Gans, M. (2004). Einfluss des sozialen und kulturellen Kapitals auf die Lesekompetenz. Ein Vergleich der PISA 2000-Daten aus Deutschland, Frankreich und der Schweiz. *Zeitschrift für Soziologie*, 33(5), 375–397. <https://doi.org/10.1515/zfsoz-2004-0502>
- Konsortium ÜGK (Hrsg.) (2019). *Überprüfung der Grundkompetenzen. Nationaler Bericht der ÜGK 2016: Mathematik 11. Schuljahr*. EDK und SRED. <https://doi.org/10.18747/PHSGcoll3/id/386>
- Konstantopoulos, S. (2008). Do Small Classes Reduce the Achievement Gap between Low and High Achievers? Evidence from Project STAR. *The Elementary School Journal*, 108(4), 275–291. <https://doi.org/10.1086/528972>
- Krueger, A.B. (1999). Experimental Estimates of Education Production Functions. *The Quarterly Journal of Economics*, 114(2), 497–532. <https://doi.org/10.1162/003355399556052>
- Krueger, A.B. (2003). Economic Considerations and Class Size. *The Economic Journal*, 113(485), 34–63. <https://doi.org/10.1111/1468-0297.00098>
- Krueger, A.B., & Whitmore, D.M. (2001). The Effect of Attending a Small Class in the Early Grades on College-test Taking and Middle School Test Results: Evidence from Project STAR. *The Economic Journal*, 111(468), 1–28. <https://doi.org/10.1111/1468-0297.00586>
- Lee, D. S., & Lemieux, T. (2015). Regression discontinuity designs in social sciences. In H. Best. & C. Wolf (Hrsg.), *Regression Analysis and Causal Inference* (S. 301–326). Sage.
- Leuven, E., Oosterbeek, H., & Ronning, M. (2008). Quasi-experimental Estimates of the Effect of Class Size on Achievement in Norway. *The Scandinavian Journal of Economics*, 110(4), 663–693. <https://doi.org/10.1111/j.1467-9442.2008.00556.x>
- Li, W., & Konstantopoulos, S. (2016). Class Size Effects on Fourth-Grade Mathematics Achievement: Evidence from TIMSS 2011. *Journal of Research on Educational Effectiveness*, 9(4), 503–530. <https://doi.org/10.1080/19345747.2015.1105893>
- Mitchell, D.E., Beach, S.A., & Badarak, G. (1989). Modeling the Relationship between Achievement and Class Size: A Re-analysis of the Tennessee Project STAR Data. *Peabody Journal of Education*, 67(1), 34–74. <https://doi.org/10.1080/01619569209538669>

- Mosteller, F. (1995). The Tennessee Study of Class Size in the Early School Grades. *Future of children*, 5(2), 113–127. <https://doi.org/10.1515/9781400851607.261>
- Muller, C., Winship, C., & Morgen, S.L. (2015). Instrumental Variables Regression. In H. Best, & C. Wolf (Hrsg.), *Regression Analysis and Causal Inference* (S. 277–299). Sage.
- Murnane, R.J., & Willett, J.B. (2011). *Methods matter. Improving Causal Inference in Educational and Social Science Research*. Oxford University Press.
- Nidegger, C. (2019). ÜGK / COFO / VECOF 2016: Competencies of Swiss pupils in mathematics (Version 1.0.0) [Data set]. FORS data service. <https://doi.org/10.23662/FORS-DS-1004-1>
- Nye, B., Hedges, L.V., & Konstantopoulos, S. (2001). The Long-Term Effects of Small Classes in Early Grades: Lasting Benefits in Mathematics Achievement at Grade 9. *The Journal of Experimental Education*, 69(3), 245–257. <https://doi.org/10.1080/00220970109599487>
- Nye, B., Hedges, L.V., & Konstantopoulos, S. (2002). Do Low-Achieving Students Benefit More from Small Classes? Evidence from the Tennessee Class Size Experiment. *Educational Evaluation and Policy Analysis*, 24(3), 201–217. <https://doi.org/10.3102/01623737024003201>
- Pedder, D. (2006). Are small classes better? Understanding relationships between class size, classroom processes and pupils' learning. *Oxford Review of Education*, 32(2), 213–234.
- Pham, G., Helbling, L., Verner, M., Petrucci, F., Angelone, D., & Ambrosetti, A. (2019). *ÜGK – COFO – VeCoF 2016 results: Technical appendices*. Pädagogische Hochschule St. Gallen (PHSG) & Service de la recherche en éducation (SRED). [www.cofo-suisse.ch/cofo-2016](http://www.cofo-suisse.ch/cofo-2016)
- Pötter, U., & Blossfeld, H. P. (2001). Causal Inference from Series of Events. *European Sociological Review*, 17(1), 21–32. <https://doi.org/10.1093/esr/17.1.21>
- Pong, S.-I., & Pallas, A. (2001). Class Size and Eighth-Grade Math Achievement in the United States and Abroad. *Educational Evaluation and Policy Analysis*, 23(3), 251–273. <https://doi.org/10.3102/01623737023003251>
- Raub, W., De Graaf, N. D., & Gérxhani, K. (2022). Rigorous Sociology. In K. Gérxhani, N.-D. De Graaf, & W. Raub (Hrsg.), *Handbook of Sociological Science. Contributions to Rigorous Sociology* (S. 2–19). Edward Elgar Publishing. <http://dx.doi.org/10.4337/9781789909432>
- Rivkin, S.G., Hanushek, E.A., & Kain, J.F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2), 417–458. <https://doi.org/10.1111/j.1468-0262.2005.00584.x>
- Schanzenbach, D. W. (2006). What Have Researchers Learned from Project STAR? *Brookings Papers on Education Policy*, 9(2006/2007), 205–228. <https://doi.org/10.1353/pep.2007.0007>
- Schubert, F., & Becker, R. (2010). Social Inequality of Reading Literacy. A longitudinal analysis with cross-sectional data of PIRLS 2001 and PISA 2000 utilizing the pair wise matching procedure. *Research in Social Stratification and Mobility*, 28(1), 109–133. <https://doi.org/10.1016/j.rssm.2009.12.007>
- Schweizerische Konferenz der kantonalen Erziehungsdirektoren. (2024). *Klassengrößen*. <https://www.edk.ch/de/bildungssystem/kantonale-schulorganisation/kantonsumfrage/a3-klassengroessen>
- Shen, T., & Konstantopoulos, S. (2017). Class size effects on reading achievement in Europe: Evidence from PIRLS. *Studies in Educational Evaluation*, 53, 98–114. <https://doi.org/10.1016/j.stueduc.2017.04.001>
- Shin, Y., & Raudenbush, S.W. (2011). The Causal Effect of Class Size on Academic Achievement: Multivariate Instrumental Variable Estimators with Data Missing at Random. *Journal of Educational and Behavioral Statistics*, 36(2), 154–185. <https://doi.org/10.3102/1076998610388632>
- Slavin, R. (1990). Class Size and Student Achievement: Is smaller better? *Contemporary Education*, 62(1), 6–12.
- Smith, M. L., & Glass, G. V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal*, 17(4), 419–433.
- Sohn, K. (2016). A review of research on Project STAR and path ahead. *School Effectiveness and School Improvement*, 27(2), 116–134. <https://doi.org/10.1080/09243453.2014.994643>
- Verner, M., & Helbling, L. (2019). *Sampling ÜGK 2016. Technischer Bericht zu Stichprobendesign, Gewichtung und Varianzschätzung bei der Überprüfung des Erreichens der Grundkompetenzen 2016*. Institut für Bildungsevaluation.
- Weiß, M. (2011). *Allgemeinbildende Privatschulen in Deutschland*. Friedrich-Ebert-Stiftung.
- Wößmann, L. (2005). Leistungsfördernde Anreize für das Schulsystem. *ifo Schnelldienst*, 58(19), 18–27.
- Wößmann, L. (2005). Educational Production in Europe. *Economic Policy*, 20(43), 446–504. <https://doi.org/10.1111/j.1468-0327.2005.00144.x>
- Wößmann, L., & West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, 50(3), 695–736. <https://doi.org/10.1016/j.eurocorev.2004.11.005>
- Zahner, C., Meyer, A.H., Moser, U., Brühwiler, C., Vellacott, M. C., Huber, M., Malti, T., Ramseier, E., Wolter, S.C., & Zutavern, M. (2002). *Für das Leben gerüstet? Die Grundkompetenzen der Jugendlichen – Nationaler Bericht der Erhebung PISA 2000*. Bundesamt für Statistik (BFS).
- Zanger, C., & Becker, R. (2019). Experiments in the Sociology of Education: Causal Inference and Estimating Causal effects in Sociological Research on Education. In R. Becker (Hrsg.), *Research Handbook on the Sociology of Education* (S. 153–171). EEP. <https://doi.org/10.4337/9781788110426>

**Schlagworte:** Klassengröße, Mathematikleistung, ÜGK 2016, Instrumental Variable Regression (2SLS), Regression Discontinuity Design (RDD), Schweiz

## **Une petite taille de classe dans une école est-elle vraiment importante ? Une réPLICATION empirique pour la Suisse sur le lien entre la taille des classes et les performances scolaires au premier cycle de l'enseignement secondaire avec les données de l'enquête COFO**

### **Résumé**

Pour la Suisse, l'étude vise à déterminer si la taille de la classe a une influence causale sur les performances en mathématiques. À cette fin, les données de COFO 2016, une enquête nationale sur les performances scolaires concernant les compétences de base des élèves de dernière année de la scolarité obligatoire dans le domaine des mathématiques, sont utilisées. Premièrement, les analyses bivariées suggèrent que les classes plus grandes plutôt que plus petites sont associées à des résultats plus élevés aux tests de mathématiques. En utilisant des méthodes d'estimation multivariée adaptées à la détection des effets causaux de la taille des classes, ce résultat est validé tout en contrôlant les influences concurrentes.

**Mots-clés :** Taille des classes, performance en mathématiques, COFO 2016, régression des variables instrumentales (2SLS), plan de discontinuité de régression (RDD), Suisse

## **Le classi di piccole dimensioni contano davvero? Una replica empirica per la Svizzera sulla relazione tra dimensioni delle classi e rendimento scolastico nella scuola secondaria inferiore utilizzando i dati VeCoF**

### **Riassunto**

Questo studio esamina se le dimensioni delle classi scolastiche abbiano un'influenza causale sul rendimento in matematica in Svizzera. A tal fine, vengono utilizzati i dati di VeCoF 2016 uno studio nazionale sul rendimento scolastico sulle competenze di base degli studenti dell'ultimo anno della scuola dell'obbligo in matematica. In primo luogo, analisi bivariate suggeriscono che classi scolastiche più numerose, piuttosto che più piccole, siano associate a punteggi più elevati nei test di matematica. Utilizzando metodi di stima multivariata adatti a scoprire gli effetti causali delle dimensioni delle classi scolastiche, questo risultato viene convalidato controllando al contempo le influenze concorrenti. In Svizzera, almeno per le scuole secondarie inferiori, non è necessario ridurre le dimensioni delle classi al fine di migliorare il rendimento scolastico degli studenti.

**Parole chiave:** Dimensione della classe, rendimento in matematica, VeCoF 2016, regressione con variabili strumentali (2SLS), disegno di discontinuità della regressione (RDD), Svizzera

## **Does the small class size really matter? An empirical replication for Switzerland on the relationship between class size and school performance in lower secondary school using ÜGK/COFO/VeCoF data**

### **Abstract**

This study examines whether the size of the school class has a causal influence on mathematics performance in Switzerland. For this purpose, data from ÜGK/COFO 2016 (Verification of the Attainment of Basic Competencies), a national school performance study on the basic skills of students in the last grade of compulsory schooling in mathematics, are used. Firstly, bivariate analyses suggest that larger rather than smaller school classes are associated with higher test scores in mathematics. Using multivariate estimation methods that are suitable for uncovering causal effects of school class size, this finding is validated while controlling competing influences. In Switzerland, at least for lower secondary schools, there is no need to reduce class sizes with the aim of improving academic performance among students.

**Keywords:** Class size, mathematics performance, ÜGK/COFO/VeCoF 2016, Instrumental Variable Regression (2SLS), Regression Discontinuity Design (RDD), Switzerland

## Anhang

Tabelle A.1

Mathematikleistung (WLE) nach Schulklassengröße am Ende der Pflichtschulzeit (9. Klasse)

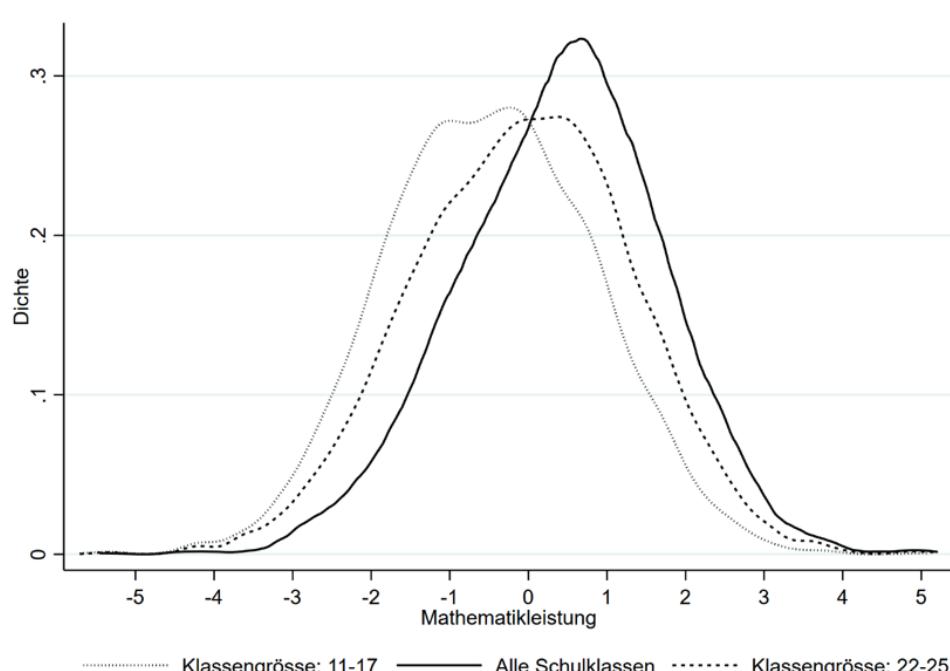
	Insgesamt	Gymnasium	Andere Schultypen
<i>Schulklassengröße</i>			
Grösse der Schulklasse	0,4759661 (0,165)**	0,2221081 (0,137)	0,4349946 (0,209)*
Grösse der Schulklasse <sup>2</sup>	-0,0124895 (0,005)*	-0,0023517 (0,004)	-0,0116138 (0,006)*
<i>Sozioökonomischer Status</i>			
Individualebene	0,005 (0,001)***	0,006 (0,001)***	0,004 (0,001)***
Schulklassenebene	0,032 (0,002)***	0,012 (0,004)**	0,029 (0,002)***
Schulebene	0,015 (0,004)***	0,014 (0,005)**	-0,004 (0,005)
Kantonalebene	0,139 (0,012)***	0,142 (0,018)***	0,144 (0,012)***
<i>Konstante</i>	-20,391 (2,190)***	-18,270 (2,597)***	-19,428 (2,021)***
Fallzahl	21.150	4.873	15.560
Population	75.930	20.158	52.510
F-Wert (6, 114)	161,88***	43,81***	67,21***
Optimale Schulklassengröße	19	47	19

\* p &lt; 0,05; \*\* p &lt; 0,01; \*\*\* p &lt; 0,001 (in Klammern: Standardfehler der Regressionskoeffizienten).

Daten: ÜGK 2016 – eigene Berechnung

Abbildung A.1

Replikation des Tennessee STAR Project: Mathematikleistung (WLE) nach Schulklassengröße



**Tabelle A.2**Robustheitstest mit RDD: Mathematikleistung nach Schulklassengröße<sup>1</sup>

	Insgesamt	Gymnasium	Gymnasium <sup>2</sup>	Andere Schule	Andere Schule <sup>2</sup>
<i>Scharfe RD-Analyse</i>					
Konventionell	-0,066 (0,105)	-0,112 (0,170)	-0,038 (0,129)	0,229 (0,179)	0,144 (0,148)
Bias-korrigiert	-0,040 (0,105)	0,327 (0,170)	-0,065 (0,129)	0,587 (0,179)**	0,245 (0,148)
Robust	-0,040 (0,176)	0,327 (0,402)	-0,065 (0,243)	0,587 (0,282)*	0,245 (0,228)
<i>rho</i>	0,656	0,675	0,830	0,564	0,637
<i>Unscharfe RD-Analyse</i>					
1. Schritt					
Konventionell	0,319 (0,171)	-0,147 (0,201)	-0,077 (0,211)	1,167 (0,234)***	1,154 (0,285)***
Bias-korrigiert	0,248 (0,171)	0,845 (0,201)***	-0,705 (0,211)***	1,122 (0,234)***	1,067 (0,285)***
Robust	0,248 (0,284)	0,845 (0,343)*	-0,705 (0,341)*	1,122 (0,455)**	1,067 (0,448)**
2. Schritt					
Konventionell	-0,225 (0,376)	1,431 (2,240)	0,480 (2,167)	0,198 (0,158)	0,125 (0,135)
Bias-korrigiert	-0,207 (0,376)	-5,640 (2,240)*	-3,162 (2,167)	0,519 (0,158)**	0,220 (0,135)
Robust	-0,207 (0,626)	-5,640 (3,783)	-3,162 (3,652)	0,519 (0,260)*	0,220 (0,208)
Fallzahl	21.502	4.923	4.309	15.808	13.515
Fallzahl, links von cutoff	1.383	459	400	751	632
Fallzahl, rechts von cutoff	20.119	4.464	3.909	15.057	12.883
<i>rho</i>	0,659	0,879	0,716	0,558	0,667

<sup>\*</sup> p < 0,05; \*\* p < 0,01; \*\*\* p < 0,001; Regression Discontinuity Design (RDD), in Klammern: Standardfehler der Koeffizienten.<sup>1</sup> Schätzungen der Behandlungseffekte: (1) Ergebnis: Mathematikleistung (WLEM). (2) Laufende Variable: Differenz von vorgegebener und tatsächlicher Schulklassengröße mit *cutoff* von 0 (Nullpunkt). (3) Behandlungsstatus: beobachtete Schulklassengröße.<sup>2</sup> Schätzung mit Kovariaten (wie in Tabelle 1: Modelle 1, 3 und 4).

Daten: ÜGK 2016 – eigene Berechnung

**Tabelle A.3***Mathematikleistung nach Schulklassengröße (RDD)*

	Modelle	Kleine Schulkasse	Grosse Schulkasse	Insgesamt
<i>Schulklassengröße</i>				
Grösse der Schulkasse	−0,065 (0,018)***	0,376 (0,162)*		
Grösse der Schulkasse <sup>2</sup>	0,003 (0,001)***	−0,007 (0,003)*		
Vorgegebene–gemessene Klassengröße			−0,020 (0,003)***	
<i>Individuelle Merkmale</i>				
Geschlecht (Referenz: männlich)	−0,130 (0,019)***	−0,125 (0,067)	−0,132 (0,019)***	
Sozialer Status	0,006 (0,001)***	0,004 (0,002)*	0,006 (0,001)***	
Migration: 2. Generation (vs. Einh.)	−0,366 (0,029)***	−0,229 (0,069)**	−0,354 (0,029)***	
Migration: 1. Generation (vs. Einh.)	−0,262 (0,042)***	−0,133 (0,097)	−0,249 (0,040)***	
Testsprache zu Hause gesprochen	0,184 (0,048)***	0,057 (0,116)	0,177 (0,044)***	
<i>Schulnoten</i>				
Mathematik	0,108 (0,023)***	0,036 (0,067)	0,099 (0,024)***	
Naturwissenschaften	0,148 (0,020)***	0,189 (0,040)***	0,167 (0,018)***	
<i>Leistungsmerkmale für Mathematik</i>				
Selbstkonzept	0,329 (0,016)***	0,380 (0,049)***	0,338 (0,016)***	
Private Unterstützung in Mathe	−0,285 (0,029)***	−0,250 (0,075)**	−0,281 (0,027)***	
Klassenwiederholung	−0,338 (0,030)***	−0,223 (0,078)**	−0,341 (0,027)***	
Idealistische Bildungsaspiration	0,441 (0,025)***	0,350 (0,071)***	0,447 (0,023)***	
<i>Schulklassenmerkmale</i>				
Anteil der Frauen	0,475 (0,140)***	−0,203 (0,231)	0,427 (0,119)***	
<i>Schulmerkmale</i>				
Sozialer Status der Schülerschaft	0,012 (0,002)***	0,016 (0,004)***	0,012 (0,002)***	
Anteil der Migranten	−0,612 (0,122)***	0,139 (0,209)	−0,577 (0,111)***	
Anzahl der Schulklassen (9. Klasse)	0,016 (0,004)***	0,030 (0,009)***	0,023 (0,004)***	
Gymnasium (vs. andere Schultypen)	0,653 (0,041)***	0,554 (0,079)***	0,695 (0,033)***	
Konstante	−2,436 (0,233)***	−7,087 (1,977)***	−2,501 (0,196)***	
Fallzahl	16.078	1.695	17.773	
Population	54.699	8.254	62.954	
Korrigiertes R <sup>2</sup>	0,4442	0,4520	0,4410	
F-Wert (#Koeffizienten, Freiheitsgrad)	330,13 (18, 102)	69,36 (18, 102)	381,71 (17, 103)	

\* p < 0,05; \*\* p < 0,01; \*\*\* p < 0,001; OLS-Regression mit BRR mit 120 Replikationen (in Klammern: Standardfehler der Koeffizienten).

Daten: ÜGK 2016 – eigene Berechnung

Tabelle A.4

Robustheitstest: Mathematikleistung (WLE) nach Selektion in die Schulklassengröße

Variablen für das Hauptmodell	Modell 1		Modell 2		
	Effekt	Selektion	Effekt	Selektion	Selektionsmodell
<i>Schulklassengröße</i>					
Grösse der Schulklasse	-0,018 (0,074)	0,220 (0,007)***		0,220 (0,007)***	<i>Schulklassengröße</i> Kantonale Klassengröße
Grösse der Schulklasse <sup>2</sup>	0,001 (0,002)				
Vorgegebene-empirische Grösse			-0,012 (0,004)***		
<i>Individuelle Merkmale</i>					
Geschlecht (Referenz: männlich)	-0,160 (0,023)***	0,328 (0,028)***	-0,160 (0,023)***	0,328 (0,028)***	<i>Elterliches Bildungsniveau</i> Mittleres vs. tiefes Niveau
Sozialer Status	0,004 (0,001)***	0,580 (0,029)***	0,004 (0,001)***	0,580 (0,029)***	Hohes vs. tiefes Niveau
Migration: 2. Generation (vs. Einh.)	-0,280 (0,033)***		-0,277 (0,033)***		
Migration: 1. Generation (vs. Einh.)	-0,134 (0,043)**		-0,134 (0,043)**		
Testsprache zu Hause gesprochen	0,082 (0,044)		0,079 (0,044)		
<i>Schulnoten</i>					
Mathematik	0,116 (0,023)***		0,118 (0,023)***		
Naturwissenschaften	0,165 (0,019)***		0,166 (0,019)***		
<i>Leistungsmerkmale für Mathematik</i>					
Selbstkonzept	0,340 (0,016)***		0,339 (0,016)***		
Private Unterstützung in Mathe	-0,239 (0,027)***		-0,239 (0,027)***		
Klassenwiederholung	-0,287 (0,031)***		-0,290 (0,031)***		
Idealistische Bildungsaspiration	0,373 (0,025)***		0,370 (0,025)***		
<i>Schulklassenmerkmale</i>					
Anteil der Frauen	0,190 (0,095)*		0,189 (0,095)*		
<i>Schulmerkmale</i>					
Sozialer Status der Schülerschaft	0,018 (0,002)***	0,044 (0,022)*	0,018 (0,002)***	0,044 (0,022)*	<i>Schulischer Kontext</i> Urban vs. gemischt
Anteil der Migranten	-0,341 (0,076)***	-0,261 (0,033)***	-0,386 (0,077)***	-0,261 (0,033)***	Ländlich vs. gemischt
Anzahl der Schulklassen (9. Klasse)	0,024 (0,003)***		0,023 (0,003)***		
Gymnasium (vs. andere Schulen)	0,610 (0,028)***		0,626 (0,027)***		
Konstante	-2,769 (0,847)**	-4,560 (0,126)***	-2,715 (0,175)***	-4,560 (0,126)***	<i>Konstante</i>
		0,168 (0,049)***		0,136 (0,049)**	$\Lambda$ (Inverses Mills Ratio)

Fallzahl	19.458	19.458
Selektierte Fälle	8.352	8.352
Wald-Chi <sup>2</sup> (d.f.)	5885,63 (18)	5892,04 (17)
Selektion ins Gymnasium in %	42,9%	42,9%

\* p < 0,05; \*\* p < 0,01; \*\*\* p < 0,001; Heckman-Korrektur für WLEM (in Klammern: Standardfehler der Koeffizienten).

Daten: ÜGK 2016 – eigene Berechnung

**Rolf Becker**, Prof. Dr. ist Ordinarius für Bildungssoziologie und Direktor der Abteilung Bildungssoziologie an der Universität Bern (seit 2004). Studium in Soziologie, Sozialpsychologie und Politikwissenschaft an der Universität Mannheim (1981–1987; Diplom); Promotion (1988–1990; Max-Planck-Institut für Bildungsforschung in Berlin); Habilitation (1999/2000; Technische Universität Dresden). Forschungsinteressen: Lebensverlaufsforschung; Bildungssoziologie; Mobilitätsforschung; Soziale Ungleichheit; Methoden empirischer Sozialforschung.

Universität Bern, Fabrikstrasse 8, CH-3012 Bern

E-Mail: rolf.becker@unibe.ch