Universität Zürich[UZH]

Institut für Bildungsevaluation
Assoziiertes Institut der Universität Zürich

Sampling ÜGK 2016

Technischer Bericht zu Stichprobendesign, Gewichtung und Varianzschätzung bei der Überprüfung des Erreichens der Grundkompetenzen 2016

Martin Verner

Laura Helbling

Zürich, Mai 2019

Anschrift

Institut für Bildungsevaluation
Assoziiertes Institut der Universität Zürich
Wilfriedstrasse 15
8032 Zürich

Tel.: 043 268 39 62
Fax: 043 268 39 67
www.ibe.uzh.ch

martin.verner@ibe.uzh.ch

Inhaltsverzeichnis

Liste der im Text verwendeten Abkürzungen

BRR:	Balanced Repeated Replication
HarmoS:	Interkantonale Vereinbarung über die Harmonisierung der obligatorischen Schule
MOS:	Measure of Size
NRA:	Non-Response Adjustment
RN:	Random Number
SI:	Sampling Interval
TCS:	Target Cluster Size
ÜGK:	Überprüfung des Erreichens der Grundkompetenzen

Liste der im Text verwendeten Kantonskürzel

AG:	Kanton Aargau
AI:	Kanton Appenzell Innerrhoden
AR:	Kanton Appenzell Ausserrhoden
BE_d:	Deutschsprachiger Teil des Kantons Bern
BE_f:	Französischsprachiger Teil des Kantons Bern
BL:	Kanton Basel-Landschaft
BS:	Kanton Basel-Stadt
FR_d:	Deutschsprachiger Teil des Kantons Freiburg
FR_f:	Französischsprachiger Teil des Kantons Freiburg
GE:	Kanton Genf
GL:	Kanton Glarus
GR:	Kanton Graubünden
JU:	Kanton Jura
LU:	Kanton Luzern
NE:	Kanton Neuenburg
NW:	Kanton Nidwalden
OW:	Kanton Obwalden
SG:	Kanton Sankt Gallen
SH:	Kanton Schaffhausen
SO:	Kanton Solothurn
SZ:	Kanton Schwyz
TG:	Kanton Thurgau
TI:	Kanton Tessin
UR:	Kanton Uri
VD:	Kanton Waadt
VS_d:	Deutschsprachiger Teil des Kantons Wallis
VS_f:	Französischsprachiger Teil des Kantons Wallis
ZG:	Kanton Zug
ZH:	Kanton Zürich

1 Einleitung

Im Frühjahr 2016 fand die erste Erhebung zur Überprüfung des Erreichens der Grundkompetenzen (ÜGK) statt. Dabei wurde auf nationaler und kantonaler Ebene analysiert, inwieweit Schülerinnen und Schüler am Ende des 11. Schuljahres (Definition nach HarmoS) die im Rahmen von HarmoS definierten Grundkompetenzen (Nationale Bildungsziele) in Mathematik erreichen.

Landesweit umfasste die interessierende Population knapp 85'000 Schülerinnen und Schüler, die in rund 1'500 Schulen unterrichtet wurden. Da die Überprüfung sämtlicher unterrichteter Schülerinnen und Schüler des 11. Schuljahres mit einem unverhältnismässig hohen Aufwand verbunden gewesen wäre, wurden teilweise Schul- oder Schülerstichproben gebildet. Der vorliegende Bericht dokumentiert die eingesetzten Stichprobenverfahren. Da sich Schulsysteme und verfügbare Informationen, die zur Ziehung einer Stichprobe notwendig sind, von Kanton zu Kanton stark unterscheiden, werden verschiedene Vorgehensweisen vorgestellt.

1.1 Vollerhebungen, einstufige und zweistufige Stichprobenverfahren

Die 26 Kantone bzw. 29 sprachregionalen Kantonsteile[1] lassen sich anhand von zwei verschiedenen Stichprobenverfahren in Gruppen einteilen:

- In den Kantonen AI, AR, BE_f, GL, JU, NW, OW, SH, UR, VS_d sowie ZG wurden keine Stichproben gezogen. Sämtliche Schulen, die eine 11. Klasse führen, wurden kontaktiert und alle der Populationsdefinition entsprechenden Schülerinnen und Schüler zur Teilnahme aufgeboten. Da aus diversen Gründen einzelne Schulen bzw. Schülerinnen und Schüler nicht an der Erhebung teilnahmen, sind die in den Abschnitten 5.3 und 0 beschriebenen Verfahren zur Korrektur der Stichprobengewichte auch für diese Kantone relevant. In späteren Kapiteln wird diese Gruppe als *Kantone mit Vollerhebung* bezeichnet.
- In den Kantonen BL, BS, FR_d, FR_f, GE, GR, NE, SO, SZ, TG, TI sowie VS_f wurde ein Stichprobenverfahren auf Schülerebene eingesetzt. Das heisst, dass alle Schulen, die eine 11. Klasse führen, zur Teilnahme aufgeboten wurden, innerhalb dieser Schulen jedoch ein bestimmter Anteil der Schülerinnen und Schüler gezogen wurde. Die Details zur Vorgehensweise bei der Ziehung von Schülerstichproben kann Kapitel 4 entnommen werden. Diese Gruppe wird

[1] Im weiteren Verlauf des vorliegenden Dokuments werden mit dem Begriff «Kanton» sprachregionale Teile der Kantone BE, FR und VS miteinbezogen.

im späteren Verlauf der Dokumentation als *Kantone mit einstufigen Stichprobenverfahren* bezeichnet.

- In den Kantonen AG, BE_d, LU, SG, VD sowie ZH wurde ein zweistufiges Stichprobenverfahren eingesetzt. Das bedeutet, dass nicht sämtliche Schulen dieser Kantone zwecks Erhebung kontaktiert wurden, sondern in einem ersten Schritt ein Stichprobenverfahren zur Ziehung von Schulen zum Einsatz kam. Die Methodik zu diesem Verfahren wird in Kapitel 3 beschrieben. In einem zweiten Schritt wurden analog zu *Kantonen mit einstufigen Stichprobenverfahren* innerhalb der teilnehmenden Schulen Schülerinnen und Schüler gezogen, weshalb die Ausführungen in Kapitel 4 für die hier beschriebenen Kantone ebenfalls ihre Gültigkeit haben. Diese Kantone werden fortan als *Kantone mit zweistufigen Stichprobenverfahren* bezeichnet.

1.2 Stichprobenfehler, Klumpen- und Designeffekte

Wird der Anteil Schülerinnen und Schüler, die den definierten Grundkompetenzen in Mathematik nicht genügen, auf Basis von Stichproben geschätzt, dann kann dieser Anteil in Abhängigkeit der gezogenen Schulen bzw. Schülerinnen und Schüler variieren. In anderen Worten: Würde der Anteil mit verschiedenen Stichproben geschätzt, dann wäre mit einer bestimmten Varianz im geschätzten Anteil zu rechnen (Stichprobenvarianz). Die Streuung dieser Anteilsschätzung widerspiegelt den Stichprobenfehler bzw. Standardfehler, mit dem auf Stichproben beruhende Schätzungen behaftet sind. Dieser Fehler wird in erster Linie von der Varianz des Zielmerkmals und vom Stichprobenumfang bestimmt.[2] Je kleiner die Varianz des Merkmals und je grösser die Stichprobe, umso kleiner fällt der Stichprobenfehler aus. Die Methoden zur Berechnung dieser Fehler bei der ÜGK 2016 sind Kapitel 6 zu entnehmen.

Werden mehrstufige Stichprobenverfahren eingesetzt, besteht im Bildungskontext das Problem, dass sich die Individuen bezüglich des zu messenden Merkmals nicht zufällig auf die Auswahleinheiten der ersten Stufe (z.B. Schulen oder Schulklassen) verteilen. So sind sich Schülerinnen und Schüler innerhalb einer Schule in diversen Merkmalen (z.B. der Leistung in Mathematik) ähnlicher als Schülerinnen und Schüler

[2] Im vorliegenden Fall handelt es sich beim Zielmerkmal um den Anteil Schülerinnen und Schüler, die den Grundkompetenzen nicht genügen. Wie der folgenden Formel entnommen werden kann, wird dabei die Varianz dann maximal, wenn sich der Anteil 0.5 nähert (p entspricht dem zu schätzenden Anteil, n dem Stichprobenumfang und σ dem Standardfehler; Kauermann & Küchenhoff, 2011):

$$\sigma_n = \sqrt{\frac{p(1-p)}{n}}$$

aus unterschiedlichen Schulen. In *Kantonen mit zweistufigen Stichprobenverfahren* führt dies dazu, dass bei der Ziehung von Schulen relativ leistungshomogene Gruppen in die Stichproben aufgenommen werden. Diese Klumpeneffekte können den oben beschriebenen Stichprobenfehler vergrössern bzw. die Messpräzision verringern. Auch Schulklassen stellen solche Klumpen dar. Da im Rahmen der ÜGK 2016 jedoch keine Schulklassen gezogen wurden (vgl. 4.4), sind hier Klumpeneffekte auf Klassenebene stichprobentheoretisch irrelevant.

Bei mehreren in Frage kommenden Stichprobendesigns gilt dasjenige als effizienter, das bei gleichbleibendem Stichprobenumfang den tieferen Stichprobenfehler aufweist. Diese Effizienz wird in der Regel mit dem Designeffekt quantifiziert, der die Stichprobenvarianz eines zu beurteilenden Designs ins Verhältnis zur erwarteten Stichprobenvarianz einer einfachen Zufallsstichprobe setzt. So lässt sich auch die Auswirkung von merkmalshomogenen Gruppen auf den Stichprobenfehler bei mehrstufigen Stichprobenverfahren mithilfe des Designeffekts beziffern. In diesem Fall setzt er sich aus der Grösse der Klumpen sowie einem Mass für die Homogenität des Merkmals zusammen.[3] Um den Stichprobenfehler bei komplexeren (zweistufigen) Stichprobenverfahren a priori zu schätzen, wird deshalb die Quadratwurzel des Designeffekts mit dem bei einer einfachen Zufallsstichprobe zu erwartenden Stichprobenfehler multipliziert (Kish, 1965).

Im Vorfeld der ÜGK 2016 wurden Intraklassenkorrelationskoeffizienten auf Basis vergangener PISA-Erhebungen geschätzt. Dabei wurde untersucht, wie stark die Anteile der Schülerinnen und Schüler, die das PISA-Kompetenzniveau 2 in Mathematik nicht erreicht haben, zwischen den bzw. innerhalb der Schweizer Schulen variieren. Dabei waren in Kantonen mit stark nach Leistung getrennten Schulsystemen höhere Intraklassenkorrelationen zu erwarten als in Kantonen mit einer weniger ausgeprägten Selektion. So wurden beispielsweise für den Kanton AG höhere Intraklassenkorrelationen ($\rho \approx 0.50$) als für die Kantone BE_d ($\rho \approx 0.30$) oder VD ($\rho \approx 0.15$) geschätzt. Diesen Effekten kann entgegengewirkt werden, indem Schulen oder Schülerinnen und Schüler vor der Ziehung in leistungshomogene Gruppen gegliedert werden. Detaillierte Informationen zu diesen Gruppierungen (Stratifizierung) sind in den Abschnitten 3.1 (Schulebene) sowie 4.3 (Schülerebene) ersichtlich.

[3] Der Designeffekt ($deff$) aufgrund komplexer Stichprobendesigns ist abhängig von der mittleren Klumpengrösse (b) und dem Intraklassenkoeffizienten (ρ) des entsprechenden Merkmals (vgl. Kish, 1965): $deff = 1 + \rho(b - 1)$

1.3 Stichprobenumfänge

Bei der Erarbeitung des Stichprobendesigns wurde von der Vorgabe ausgegangen, dass die maximale Gesamtstichprobengrösse von 25'000 Schülerinnen und Schülern und eine Gesamtzahl von 1'400 bis 1'500 Testsitzungen nicht überschritten werden sollte. Darüber hinaus war es das Ziel, eine kantonal vergleichbare Schätzpräzision zu erreichen, während Auswertungen auf nationaler Ebene in möglichst kleinen Standardfehlern resultieren. Um einen entsprechenden Kompromiss zwischen maximaler Schätzpräzision auf kantonaler und nationaler Ebene zu gewährleisten, wurde darauf verzichtet, die Stichprobengrösse gleichmässig auf die 29 Kantonsteile aufzugliedern. Damit möglichst präzise Ergebnisse auf kantonaler Ebene erzielt und kantonsspezifische Stichprobenverfahren ausgearbeitet werden konnten, wurde die Population in 29 Schichten aufgeteilt, die jeweils einem expliziten Stratum mit separatem Ziehungsdesign entsprachen (vgl. 3.1).

In Tabelle 1.1 werden Populationsgrössen für *Kantone mit Vollerhebungen* dargestellt. Aufgrund von Abwesenheiten einzelner Schülerinnen und Schüler, Schulverweigerungen und von technischen Problemen entspricht die Anzahl getesteter Schülerinnen und Schüler nicht in allen Kantonen dem Populationsumfang.

Tabelle 1.1: Anzahl getesteter Schülerinnen und Schüler sowie Populationsumfänge getrennt nach Kanton für *Kantone mit Vollerhebungen*

Kanton	Populationsumfang	Realisierte Stichprobe
AI	204	202
AR	514	482
BE_f	814	749
GL	391	376
JU	815	683
NW	425	410
OW	443	437
SH	708	666
UR	366	357
VS_d	812	784
ZG	1'197	1'142
Total	*6'689*	*6'288*

Anmerkungen: Bei den Stichprobengrössen handelt es sich um die Anzahl tatsächlich am Test teilnehmender Schülerinnen und Schüler. Die Populationsumfänge beruhen auf der jeweiligen kantonalen Summe der Schülergewichte und beziehen sich demnach auf die *ÜGK-Population* (vgl. Kapitel 2). Abhängig von Ausschlüssen und der Güte der für die Ziehung von Schulen verwendeten Schülerbestandslisten, können diese Zahlen von den tatsächlichen, kantonalen Populationsumfängen abweichen.

Tabelle 1.2: Anzahl getesteter Schülerinnen und Schüler sowie Populationsumfänge getrennt nach Kanton für *Kantone mit ein-* oder *zweistufigen Stichprobenverfahren*

Kanton	ÜGK-Population	Realisierte Stichprobe
Kantone mit einstufigen Stichprobenverfahren		
BL	2'588	703
BS	1'427	628
FR_d	894	717
FR_f	2'750	743
GE	4'530	665
GR	1'826	925
NE	1'883	648
SO	2'389	746
SZ	1'634	755
TG	2'700	996
TI	3'205	695
VS_f	2'556	755
Total	*28'382*	*8'976*
Kantone mit zweistufigen Stichprobenverfahren		
AG	6'903	1'112
BE_d	8'806	1'093
LU	4'002	1'093
SG	4'805	1'137
VD	7'960	1'014
ZH	13'309	1'710
Total	*45'785*	*7'159*

Anmerkungen: Bei den Stichprobengrössen handelt es sich um die Anzahl tatsächlich am Test teilnehmender Schülerinnen und Schüler. Die Populationsumfänge beruhen auf der jeweiligen kantonalen Summe der Schülergewichte und beziehen sich demnach auf die *ÜGK-Population* (vgl. Kapitel 2). Abhängig von Ausschlüssen und der Güte der für die Ziehung von Schulen verwendeten Schülerbestandslisten, können diese Zahlen von den tatsächlichen, kantonalen Populationsumfängen abweichen.

Die *Kantone mit einstufigen Stichprobenverfahren* zeichneten sich dadurch aus, dass eine vergleichsweise hohe Anzahl Schülerinnen und Schüler den Unterricht in der 11. Klasse besucht hat, die Schülerinnen und Schüler jedoch auf verhältnismässig wenige Schulen verteilt waren. Die Bildung einer Schulstichprobe hätte in diesen Kantonen dazu geführt, dass zur Erreichung des anvisierten Stichprobenumfangs ein relativ hoher Anteil der Schülerinnen und Schüler einer gezogenen Schule am Test hätte teilnehmen müssen. Der durch die Ziehung von Schulen entstehende Klumpeneffekt hätte die Schätzpräzision jedoch derart beeinträchtigt, dass es sich stattdessen anbot, sämtliche Schulen der entsprechenden Kantone zu berücksichtigen (Vollerhebung auf Ebene der Schulen) und innerhalb der einzelnen Schulen jeweils eine Zufallsstichprobe von Schülerinnen und Schülern zu ziehen. Der Richtwert für Stichprobengrössen für *Kantone mit einstufigen Stichprobeverfahren* betrug ursprünglich 700 Schülerinnen und Schüler. Da abwesende, ausgeschlossene oder verweigernde Schülerinnen

und Schüler nicht einberechnet werden und die definitiven Stichprobenumfänge zusätzlich von der Computerinfrastruktur der teilnehmenden Schulen abhängig waren (vgl. 4.2), weichen die in Tabelle 1.2 dargestellten Zahlen vom Richtwert ab.

Um Klumpeneffekten entgegenzuwirken, wurden in *Kantonen mit zweistufigen Stichprobenverfahren* verhältnismässig mehr Schülerinnen und Schüler gezogen. Die entsprechenden Stichproben- sowie Populationsumfänge sind ebenfalls in Tabelle 1.2 aufgeführt. Der ursprüngliche Richtwert entsprach einem Stichprobenumfang von 1'000 Schülerinnen und Schülern pro Kanton. Um einem aufgrund variierender Gewichte wachsenden Stichprobenfehler (vgl. hierzu Kish, 1992) auf nationaler Ebene entgegenzuwirken, wurden die Stichprobenumfänge in Kantonen mit grossen Populationen teilweise weiter erhöht. Dies führte vor allem im Kanton ZH zu einem deutlich höheren Stichprobenumfang. Auch im Kanton BE_d wäre ein höherer Stichprobenumfang durchaus sinnvoll gewesen. Die sehr hohe Anzahl kleiner Schulen im Kanton BE_d (die Anzahl Schulen mit 11. Schuljahren ist in BE_d höher als in ZH) führte jedoch dazu, dass in BE_d aus ressourcentechnischen Gründen der Stichprobenumfang auf knapp 1'100 Schülerinnen und Schüler begrenzt werden musste.

1.4 Stichprobengewichte

Für jede an der ÜGK 2016 teilnehmende Schülerin und jeden teilnehmenden Schüler wurde ein Stichprobengewicht berechnet. Die Schülergewichte sind ein Mass für die Anzahl Schülerinnen und Schüler in der Population, die durch die entsprechenden Schülerinnen und Schüler in der Stichprobe repräsentiert werden. Mit anderen Worten: Die Summe der Gewichte sämtlicher an der ÜGK 2016 teilnehmender Schülerinnen und Schüler entspricht näherungsweise dem Umfang der *ÜGK-Population* (vgl. Kapitel 2). Die Gewichte kompensieren primär die unterschiedlichen individuellen Auswahlwahrscheinlichkeiten der einzelnen Schülerinnen und Schüler, können jedoch abhängig von Absenzen und Verweigerungen nachträglich korrigiert worden sein. Einzelheiten zur Gewichtung werden in Kapitel 5 berichtet.

2 Population

In diesem Kapitel wird die interessierende Population bzw. diejenige Gruppe von Schülerinnen und Schülern, deren erhobene Merkmale durch den Datensatz ÜGK 2016 repräsentiert wird, definiert. Dies beinhaltet auch die Erläuterung diverser Ausschlüsse auf Schul- sowie Schülerebene.

Bei landesweiten oder internationalen Erhebungen der Schulleistung werden in der Regel Alterskohorten definiert oder ein interessierendes Schuljahr festgelegt. Da bei der ÜGK 2016 die Schülerinnen und Schüler am Ende des 11. Schuljahres im Mittelpunkt standen, bot sich eine auf dem Schuljahr beruhende Populationsdefinition geradezu an. Verglichen mit der Methode der Alterskohorte sind die Festlegung einer Population, das Ziehen einer Stichprobe sowie die Durchführung der Erhebung bei einer durch das Schuljahr definierten Population vergleichsweise einfach (Rust, 2014).

Dementsprechend umfasste die erwünschte Population der ÜGK 2016 sämtliche Schülerinnen und Schüler, die in einer nach Schweizer Recht organisierten Schule in der 11. Klasse unterrichtet wurden. Gemäss dieser Definition waren private Schulen (unabhängig vom Subventionsgrad) grundsätzlich auch Teil der Population. Lediglich Schulen, die auf Basis von ausländischen Programmen oder in keiner Landessprache unterrichten, waren in dieser Definition nicht berücksichtigt. In der Schweiz trifft dies auf eine äusserst kleine Anzahl internationaler Schulen zu, die ausschliesslich englisch- oder japanischsprachigen Unterricht anbieten.

Es war ein erklärtes Ziel der ÜGK, dass die Stichprobe die so definierte Population möglichst lückenlos abdeckt. Dennoch waren Ausschlüsse sowohl auf Schulebene als auch auf der Ebene der Schülerinnen und Schüler – vorwiegend aus erhebungspraktischen Gründen – unvermeidbar. Die tatsächlich untersuchte Population, auf die sich die gewichteten Ergebnisse der ÜGK 2016 beziehen (in der Folge *ÜGK-Population* genannt), umfasste aufgrund der im nächsten Abschnitt erläuterten Ausschlüsse weniger Schülerinnen und Schüler als die erwünschte Population.

2.1 Ausschlüsse auf Schulebene

Bei der ÜGK 2016 wurden auf Schulebene Sonderschulen aus diversen Gründen ausgeblendet:

- Die grosse Mehrheit der in Sonderschulen unterrichteten Schülerinnen und Schüler kann nicht einem Schuljahr zugeordnet werden, was die Bildung einer klar definierten Zielgruppe deutlich erschwert.
- Den wenigsten Kantonen stehen Informationen über die Häufigkeit bestimmter Behinderungsformen bzw. Lern- und Verhaltensschwierigkeiten an einzelnen

Sonderschulen zur Verfügung. Ohne diese Informationen lässt sich nicht feststellen, an welchen Sonderschulen eine Testdurchführung im Sinne einer objektiven und vor allem zumutbaren Erhebung überhaupt möglich ist.

- Die Mathematikaufgaben wurden nicht im Hinblick auf Sonderschulen entwickelt. Zahlreiche in Sonderschulen unterrichtete Schülerinnen und Schüler könnten die Aufgaben ohne fremde Unterstützung nicht bearbeiten.

Da sich an Sonderschulen unterrichtete Schülerinnen und Schüler nicht einem Schuljahr zuordnen lassen, kann nur grob geschätzt werden, wie viele Schülerinnen und Schüler der erwünschten Population ÜGK 2016 durch diesen Ausschluss betroffen waren. Hierbei handelt es sich also um in Sonderschulen unterrichtete Schülerinnen und Schüler, die in einer 11. Klasse gewesen wären, wenn sie eine Regelschule besuchen würden. Mit der Annahme, dass der Anteil Sonderschülerinnen und Sonderschüler zwischen einer bestimmten Alterskohorte (Jahrgang) und einem Schuljahr vergleichbar ist, kann davon ausgegangen werden, dass knapp über zwei Prozent der erwünschten Population in Sonderschulen unterrichtet wurden.[4] Die geschätzten Anteile für die einzelnen Kantone variieren relativ stark und werden in der zweitletzten Spalte in Tabelle A.1 (siehe Anhang A) dargestellt.

2.2 Ausschlüsse auf Schülerebene

Innerhalb der zur Teilnahme bestimmten Schulen hatten die jeweiligen Schulleitungen bzw. Lehrpersonen die Möglichkeit, einzelne Schülerinnen und Schüler, beruhend auf den folgenden Kriterien, von der Erhebung zu dispensieren:

- Kognitiv beeinträchtigte Schülerinnen und Schüler, deren Beeinträchtigung gemäss zuständigen Lehrpersonen eine valide Testdurchführung verunmöglichte, wurden ausgeschlossen. Hierzu gehören in Regelschulen unterrichtete Schülerinnen und Schüler, die auf emotionaler oder kognitiver Ebene den allgemeinen Anweisungen des Tests nicht folgen können und für die dementsprechend eine Testdurchführung als nicht zumutbar eingestuft wurde.
- Funktional beeinträchtigte Schülerinnen und Schüler, deren körperliche Beeinträchtigung die Validität der Ergebnisse einschränken könnte, wurden ebenfalls nicht einbezogen. Hierbei handelt es sich hauptsächlich um Schülerinnen und Schüler mit Körper- oder Sehbehinderungen.
- Schliesslich wurden auch Schülerinnen und Schüler mit sehr schlechten Kenntnissen der Testsprache ausgeschlossen. Notwendige

[4] Vgl. die Statistik der Lernenden Schuljahr 2015/16 des Bundesamts für Statistik.

Ausschlusskriterien waren, dass (1) die Muttersprache nicht der Testsprache entsprach, (2) die sprachlichen Schulleistungen deutlich eingeschränkt waren und (3) die Schülerin oder der Schüler weniger als ein Jahr in der Testsprache unterrichtet wurde.

Die Schulen wurden ausdrücklich darauf hingewiesen, dass schlechte Schulleistungen oder disziplinarische Probleme keinen Ausschlussgrund darstellen. Dennoch variieren sowohl die Anzahl der als auch die Gründe für Ausschlüsse innerhalb Schulen zwischen den Kantonen relativ stark. Dabei wird deutlich, dass die Ausschlussquoten innerhalb von Schulen in der deutschsprachigen Schweiz (gesamthaft 1 Prozent) tiefer sind als in der Romandie (2.5 Prozent) oder im Tessin (2.4 Prozent).

Es soll deutlich darauf hingewiesen werden, dass Tabelle A.1 lediglich der Schätzung kantonaler Ausschlussquoten dient. Die Zahlen können nicht zur Berechnung der Häufigkeit einzelner Behinderungsformen verwendet werden, da die grosse Mehrheit der Schülerinnen und Schüler mit besonderen Lernbedürfnissen am Test teilgenommen hat. Darüber hinaus wurden dem Test ferngebliebene Schülerinnen und Schüler, die nicht explizit von der Schulleitung bzw. der Lehrperson als ausgeschlossen kommuniziert worden waren, nicht ausgeschlossen, sondern als «abwesend» vermerkt und dementsprechend bei der Korrektur der Stichprobengewichte berücksichtigt (vgl. 0).

3 Schulstichproben

Die hohe Anzahl Schulen in den Kantonen AG, BE_d, LU, SG, VD und ZH bedingte die Ziehung von Schulstichproben. Dazu wurde beim Bundesamt für Statistik (BfS) eine Liste sämtlicher Schulen, die Schülerinnen und Schüler im 11. Schuljahr unterrichten, eingefordert. Für sämtliche Schulen enthielt diese Liste – nebst Schulnamen und Adresse – Angaben zu Trägerschaft, unterrichteten Schulprogrammen sowie zur Anzahl unterrichteter Schülerinnen und Schüler im 11. Schuljahr. Schliesslich wurde die Liste der wählbaren Schulen mithilfe der kantonalen Bildungsstatistiken teilweise ergänzt bzw. aktualisiert.

3.1 Stratifizierung der Liste wählbarer Schulen

Um den Einsatz unterschiedlicher Stichprobenverfahren in den verschiedenen Kantonen zu ermöglichen, die Effizienz dieser Verfahren zu erhöhen und um sicherzustellen, dass alle Teile der Population adäquat in der Stichprobe vertreten waren, wurde in *Kantonen mit zweistufigen Stichprobenverfahren* die Liste wählbarer Schulen vor dem Ziehungsprozess nach bestimmten Merkmalen geschichtet und sortiert.

Analog zu den im Rahmen von PISA verwendeten Stichprobenverfahren (OECD, 2017) wurde bei der Vorbereitung der Schullisten sowohl auf explizite als auch auf implizite Stratifizierungsmethoden zurückgegriffen. Erstere beinhalten die Bildung von Schichten (Strata), die im Verlauf des Stichprobenprozesses unabhängig voneinander behandelt werden (vgl. Rust, 2014, S. 126; Meinck, 2015). So bildet jeder Kanton ein explizites Stratum, was den Einsatz kantonsspezifischer Verfahren ermöglichte. Um die Effizienz des Verfahrens zu erhöhen und adäquate Anteile von Privatschulen und bestimmten Schulprogrammen zu gewährleisten, wurden die einzelnen Kantone auf weitere drei bis vier Strata auf Basis von kantonalen Schulprogrammen und der Schulträgerschaft aufgeteilt.

Die implizite Stratifizierungsmethode bezieht sich auf die Sortierung der Schullisten nach bestimmten Schulmerkmalen. Eine adäquate Sortierung innerhalb der expliziten Strata kann dann zu einer Reduktion des Stichprobenfehlers führen, wenn die Ziehung der Schulen auf eine systematische (vgl. Rust, 2014, S. 129) Art und Weise durchgeführt wird. Der Begriff «systematisch» bedeutet in diesem Zusammenhang, dass ein auf einer Zufallszahl beruhendes Ziehungsintervall definiert wird, mit dem die sortierten Listen «durchgezählt» und die entsprechenden «Treffer» als gezogene Einheiten gelten (vgl. 3.3).

Innerhalb der expliziten Strata eines Kantons wurden die Schulen dementsprechend nach allfälligen Schulprogrammen oder -modellen sowie – um eine Stichprobe mit ausschliesslich kleinen oder grossen Schulen zu verhindern – nach der geschätzten Anzahl unterrichteter Schülerinnen und Schüler sortiert. In Strata, in denen mehrere Programme pro Schule unterrichtet werden, wurden Anteile von Schülerinnen und

Schülern, die bestimmte Programme besuchen, als implizite Stratifizierungsvariable verwendet. So wurden beispielsweise die Schulen des Kantons AG, in welchen sowohl Real- als auch Sekundarschüler unterrichtet werden, nach den Anteilen auf Realniveau unterrichteter Schülerinnen und Schüler sortiert. Tabelle B.1 in Anhang B fasst die verwendeten Stratifizierungsvariablen für *Kantone mit zweistufigen Stichprobenverfahren* zusammen.

3.2 Aufteilung der Stichprobe auf die expliziten Strata

Nachdem die expliziten Strata und die kantonalen Stichprobenumfänge bestimmt waren, stellte sich die Frage nach der innerkantonalen Aufteilung der Anzahl zu ziehender Schulen auf die verschiedenen Strata. Häufig geschieht diese Aufteilung proportional zur Anzahl unterrichteter Individuen in den einzelnen Strata: Im populationsreichsten Stratum werden die meisten Schulen bzw. Schülerinnen und Schüler gezogen und umgekehrt. Bei dieser *proportionalen Aufteilung* werden allfällige Unterschiede in der Varianz des Zielmerkmals zwischen den Strata nicht berücksichtigt. Da anzunehmen war, dass sich die Anteile an Schülerinnen und Schülern, welche die Grundkompetenzen nicht erreichen, stark zwischen leistungsbezogenen, kantonalen Schulprogrammen unterscheiden, wurde für die ÜGK 2016 eine alternative Methode zur rein proportionalen Aufteilung gewählt.

Der Grund für diesen Entscheid liegt in der Varianz des Zielmerkmals (dem Anteil Schülerinnen und Schüler, deren Leistungen den Grundkompetenzen nicht genügen; vgl. 1.2). Es wurde angenommen, dass äusserst wenige Schülerinnen und Schüler, die ein kantonales Programm mit hohen Anforderungen besuchen, die Grundkompetenzen in Mathematik nicht erreichen. Gemäss Schätzungen, die auf PISA 2012 beruhen, dürfte dieser Anteil unter 1 Prozent liegen. Mit anderen Worten: In Gymnasien und Kantonsschulen war kaum Varianz im Zielmerkmal zu erwarten. Vor allem in Schulen bzw. Programmen mit Grundanforderungen war mit deutlich höheren Anteilen und dementsprechend auch mehr Varianz im Zielmerkmal zu rechnen. Um eine möglichst hohe Präzision der Schätzung des Zielmerkmals zu erzielen, wurde deshalb zunächst für jedes Stratum, beruhend auf der jeweils erwarteten Varianz des Merkmals, eine Anzahl zu ziehender Schulen mittels der *optimalen Aufteilung* gemäss Neyman (vgl. Lehtonen & Pahkinen, 1995, S. 70; Lohr, 2010, S. 89) bestimmt:

$$n_h = \frac{n(N_h \cdot S_h)}{(\sum N_i \cdot S_i)}$$

Die Anzahl zu ziehender Schulen n_h in Stratum $\boldsymbol{h}$ ergibt sich dabei aus der kantonalen Stichprobengrösse n, dem Populationsumfang N_h in Stratum $\boldsymbol{h}$ und der Zielmerkmal-Standardabweichung S_h in Stratum $\boldsymbol{h}$. Der Nenner in der Formel enthält die Summe der Produkte von Population und Standardabweichung für alle in der Aufteilung

berücksichtigten Strata. Wie dies meistens in der Praxis der Fall ist, war auch im vorliegenden Fall S_h nicht bekannt und musste geschätzt werden.

Basis dieser Schätzungen waren Anteile der Schülerinnen und Schüler, die im Rahmen vergangener PISA-Erhebungen das Kompetenzniveau 2 in Mathematik nicht erreicht hatten. Gemäss PISA besteht für diese Schülerinnen und Schüler ein erhöhtes Risiko, dass der Übergang von der Schule ins Arbeitsleben nicht reibungslos verläuft und die Nutzung von Fort- und Weiterbildungsangeboten mit erheblichen Problemen verbunden ist (OECD, 2017). Die Schätzungen für die oben erwähnte *optimale Aufteilung* beruhen dementsprechend auf der Annahme, dass sich die beiden Gruppen – Schülerinnen und Schüler, die das PISA-Kompetenzniveau 2 nicht erreichen, sowie Schülerinnen und Schüler, die den HarmoS-Grundkompetenzen nicht genügen – zu grossen Teilen überschneiden. Teilweise konnte auf kantonal repräsentative PISA-Stichproben zurückgegriffen werden. Falls dies nicht möglich war, beruhten die Schätzungen auf der jeweiligen Sprachregion des Kantons.

Die aus der *optimalen Aufteilung* resultierenden Stichprobengrössen pro Stratum innerhalb eines Kantons waren äusserst ungleich. Da die Anzahl in Programmen mit hohen Anforderungen unterrichteter Schülerinnen und Schüler, die im Rahmen von PISA das Kompetenzniveau 2 nicht erreicht hatten, nahe Null ist, wären gemäss der *optimalen Aufteilung* jeweils äusserst wenige Schulen mit progymnasialem Unterricht in die Stichprobe aufgenommen worden. Während dies zwar zu präzisen Schätzungen des Zielmerkmals geführt hätte, hätten weitere Analysen – beispielsweise mit Fragebogendaten der progymnasial unterrichteten Schülerinnen und Schüler – unter einem Mangel an Präzision gelitten bzw. Analysen über die genannte Population vollständig verunmöglicht. Aus diesem Grund wurden schliesslich Schulen mittels eines Kompromisses zwischen *proportionaler* und *optimaler Aufteilung* gezogen. Dabei wurde grundsätzlich eine *optimale Aufteilung* angewendet, jedem Stratum mussten aber gleichzeitig mindestens 60 Prozent der Schulen zugeteilt werden, die bei einer *proportionalen Aufteilung* resultiert hätten. Die Anzahl gezogener Schulen pro Stratum und *Kanton mit zweistufigen Stichprobenverfahren* kann der Tabelle B.1 in Anhang entnommen werden.

3.3 Systematische Ziehung der Schulen per PPS-Verfahren

Analog zu etablierten, internationalen *Large-Scale Assessments* (z.B. *Trends in International Mathematics and Science Study*, TIMSS; *Progress in International Reading Literacy Study*, PIRLS; *Program of International Student Assessment*, PISA) wurden in *Kantonen mit zweistufigen Stichprobenverfahren* in einem ersten Schritt Schulen proportional zu ihrer Grösse (*Probability Proportional to Size*; *PPS*; z.B. Rust, 2014) gezogen und in einem zweiten Schritt eine bestimmte Anzahl Schülerinnen und Schüler pro Schule zur Teilnahme an der Erhebung aufgeboten. Aus diesem Grund wurde jeder Schule eine *Measure of Size* (*MOS*) zugeordnet, die grundsätzlich der geschätzten Anzahl im 11.

Schuljahr unterrichteter Schülerinnen und Schüler entsprach (mit Ausnahme kleiner Schulen, vgl. 3.5). In der grossen Mehrheit der Kantone enthielten die Schullisten Angaben aus älteren Schuljahren. Dementsprechend kann die *MOS* von der tatsächlichen Anzahl unterrichteter Schülerinnen und Schüler abweichen. Informationen zur Qualität dieser Schätzungen können den Tabellen B.2 und B.3 im Anhang entnommen werden.

Anstatt einer einfachen Zufallsstichprobe wurde ein systematisches Verfahren eingesetzt. Dieses Verfahren beinhaltet die systematische Sortierung der Listen sämtlicher Einheiten in der Population nach bestimmten Merkmalen (vgl. implizite Stratifizierung, 3.1). Grundsätzlich werden dabei Samplingintervalle (*SI*) definiert, die sich aus dem Populationsumfang *N* dividiert durch die erwünschte Stichprobengrösse *n* berechnen. Vor der Ziehung wird eine Startzahl (*RN*) entsprechend einem Zufallswert zwischen 0 und *SI* definiert. Die gezogenen Einheiten in der Liste entsprechen den «Auswahlnummern» *RN*, *RN* + *SI*, *RN* + 2*SI* usw. bis *RN* + (*n* - 1) *SI* (vgl. Rust, 2014, S. 129).

Im konkreten Fall der ÜGK 2016 wurde das systematische Prinzip auf die PPS-Ziehung von Schulen in *Kantonen mit zweitstufigen Stichprobenverfahren* übertragen: Da *N* der Summe von *MOS* entsprach (MOS_{tot}), ergab sich für jedes explizite Stratum *SI* aus MOS_{tot} dividiert durch die Anzahl der zu ziehenden Schulen *n*. Schulen, deren *MOS* gleich oder grösser *SI* war, hatten eine Wahrscheinlichkeit von 1, um in die Stichprobe zu gelangen, und wurden deshalb in jeweils separate Strata – *certainty strata* – überführt, bevor MOS_{tot} und *SI* neu berechnet wurden. Dieser iterative Prozess wurde so lange wiederholt, bis die *MOS* aller Schulen kleiner war als *SI*.

Anschliessend wurde für jedes explizite Stratum eine auf vier Kommastellen gerundete Zufallszahl zwischen 0 und 1 (*RN*) generiert. Das Produkt aus *RN* und *SI* entsprach der ersten «Auswahlnummer» jedes Stratums. Die erste zu ziehende Schule war dementsprechend die erstgelistete Schule, deren kumulative *MOS* (MOS_{cum}) gleich oder grösser als die «Auswahlnummer» war. Wurde zu dererstenn «Auswahlnummer» *SI* hinzuaddiert (*RN SI* + *SI*), entsprach dies der zweiten «Auswahlnummer». Die dritte «Auswahlnummer» war die Summe aus zweiter «Auswahlnummer» und *SI* (*RN SI* + 2 *SI*) usw. bis zur letzten «Auswahlnummer» *RN SI* + (*n* - 1) *SI*. Die jeweils erstgelisteten Schulen, deren MOS_{cum} gleich oder grösser als die «Auswahlnummern» war, fielen in die Stichprobe. Diese «Auswahlnummern» wurden unabhängig für jedes explizite Stratum berechnet, indem jeweils neue *RN* generiert wurden.

3.4 Ersatzschulen

Sofern es die Anzahl vorhandener Schulen pro Stratum erlaubte, wurden jeder gezogenen Schule zwei Ersatzschulen zugeordnet. Damit die Ersatzschulen den

erstgezogenen Schulen hinsichtlich impliziter Stratifizierungsvariablen (Schulgrösse, unterrichtete Programme) möglichst ähnlich waren, wurden jeweils diejenigen Schulen als Ersatzschulen bestimmt, die in der stratifizierten Liste unmittelbar vor und nach der gezogenen Schule aufgeführt waren. Vor allem bei grösseren Schulen war diese Methode nicht anwendbar, weil manchmal zwei aufeinanderfolgende Schulen in die Stichprobe fielen. In solchen Fällen wurde teilweise dieselbe Schule als Ersatzschule für mehrere erstgezogene Schulen definiert. Ersatzschulen wurden nur dann zur Teilnahme aufgefordert, wenn die ursprünglich gezogene Schule die Teilnahme verweigerte. Inwieweit Ersatzschulen zum Einsatz kamen, kann Tabelle B.4 im Anhang entnommen werden.

3.5 Umgang mit kleinen Schulen

Während der Bildung der diversen Strata auf Schulebene wurden die Schulen zusätzlich vier Kategorien, beruhend auf der geschätzten Anzahl unterrichteter Schülerinnen und Schüler, zugeteilt. Der Richtwert der zu ziehenden Schülerinnen und Schüler in *Kantonen mit zweistufigen Stichprobenverfahren* betrug 20 (*Target Cluster Size, TCS*; vgl. 4.2). Sämtliche Schulen mit mindestens 20 *(= TCS)* unterrichteten Schülerinnen und Schülern im 11. Schuljahr wurden als *gross* eingestuft. Schulen mit einer Schülerzahl zwischen 10 *(= TCS/2)* und 20 galten als *mittelgross*, solche mit weniger als zehn, aber mindestens drei Schülerinnen und Schülern wurden als *klein* bezeichnet. Die restlichen Schulen wurden als *sehr klein* eingestuft. Sämtliche gezogenen Schulen wurden kontaktiert und über ihre Teilnahme an der ÜGK 2016 informiert. Betrug die Anzahl tatsächlich im 11. Schuljahr unterrichteter Schülerinnen und Schüler jedoch weniger als vier, wurde aus ökonomischen Gründen auf eine Testdurchführung verzichtet. Der Ausfall dieser Schulen wurde auf Schulebene mit Anpassungen der Stichprobengewichte kompensiert (vgl. 5.3).

Enthielten Strata Schulen mit weniger als 20 Schülerinnen und Schülern, bestand das Risiko, dass die Anzahl gezogener Schülerinnen und Schüler dem erwünschten Stichprobenumfang nicht genügt. Darüber hinaus kann eine Schulstichprobe mit zahlreichen kleinen Schulen zu einem unverhältnismässig hohen Aufwand führen, da jeweils Testsitzungen mit äusserst wenigen Schülerinnen und Schülern organisiert werden müssen. Um diese Probleme zu beheben, wurde ein Verfahren angewendet, das sich an den Vorgehensweisen vergangener PISA-Erhebungen orientiert (OECD, 2017, S. 77).

Bei einer verhältnismässig grossen Anzahl kleiner Schulen innerhalb eines Stratums wurde die Auswahlwahrscheinlichkeit *kleiner* sowie *sehr kleiner* Schulen um die

Faktoren 0.5 bzw. 0.25 verringert[5], während gleichzeitig die Auswahl *grosser* Schulen proportional erhöht wurde. Zusammengefasst enthielt dieses Verfahren die folgenden Überprüfungen:

- Wenn der Anteil von in *kleinen* sowie *sehr kleinen* Schulen unterrichteten Schülerinnen und Schülern 1 Prozent oder mehr betrug, wurden diese Schulen unterrepräsentiert und die Anzahl zu ziehender Schulen erhöht.
- Wenn der Anteil von in *kleinen* sowie *sehr kleinen* Schulen unterrichteten Schülerinnen und Schülern unter 1 Prozent lag, der Anteil in *mittelgrossen* Schulen jedoch mindestens 4 Prozent entsprach, wurde lediglich die Anzahl zu ziehender Schulen erhöht. Auf ein *Undersampling kleiner* sowie *sehr kleiner* Schulen wurde in diesen Fällen verzichtet.

War keine dieser Bedingungen gegeben, war die Wahrscheinlichkeit gering, dass der relativ kleine Anteil *kleiner* und *sehr kleiner* Schulen die gewünschte Stichprobengrösse in einem relevanten Ausmass reduziert. In diesem Fall wurden weder Schulen unterrepräsentiert noch der Stichprobenumfang angehoben. Die detaillierten Berechnungsschritte dieser Überprüfungen sind in Tabelle B6 im Anhang dargestellt.

[5] Die *MOS* von *mittelgrossen* Schulen betrug stets 20. Dementsprechend war beispielsweise die Auswahlwahrscheinlichkeit einer Schule mit geschätzten 12 Schülerinnen und Schülern dieselbe wie bei einer Schule mit geschätzten 20 unterrichteten Schülerinnen und Schülern. Im Falle einer Unterrepräsentation wurden die *MOS* von *kleinen* sowie *sehr kleinen* Schulen auf 10 (*TCS*/2) bzw. 5 (*TCS*/4) gesetzt.

4 Schülerstichproben

Während in den Kantonen AG, BE_d, LU, SG, VD und ZH – anhand des in Kapitel 3 beschriebenen Stichprobenverfahrens – Schulen zur Teilnahme an der ÜGK 2016 gezogen wurden, nahmen in den restlichen Kantonen sämtliche Schulen an den Erhebungen teil (verweigernde Schulen ausgenommen). In *Kantonen mit Vollerhebungen* wurden zudem alle Schülerinnen und Schüler zum Test aufgeboten. In der Mehrzahl der Kantone *(Kantone mit ein-* oder *zweistufigen Stichprobenverfahren)* wurde jedoch auf Schülerebene ein Stichprobenverfahren angewendet, das in den folgenden Abschnitten beschrieben wird.

4.1 Listen wählbarer Schülerinnen und Schüler

Nachdem die teilnehmenden Schulen bestimmt waren, wurden diese erstmals brieflich kontaktiert. Nebst dem Versand eines allgemeinen Informationsschreibens zur Studie wurden – mit Ausnahme der Kantone GE, NE, TI und VD, bei welchen die Schülerlisten mithilfe eines zentralen, kantonalen Registers erstellt wurden – Listen sämtlicher Schülerinnen und Schüler, die im 11. Schuljahr unterrichtet wurden, eingefordert. Die Schulleitungen wurden gebeten, die Namenslisten mit den folgenden Informationen zu ergänzen:

- Geschlecht
- Geburtsdatum
- Kantonales Schulprogramm
- Angaben zu allfälligen Niveauzugehörigkeiten oder Lernzielbefreiungen
- Klassenbezeichnung
- Name Klassenlehrperson

Diese Listen bildeten die Grundlage zur Ziehung der teilnehmenden Schülerinnen und Schüler.

4.2 Stichprobenumfänge auf Schülerebene

Vor der Stichprobenziehung auf Schülerebene war es notwendig die Anzahl der an der Studie teilnehmenden Schülerinnen und Schüler pro Schule zu bestimmen (*Target Cluster Size, TCS*). Zusätzlich zu dem in Abschnitt 1.2 beschriebenen Designeffekt aufgrund komplexer Stichprobendesigns können auch zu stark variierende Stichprobengewichte zu Einbussen in der Schätzpräzision führen (vgl. Kish, 1995; Liu, Iannacchione & Byron, 2002; Le, Brick & Kalton, 2002). Aus diesem Grund und weil sich die Stichprobengewichte umgekehrt proportional zur Auswahlwahrscheinlichkeit verhalten (vgl. Kapitel 5), wurde innerhalb jedes expliziten Stratums versucht, die

Schülerinnen und Schüler mit einer möglichst vergleichbaren Auswahlwahrscheinlichkeit zu ziehen.

In *Kantonen mit zweistufigen Stichprobenverfahren* führte eine konstante *TCS* aufgrund der *PPS*-Selektion (vgl. 3.3) auf Schulebene zu vergleichbaren Stichprobengewichten: Die Auswahlwahrscheinlichkeiten grosser Schulen waren verhältnismässig hoch, während die Wahrscheinlichkeit einer einzelnen Schülerin bzw. eines einzelnen Schülers, in die schulinterne Stichprobe zu gelangen, bei grossen Schulen relativ gering war (vgl. Rust, 2014, S. 130). Auf theoretischer Ebene ist das Produkt dieser beiden Wahrscheinlichkeiten für alle Schülerinnen und Schüler – unabhängig von der Schulgrösse – identisch, wenn in jeder Schule die gleiche Anzahl Schülerinnen und Schüler in die Stichprobe aufgenommen wird (mit Ausnahme *kleiner* bzw. *sehr kleiner* Schulen; vgl. 3.5). Erfahrungen aus PISA 2015 haben gezeigt, dass den Schulen in der Schweiz häufig Computerräume mit 15 bis 25 Arbeitsplätzen zur Verfügung stehen. Im Sinne einer optimalen Kosteneinsparung (Optimierung der Anzahl Testsitzungen) und um allzu grosse Klumpeneffekte zu vermeiden (vgl. 1.2), wurde eine *TCS* von 20 Schülerinnen und Schülern als adäquat betrachtet.

In *Kantonen mit einstufigen Stichprobenverfahren* hatten alle Schulen dieselbe Wahrscheinlichkeit von $p = 1$, um in die Stichprobe zu gelangen. Um auch hier möglichst identische Auswahlwahrscheinlichkeiten zu erzielen, wurde in den entsprechenden Schulen ein bestimmter Anteil der Schülerinnen und Schüler pro Schule ausgewählt. Wenn die Population eines Kantons beispielsweise 2'100 Schülerinnen und Schüler umfasste und die gewünschte Stichprobengrösse 700 entsprach, wurde in jeder Schule ein Drittel der Lernenden in die Stichprobe aufgenommen. Dies hatte zur Folge, dass in *Kantonen mit einstufigen Stichprobenverfahren* die Anzahl getesteter Schülerinnen und Schüler pro Schule stärker variierte als in *Kantonen mit zweistufigen Stichprobenverfahren*.

Sowohl bei der *TCS* als auch bei den im oberen Abschnitt erwähnten Anteilen handelt es sich lediglich um Richtwerte. Die definitive Anzahl gezogener Schülerinnen und Schüler pro Schule war zusätzlich abhängig von der jeweils zur Verfügung stehenden Infrastruktur. Aus ökonomischen Gründen bzw. um die Anzahl benötigter Testsitzungen zu verringern, wurde versucht, die Computerräume der teilnehmenden Schulen auszulasten und Testsitzungen mit einer sehr kleinen Anzahl Schülerinnen und Schüler zu verhindern. Dementsprechend wurden die Stichprobenrichtwerte der Anzahl Computer, die in einem Raum zur Verfügung standen, angepasst:

- War beispielsweise eine Erhebung mit 20 Schülerinnen und Schülern geplant, während die Schule einen Computerraum mit 24 Geräten zur Verfügung stellen konnte, wurden 24 Schülerinnen und Schüler in die Stichprobe aufgenommen.
- Umgekehrt wurde der Stichprobenumfang teilweise auch verringert. Standen nur wenige Geräte zur Verfügung, sodass der Stichprobenumfang um mehr als 20 Prozent hätte verringert werden müssen, wurden zusätzliche Testsitzungen

durchgeführt. So wurden beispielsweise in einer Schule mit einem geplanten Stichprobenumfang von 40 und einem Computerraum mit 25 Arbeitsplätzen schliesslich zwei Testsitzungen durchgeführt und 50 Schülerinnen und Schüler einbezogen.

Dies führte dazu, dass generell mehr Schülerinnen und Schüler in die Stichprobe aufgenommen wurden als ursprünglich geplant, die Anzahl Testsitzungen jedoch markant reduziert werden konnte. Andererseits führten die schulspezifischen Anpassungen bei der Anzahl getesteter Schülerinnen und Schüler zu einer Erhöhung in der Varianz der Stichprobengewichte. Die entsprechenden Designeffekte aufgrund variierender Stichprobengewichte wurden geschätzt (Le, Brick & Kalton, 2002) und betrugen für die grosse Mehrheit der Strata weniger als 1.1.[6] Darüber hinaus konnte diesen Designeffekten in den meisten Fällen mit dem leicht höheren Stichprobenumfang entgegengewirkt werden.

Nebst der Anpassung der Anzahl zu ziehender Schülerinnen und Schüler pro Schule an die verfügbare Computerinfrastruktur konnten drei weitere Aspekte zu einer erhöhten Varianz in den Stichprobengewichten führen:

- Aufgrund von Schätzungen auf Basis von Schullisten vergangener Schuljahre wich in *Kantonen mit zweistufigen Stichprobenverfahren* die *MOS* teilweise von der tatsächlichen Anzahl unterrichteter Schülerinnen und Schüler ab. Dies bedeutet, dass die Auswahlwahrscheinlichkeit der Schule nicht immer genau proportional zu den Schülerbeständen war. Im Rahmen von PISA werden die Schulgewichte bei extremen Abweichungen zwischen erwarteten und tatsächlichen Schülerbeständen deshalb angepasst (OECD, 2017, S. 118). Die entsprechenden Kriterien trafen bei der ÜGK 2016 jedoch in keinem Fall zu, weshalb die Schulgewichte stets auf Basis der ursprünglichen Auswahlwahrscheinlichkeit belassen wurden.
- Lag in Kantonen mit Schulstichproben die Anzahl unterrichteter Schülerinnen und Schüler unter dem *TCS*-Richtwert (*TCS* = 20), wurden sämtliche Schülerinnen und Schüler zum Test aufgeboten bzw. betrug die Auswahlwahrscheinlichkeit auf Schülerebene dann immer $p = 1$. Dementsprechend wurde bei *mittelgrossen* Schulen die *MOS* ebenfalls konstant gehalten (*MOS* = 20). Wurden jedoch *kleine* bzw. *sehr kleine* Schulen unterrepräsentiert (*TCS* = 10 bzw. *TCS* = 5, vgl. 3.5), führte dies zu einer leicht erhöhten Varianz in den Stichprobengewichten.
- Ähnlich wie das bei der in Abschnitt 3.2 beschriebenen Aufteilung der Schulen auf Strata der Fall war, wurden innerhalb von Schulen, die verschiedene

[6] Der Standardfehler vergrössert sich um den Faktor 1.05 (= Quadratwurzel aus 1.1).

Schulprogramme unterrichten, die Schülerinnen und Schüler unterschiedlichen Strata zugeordnet. Da sich die Auswahlwahrscheinlichkeiten einzelner Schülerinnen und Schüler zwischen diesen Strata unterschieden, führte die *optimale Aufteilung* ebenfalls zu einer erhöhten Varianz in den Stichprobengewichten. Das Ziel der *optimalen Aufteilung* war jedoch eine erhöhte Schätzpräzision, die den Designeffekt aufgrund variierender Stichprobengewichte mehr als wettmachen sollte. Die entsprechende Methode zur Stratifizierung innerhalb von Schulen wird im folgenden Abschnitt 4.3 dargestellt.

Nach den Erhebungen war es aufgrund von Verweigerungen und Abwesenheiten teilweise notwendig, die Stichprobengewichte anzupassen. Die entsprechenden Korrekturfaktoren werden in den Abschnitten 5.3 und 0 erläutert.

4.3 Stratifizierung auf Schülerebene

Ähnlich wie dies bei der Liste der wählbaren Schulen der Fall war (vgl. 3.1), wurden die Listen wählbarer Schülerinnen und Schüler nach bestimmten, mit der Schulleistung in Zusammenhang stehenden Merkmalen sortiert. In der überwiegenden Mehrheit der Listen wurden das kantonale Schulprogramm, das Geschlecht und die Klassenzugehörigkeit als Stratifizierungsvariablen verwendet.

In Schulen mit variierendem Schulprogramm wurden zudem bis zu drei Leistungsgruppen (explizite Strata) gebildet, damit die Auswahlwahrscheinlichkeit auf Grundlage des kantonalen Schulprogramms angepasst werden konnte. Analog zur Aufteilung der Schulstichprobe auf explizite Strata (vgl. 3.2) wurde eine Mischung aus *optimaler* und *proportionaler Aufteilung* angewendet. Dabei wurden in Strata mit der höchsten erwarteten Varianz im Zielmerkmal (z.B. Real- oder Sonderklassen) zusätzliche Schülerinnen und Schüler gezogen, während in Strata mit einer niedrigen zu erwartenden Varianz (z.B. Klassen mit gymnasialem Unterricht) der Stichprobenumfang leicht verringert wurde. In Sonderklassen (in Regelschulen) unterrichtete Schülerinnen und Schüler wurden dabei stets in die Gruppe mit den erwartet niedrigsten Leistungen integriert.

4.4 Ziehung der Schülerinnen und Schüler

Die Ziehung der Schülerinnen und Schüler wurde mithilfe der Software IBM SPSS Complex Samples 20 durchgeführt. Dazu wurden sämtliche Listen wählbarer Schülerinnen und Schüler separat eingelesen, die Listen nach den in Abschnitt 4.3 beschriebenen Merkmalen sortiert bzw. gruppiert und die Anzahl zu ziehender Schülerinnen und Schüler pro Stratum festgesetzt, bevor die entsprechende Stichprobe gezogen wurde.

Die Systematik der Methode der Stichprobenziehung auf Schülerebene ist mit der in Abschnitt 3.3 dargestellten Ziehung – mit Ausnahme des *PPS*-Verfahrens – vergleichbar: Mithilfe eines zuvor definierten Samplingintervalls wurde innerhalb jedes Stratums die Liste der wählbaren Schülerinnen und Schüler «durchgezählt» und die entsprechenden «Treffer» zur Studienteilnahme aufgeboten. Die daraus resultierende Liste der gewählten Schülerinnen und Schüler wurde anschliessend an die entsprechenden Schulen zur Information und Kontrolle geschickt.

5 Stichprobengewichte

Die zur Berechnung der Stichprobengewichte verwendeten Methoden orientieren sich an international etablierten Schulleistungsstudien wie PISA, *Trends in International Mathematics and Science Study* (TIMSS) oder *Progress in International Reading Literacy Studies (*PIRLS*)*. Die entsprechenden statistischen Theorien können beispielsweise bei Cochran (1977) oder Lohr (2010) nachgelesen werden.

Für die Analysen der im Rahmen der ÜGK gewonnenen Daten, der Berechnungen adäquater Schätzer des Stichprobenfehlers sowie um valide Schlussfolgerungen in Bezug auf die untersuchte Population treffen zu können, ist der Einbezug von Stichprobengewichten zwingend erforderlich (vgl. Anhang C). Für sämtliche an der ÜGK 2016 teilnehmenden Schülerinnen und Schüler wurden individuelle Schülergewichte (vgl. 5.5) sowie weitere Variablen berechnet, welche die Berechnung von Standardfehlern, Signifikanztests oder Konfidenzintervallen erlauben. Die Auswahlwahrscheinlichkeiten der teilnehmenden Schülerinnen und Schüler unterscheiden sich zum Teil beträchtlich. Da die einzelnen Schülerinnen und Schüler deshalb einen jeweils unterschiedlich grossen Anteil der Population repräsentieren, ist es unbedingt notwendig, die Gewichte in sämtlichen Analysen zu berücksichtigen.

Aufgrund der unterschiedlichen Auswahlsätze zwischen den Kantonen variieren die Gewichte auf nationaler Ebene relativ stark. Innerhalb der Kantone ist diese Varianz deutlich kleiner. Nebst den bereits in Abschnitt 4.2 gelisteten Gründen für variable Stichprobengewichte sind die schliesslich gültigen Stichprobengewichte auch aufgrund von *Non-Response* nicht für jede Schülerin und jeden Schüler innerhalb eines expliziten Stratums identisch:

- Gewählte Schülerinnen und Schüler, die nicht ausgeschlossen worden waren und aufgrund von Verweigerung, Krankheit oder sonstigen Gründen nicht an der Erhebung teilnahmen (*Non-Response* auf Ebene der Schülerinnen und Schüler), wurden mit Gewichtsanpassungen seitens der teilnehmenden Schülerinnen und Schüler kompensiert (*Non-Response-Adjustment; NRA*).
- Gezogene Schulen, die eine Studienteilnahme verweigerten und für die in nützlicher Frist keine Ersatzschulen gefunden werden konnten (*Non-Response* auf Schulebene), wurden mit Gewichtsanpassungen seitens der teilnehmenden Schulen kompensiert.

Das Gewicht des Schülers j oder der Schülerin j in der Schule i setzt sich aus dem Basisgewicht der Schule (w_{1i}), dem Basisgewicht innerhalb der Schule (w_{2ij}) und zwei Korrekturfaktoren zusammen (f_{1i} und f_{2ij}).

$$W_{ij} = w_{1i}w_{2i}f_{1i}f_{2ij}$$

Die einzelnen Komponenten dieses Produkts werden in den folgenden Abschnitten näher erläutert.

5.1 Basisgewicht der Schule

Für Schulen in *Kantonen mit einstufigen Stichprobenverfahren* oder *mit Vollerhebung* entspricht das Basisgewicht der Schule stets $w_{1i} = 1$, da alle Schulen an der Erhebung teilgenommen haben und die einzelnen Schulen grundsätzlich nicht andere Schulen in der Population repräsentieren. In *Kantonen mit zweistufigen Stichprobenverfahren* entspricht das Basisgewicht der Schule $w_{1i} = SI / MOS$ (vgl. 3.3 – 3.5), sofern $MOS < SI$. Wenn $MOS \geq SI$ gilt, dann beträgt das Basisgewicht der Schule $w_{1i} = 1$, da diese mit einer Wahrscheinlichkeit von $p = 1$ in die Stichprobe aufgenommen wurde. Beispielsweise würde eine Schule mit $MOS = 50$ in einem Kanton mit $MOS_{tot} = 5'000$ und einem Stichprobenumfang von 40 Schulen das Basisgewicht $w_{1i} = 2.5$ erhalten ($SI / MOS = MOS_{tot} / n / MOS = 5'000 / 40 / 50$) und würde dementsprechend 2.5 Schulen im entsprechenden expliziten Stratum repräsentieren.

5.2 Basisgewicht innerhalb der Schule

Das Basisgewicht innerhalb der Schule steht für die Anzahl Schülerinnen und Schüler, die ein gezogener Schüler oder eine gezogene Schülerin in seiner bzw. ihrer Schule repräsentiert. Da in zahlreichen Schulen explizite Strata abhängig vom kantonalen Schulprogramm gebildet und die Schulen innerhalb dieser Strata nicht auf eine proportionale Art und Weise gezogen wurden, ist das Basisgewicht nicht für alle Schülerinnen und Schüler einer Schule identisch (wie beispielsweise bei PISA). Das Basisgewicht innerhalb der Schule ergibt sich aus dem Kehrwert der Auswahlwahrscheinlichkeit.

Eine Oberstufenschule könnte beispielsweise 62 Schülerinnen und Schüler in Sekundarklassen sowie 68 Schülerinnen und Schüler in Realklassen unterrichten, wobei der Schule 25 Computer zur Verfügung stehen. Die entsprechende Schätzung der Varianz innerhalb dieser Gruppen bzw. Strata könnte dazu führen, dass 10 Schülerinnen und Schüler aus Sekundarklassen und 15 Schülerinnen und Schüler aus Realklassen in die Stichprobe aufgenommen werden. Die individuellen Auswahlwahrscheinlichkeiten betragen demnach 0.16 (10/62) auf Sekundarniveau und 0.22 (15/68) auf

Realniveau. Da die Basisgewichte innerhalb der Schule dem Kehrwert dieser Auswahlwahrscheinlichkeiten entsprechen, würden diese w_{2i} = 6.25 für Schülerinnen und Schüler aus Sekundarklassen und w_{2i} = 4.55 für Schülerinnen und Schüler aus Realklassen betragen.

5.3 Non-Response-Korrekturfaktor auf Schulebene

Ein kleiner Teil der Schulen verweigerte die Teilnahme an der ÜGK 2016 oder konnte aufgrund technischer Probleme bzw. ungenügender Infrastruktur nicht getestet werden. In Fällen rechtzeitig kommunizierter Teilnahmeverweigerungen oder technischer Probleme wurden Ersatzschulen (vgl. 3.4) kontaktiert. In einigen Fällen war es aus zeitlichen Gründen nicht mehr möglich, Ersatzschulen zur Teilnahme aufzubieten. Derartige Ausfälle wurden mittels *NRA* kompensiert.

Dazu wurden zunächst ähnliche Schulen innerhalb eines Kantons gruppiert, wobei häufig auf die für die Schulstichprobenziehung explizite Stratifizierung zurückgegriffen wurde. Zusätzlich wurde bei der Bildung solcher *NRA*-Zellen *(Non-Response-Adjustment cells)* darauf geachtet, dass (1) in den teilnehmenden Schulen gesamthaft eine Mindestanzahl von 60 Schülerinnen und Schülern unterrichtet wurden und (2) der resultierende Korrekturfaktor kleiner als 2 blieb. Ersteres war für allfällige Korrekturen auf Ebene der Schülerinnen und Schüler Voraussetzung. Korrekturfaktoren, die das Gewicht einer Schule mehr als verdoppelten, wurden verhindert, da diese zu einer zu starken Variabilität in den Gewichten und einer entsprechenden Zunahme der Stichprobenvarianz geführt hätten (vgl. OECD, 2017).

Innerhalb der definierten *NRA*-Zellen wurde der Korrekturfaktor gemäss folgender Formel berechnet:

$$f_{1i} = \frac{\sum_{k \in A(i)} w_{1k} n(k)}{\sum_{k \in P(i)} w_{1k} n(k)}$$

Die Summe im Zähler bezieht sich auf die mit dem Schulgewicht gewichtete Anzahl Schülerinnen und Schüler *sämtlicher gezogener (A)* Schulen und repräsentiert somit die Population der Schülerinnen und Schüler in der jeweiligen *NRA*-Zelle. Der Nenner hingegen beinhaltet die gewichtete Summe der Schülerinnen und Schüler aus *teilnehmenden* Schulen *(P)*. Der aus dieser Formel resultierende Korrekturfaktor wurde mit jedem Schulgewicht multipliziert, damit die gesamte Population der jeweiligen *NRA*-Zelle durch die teilnehmenden Schulen repräsentiert wurde. Non-Response-Korrekturfaktor auf Ebene der Schülerinnen und Schüler

Um nichtteilnehmende Schülerinnen und Schüler zu kompensieren, wurden – analog der *NRA* auf Schulebene – Gruppen ähnlicher Schülerinnen und Schüler gebildet. Dazu wurden Schülerinnen und Schüler mit identischen kantonalen

Schulprogrammen sowie demselben Geschlecht der gleichen *NRA*-Zelle zugeordnet. Zur Vermeidung allzu grosser Gewichtskorrekturen wurden ausschliesslich Zellen mit mindestens 15 teilnehmenden Schülerinnen und Schülern gebildet. Dies führte dazu, dass die entsprechenden Zellen nur in Ausnahmefällen innerhalb einer Schule gebildet werden konnten.[7] Wurden Schülerinnen und Schüler aus verschiedenen Schulen gruppiert, wurde darauf geachtet, dass die entsprechenden Schulen möglichst gleich gross waren. Innerhalb der *NRA*-Zellen wurden Korrekturfaktoren gemäss folgender Formel berechnet.

$$f_{2i} = \frac{\sum_{k \in A(i)} f_{1i} w_{1i} w_{2ik}}{\sum_{k \in P(i)} f_{1i} w_{1i} w_{2ik}}$$

Für jeden Schüler und jede Schülerin wurde das Produkt aus Basisgewicht der Schule (korrigiert für *Non-Response* auf Schulebene; $f_{1i}\ w_{1i}$) und Basisgewicht innerhalb der Schule (w_{2ik}) gebildet. Die Summe im Zähler der Formel enthält die Gewichte sämtlicher gezogener Schülerinnen und Schüler mit Ausnahme der ausgeschlossenen Fälle (z.B. kognitive Beeinträchtigung, vgl. 2.2). Aus der Division dieser Summe mit der Summe der Gewichte der teilnehmenden Schülerinnen und Schüler ergeben sich die Korrekturfaktoren für jede *NRA*-Zelle. Mit anderen Worten: Die Gewichte der teilnehmenden Schülerinnen und Schüler wurden dann erhöht, wenn hinsichtlich des Geschlechts, des kantonalen Schulprogramms und der Schulgrösse «ähnliche» Schülerinnen und Schüler abwesend waren. Ausgeschlossene Schülerinnen und Schüler wurden nicht mittels *NRA* kompensiert.

5.4 Verschiedene Gewichtungsvariablen im ÜGK-Datensatz

Die im Datensatz enthaltenen Gewichtungsvariablen sind am Präfix *smp_w* erkennbar. Das weiter oben beschriebene – für *Non-Response* korrigierte – Schülergewicht (W_{ij}) entspricht der Variablen *smp_w_nrastubw*. Auf Populationen schliessende Analysen mit auf der Schülerebene erhobenen Daten sollten stets mit dieser Variablen gewichtet werden.

Der bei der ÜGK 2016 eingesetzte Schülerfragebogen beruht auf einem modularen Ansatz mit zwei unterschiedlichen Fragebogenversionen, die hälftig und zufällig auf die Schülerstichprobe verteilt wurden (vgl. Sacchi & Oesch, 2017). Nebst Fragen, die

[7] Damit einer *NRA*-Zelle mindestens 15 teilnehmende Knaben oder Mädchen zugeordnet werden konnten, waren schulübergreifende Zellenbildungen nicht zu vermeiden. Nur in Schulen, in denen eine verhältnismässig grosse Anzahl Schülerinnen und Schüler getestet wurde (*Kantone mit einstufigen Stichprobenverfahren*), war die Bildung von *NRA*-Zellen innerhalb einer einzelnen Schule möglich.

in beiden Fragebogenversionen vorgekommen sind, hat die eine Hälfte der Schülerinnen und Schüler hauptsächlich Fragen zum Mathematikunterricht bearbeitet, während die andere Hälfte vor allem Fragen zu nachobligatorischen Schulübertritten beantwortet hat. Um einen Rückschluss entsprechender Ergebnisse auf die Population zu ermöglichen, wurden Schülergewichte für die beiden Fragebogenversionen berechnet (*smp_w_qmath*, *smp_w_qtree*). Zur Analyse der mit den beiden Fragebogenversionen erhobenen Daten sind stets die entsprechenden Gewichtungsvariablen einzubeziehen.

Bestimmte komplexere Analysemethoden erfordern den Einbezug von nach Untersuchungsebene getrennten Gewichten. So entspricht die Variable *smp_w_wscstu* dem Schülergewicht innerhalb der Schule und quantifiziert dadurch die Anzahl Schülerinnen und Schüler derselben Schule, die durch einen erhobenen Fall repräsentiert werden (ohne Korrektur für *Non-Response*). Die Variable *smp_w_nraschbw* widerspiegelt hingegen die Schulgewichte und entspricht somit dem Kehrwert der Auswahlwahrscheinlichkeit von Schulen. In Strata mit Schulen, welche die Teilnahme verweigert haben, wurden die Schulgewichte entsprechend korrigiert.

6 Berechnung der Stichprobenvarianz

Bei den im vorliegenden Bericht vorgestellten Stichprobenverfahren handelt es sich jeweils um eine Zufallsauswahl. Dadurch besitzt die Stichprobe eine Wahrscheinlichkeitsverteilung und ermöglicht die Anwendung inferentieller Statistik (z.B. Signifikanztests, Schätzer mit Konfidenzintervallen usw.; vgl. von der Lippe & Kladroba, 2002). Die Zufallskomponente führt dazu, dass die Anteile der Schülerinnen und Schüler, welche die Grundkompetenzen erreichen, aber auch beliebige, auf dem ÜGK-Datensatz beruhende Schätzer von der Stichprobe abhängig sind und somit je nach Auswahl von Schulen oder Schülerinnen und Schülern variieren können. Die Stichprobenvarianz beziffert, inwiefern sich die Ergebnisse ändern würden, wenn die Grundkompetenzen auf Basis anderer Schülerinnen und Schüler der Population überprüft worden wären.

Zur Berechnung der Stichprobenvarianz ist es zwingend erforderlich, das komplexe Stichprobendesign sowie die entsprechende Gewichtung zu berücksichtigen. Eine Übersicht entsprechender Methoden bieten beispielsweise Wolter (1985) oder Lee, Forthofer und Lorimor (1989). Sowohl aus praktischen als auch historischen Gründen haben sich im Rahmen nationaler (z.B. *National Assessment of Educational Progress; NAEP*) oder internationaler (z.B. *PISA, TIMSS* oder *PIRLS*) *Large-Scale-Assessments* die Replikationsverfahren als Standard zur Schätzung der Stichprobenvarianz etabliert (vgl. Rust, 2014). Dementsprechend wird in Abschnitt 6.1 ein bestimmtes Replikationsverfahren, auf welchem die im Datensatz enthaltenen *Replicate Weights* beruhen, näher vorgestellt. Dennoch kann für die Varianzschätzung auch auf Linearisierungsverfahren (Demnati & Rao, 2004) zurückgegriffen werden. Eine kurze Gegenüberstellung von Replikations- und Linearisierungsverfahren – auch in Abhängigkeit von statistischer Analysemethode und Art des Schätzers – bietet Valliant (2007).

Im Rahmen von Analysen, die sich auf Leistungsvariablen beziehen (z.B. Anteile der Schülerinnen und Schüler pro Kanton, die den Grundkompetenzen nicht genügen), sollte zusätzlich stets der mit den Tests verbundene Messfehler berichtet werden. Dieser kann mittels der *Plausible Values* (vgl. Angelone & Keller, 2019) der skalierten Leistungswerte berechnet werden. Der Standardfehler eines entsprechenden Schätzers setzt sich dementsprechend aus Stichprobenvarianz und Messfehler zusammen. Da Berechnungen mit *Plausible Values* unter gleichzeitiger Verwendung von Varianzschätzverfahren komplexe und rechenintensive Analysen darstellen, sind im Anhang C ein kurze Auswertungsbeispiele mithilfe von *SPSS* und einem *R*-Programmpaket dargestellt.

6.1 Die «Balanced Repeated Replication»-Methode

Die Grundidee von Replikationsverfahren besteht darin, die Stichprobenvarianz eines Ergebnisses zu berechnen, indem dieses mehrmals – mit einer jeweils

unterschiedlichen Gewichtung einzelner Studienteilnehmer – geschätzt wird. Die Stichprobenvarianz wird aus der Variabilität des mehrmals berechneten Ergebnisses abgeleitet. Die bei der ÜGK 2016 verwendete bzw. im ÜGK-Datensatz integrierte Methode zur Schätzung der Stichprobenvarianz nennt sich *Balanced Repeated Replication* (*BRR*; vgl. Rust, 1985; Rust & Rao, 1996). Analog zur Vorgehensweise bei PISA wurde die Variante des Verfahrens gewählt, die als *Fay's Methode* bekannt ist (vgl. Judkins, 1990). Der Datensatz ÜGK 2016 enthält 120 zusätzliche Gewichtsvariablen bzw. unterschiedliche Kombinationen von Schülergewichten (*Replicate Weights*), die jeweils neue Varianten der Stichprobe darstellen und sich durch eine veränderte Gewichtung der gezogenen Schülerinnen und Schüler von der anfänglichen Stichprobe unterscheiden. Die Stichprobenvarianz eines Ergebnisses wird beruhend auf der Variabilität der entsprechenden Werte zwischen den 120 neugebildeten Varianten der Stichprobe geschätzt. Konkret bedeutet dies, dass die Stichprobenvarianz beliebiger – mit der ÜGK 2016 berechneter – Schätzer mittels der folgenden Formel berechnet werden kann:

$$\mathrm{V}_{BRR}(X^*) = \frac{1}{30}\sum_{t=1}^{120}\{(X_t^* - X^*)^2\}$$

Dabei entspricht X^*_t der Schätzung des interessierenden Merkmals, beruhend auf der wiederholten Über- und Untergewichtung mithilfe der neu erstellten Varianten der Stichprobe, und X^* der Schätzung, basierend auf der anfänglichen Stichprobe (Ausgangsgewichtung). Die Erstellung der 120 alternativen Gewichtungen folgt einer bestimmten Methode, die im folgenden Abschnitt erläutert wird.

6.2 Berechnung der «Replicate Weights»

Die Berechnung der *Replicate Weights* beinhaltet die Bildung von Paaren aus Stichprobeneinheiten, die in der Summe stets dasselbe Gewicht behalten, bei denen die einzelnen Stichprobeneinheiten allerdings unterschiedlich stark gewichtet werden Im Rahmen der ÜGK 2016 wurden dazu teilnehmende Schulen (in *Kantonen mit zweistufigen Stichprobenverfahren*) bzw. Schülerinnen und Schüler (in *Kantonen mit einstufigen Stichprobenverfahren*) auf Grundlage der Stratifizierungsvariablen gepaart, bevor innerhalb jedes Paares eine Einheit stärker und eine Einheit schwächer gewichtet wurde. Dazu wurden die entsprechenden Gewichte innerhalb jedes Paares mit den Faktoren 0.5 bzw. 1.5 multipliziert (*Fay's Variante* der *BRR*-Methode; vgl. Judkins, 1990), sodass die Summe der Gewichte innerhalb jedes Schulpaares unverändert blieb. Dieses Vorgehen wurde 120-mal wiederholt, wobei jeweils neue Kombinationen aus über- und untergewichteten Einheiten entstanden. Im Detail enthielt die Berechnung der *Replicate Weights* für die ÜGK 2017 die folgenden Schritte:

- In *Kantonen mit zweistufigen Stichprobenverfahren* wurden auf Basis von expliziter und impliziter Stratifizierung Schulpaare gebildet. Die Paare repräsentieren sogenannte Varianzstrata, die lediglich zur Schätzung der Stichprobenvarianz dienen. Diese wurden in 116 grössere Varianzstrata zusammengefasst.
- Innerhalb der Stichprobenhälften wurde stets das Basisgewicht einer Schule (w_{1i}, vgl. 5.1) mit dem Faktor 1.5 übergewichtet, während dasjenige der anderen Schule mit dem Faktor 0.5 untergewichtet wurde. Die Verteilung dieser Gewichtungsfaktoren folgte einer Hadamardmatrix (orthogonale Matrix, vgl. Lee, Forthofer & Lorimor, 1989, S. 30), die in 120 unterschiedlichen Kombinationen von Stichprobenhälften resultierte.
- Bei einer ungeraden Anzahl Schulen innerhalb eines expliziten Stratums bildete jeweils die Schule mit der höchsten Ziehungswahrscheinlichkeit (kleinstes Basisgewicht) zusammen mit den Schulen aus den *certainty strata* (vgl. 3.3) eine eigene Gruppe. Für diese Gruppe wurden Stichprobenhälften nicht auf Schul-, sondern auf Schülerebene gebildet und ebenfalls 120 *Replicate Weights* berechnet.
- Analog zu den *Replicate Weights* auf Schulebene wurden für *Kantone mit einstufigen Stichprobenverfahren* sowie *Kantone mit Vollerhebungen* Paare auf Schülerebene gebildet. Die auf Basis der impliziten Stratifizierung gebildeten Paare wurden anschliessend zu 116 grösseren Gruppen zusammengefasst.
- Die Basisgewichte der Schülerinnen und Schüler wurden in unterschiedlicher Kombination wiederum hälftig um den Faktor 1.5 erhöht bzw. um den Faktor 0.5 reduziert. Hieraus ergaben sich erneut 120 *Replicate Weights*. Bei einer ungeraden Anzahl Schülerinnen und Schüler in einer Schule blieb das Basisgewicht eines einzelnen Schülers oder einer einzelnen Schülerin über die *Replicate Weights* hinweg gleich.
- Die 120 neu gebildeten Gewichte *(Replicate Weights)* wurden schliesslich gemäss den in Abschnitt 5.3 und 0 beschriebenen Korrekturen für Non-Response auf Schul- sowie Schülerebene separat angepasst.

7 Literatur

Angelone, D. & Keller, F. (2019). *ÜGK 2016 Mathematik. Technische Dokumentation zur Testentwicklung und Skalierung.* Aarau: Geschäftsstelle der Aufgabendatenbank EDK (ADB).

Cochran, W.G. (1977). *Sampling Techniques.* New York: John Wiley and Sons.

Demnati, A., & Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology, 30*, 17-26.

Judkins, D.R. (1990). Fay's Method of Variance Estimation. *Journal of Official Statistics*, *6*(3), 223-239.

Kauermann, G., & Küchenhoff H. (2011). *Stichproben*. Heidelberg: Springer.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Kish, L. (1992). Weighting for Unequal Pi. *Journal of Official Statistics, 8*, 183-200.

Kish, L. (1995). Methods for Design Effects. *Journal of Official Statistics, 11*, 55-77.

Le, T., Brick, M., & Kalton, G. (2002). Decomposing Design Effects. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association, 2007-2012.

Lee, S.L., Forthofer, R.N., & Lorimor, R.J. (1989). *Analyzing Complex Survey Data* (Sage University Paper series on Quantitative Applications in the Social Sciences, No. 07-064). Newbury Park, CA: Sage.

Lehtonen, R., & Pahkinen, E.J. (1995). *Practical Methods for Design and Analysis of Complex Survey. Chichester*: John Wiley and Sons.

Liu, J., Iannacchione, V., & Byron, M. (2002). Decomposing Design Effects for Stratified Sampling. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association, 2124-2126.

Lohr, S.L. (2010). *Sampling: Design and Analysis*. Boston, MA: Brooks/Cole.

Meinck, S. (2015). Computing Sampling Weights in Large-scale Assessments in Education. Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach. Retrieved December 21, 2017, from http://surveyinsights.org/?p=5353

OECD. (2017). PISA 2015 Technical Report. Paris: OECD Publishing.

Pham, G. (2019). *ÜGK 2017 – Technical report: Student questionnaire data*. St. Gallen: Pädagogische Hochschule St. Gallen.

Robitzsch, A. & Oberwimmer, K. (2019). *BIFIEsurvey: Tools for survey statistics in educational assessment. R package version 3.2-25*. Verfügbar unter: https://CRAN.R-project.org/package=BIFIEsurvey [3.05.2019].

Rust, K., & Rao, J.N.K. (1996). Variance Estimation for Complex Surveys Using Replication Techniques. *Survey Methods in Medical Research, 5*, 283-310.

Rust, K. (1985). Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*, *1*, 381-397.

Rust, K. (2014). Sampling, Weighting, and Variance Estimation in International Large-Scale Assessments. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis (pp. 117-153). Boca Raton, FL: CRC Press.

Rust, K. (1985). Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*, *1*, 381-397.

Sacchi, S., & Oesch, D. *ÜGK 2016: Assessment of mathematics skills. Documentation of questionnaire-based scales.* Bern: TREE, Universität Bern.

Valliant, R. (2007). An Overview of the Pros and Cons of Linearization versus Replication in Establishment Surveys. Papers presented at the ICES-III, 929–940.

von der Lippe, P. & Kladroba, A. (2002). Repräsentativität von Stichproben. *Marketing ZFP – Journal of Research and Management, 24*, 139–144.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer.

8 Anhang A: Kennzahlen zur Stichprobenziehung

8.1 Ausschluss- und Ausschöpfungsquoten

Wie in Kapitel 2 beschrieben, galten bestimmte Schülerinnen und Schüler der erwünschten Population als «nicht erreichbar» und hatten eine Wahrscheinlichkeit von $p = 0$ in die Stichprobe aufgenommen zu werden. Dies betraf einzelne Schülerinnen und Schüler aus Regelschulen, die von den zuständigen Lehrpersonen ausgeschlossen worden waren sowie sämtliche in Sonderschulen unterrichteten Schülerinnen und Schüler. Die entsprechenden Ausschlussquoten werden in der Tabelle A.1 getrennt nach Kanton aufgeführt. Die geschätzten Ausschöpfungsquoten widerspiegeln den Anteil der *erwünschten Population*, der durch die kantonalen Stichproben abgedeckt werden konnte (*ÜGK-Population*). Da über die ausgeschlossenen Schülerinnen und Schüler anhand der ÜGK 2016 keine Aussagen getroffen werden können, sollten die Ausschöpfungsquoten als Interpretationshilfe bei kantonalen Leistungsvergleichen herangezogen werden.

In zahlreichen Kantonen können in Sonderschulen unterrichtete Schülerinnen und Schüler nicht einem bestimmten Schuljahr zugeordnet werden. Die kantonalen Anteile in Sonderschulen unterrichteter Schülerinnen und Schüler wurden deshalb auf Basis der *SDL* aus dem Schuljahr 2015/16 geschätzt, indem in Sonderschulen unterrichtete 15-Jährige ins Verhältnis zur gesamten 15-jährigen Schülerschaft eines Kantons gesetzt wurden.

Tabelle A.1: Anzahl bzw. Anteile der bei der ÜGK 2016 ausgeschlossenen Schülerinnen und Schüler sowie geschätzte Ausschöpfungsquoten getrennt nach Kanton.

	In Regelschulen ausgeschlossene SuS getrennt nach Ausschlussgrund (absolute Anzahl)										In Regelschulen ausgeschlossene SuS (Total)		Ausgeschlossene SuS der erwünschten Population in %		Geschätzte Ausschöpfungsquote der erwünschten Population in %
Kanton	*a)*	*b)*	*c)*	*d)*	*e)*	*f)*	*g)*	*h)*	*i)*	*j)*	ungew.	gew.	Regel-schulen	Sonder-schulen	
AG	8	1	1	2	0	0	0	0	0	3	15	57	0.8	2.4	96.8
AI	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0	100.0
AR	3	0	0	0	0	0	0	0	1	0	4	4	0.8	4.3	94.9
BE_d	8	3	0	1	0	0	0	0	0	13	25	136	1.5	1.9	96.6
BE_f	3	2	0	0	0	0	0	0	0	0	5	5	0.6	0.4	99.0
BL	2	0	6	1	0	0	0	1	0	4	14	38	1.5	1.8	96.7
BS	1	1	7	2	0	0	0	0	1	1	13	23	1.6	1.6	96.8
FR_d	3	3	1	0	0	0	0	0	2	1	10	13	1.5	0.7	97.8
FR_f	10	0	1	0	0	0	0	0	3	4	18	54	2.0	2.2	95.9
GE	11	3	1	1	1	0	1	0	0	5	23	104	2.3	1.3	96.4
GL	0	0	0	0	0	0	0	0	0	0	0	0	0.0	2.8	97.3
GR	0	0	3	0	0	0	0	0	0	1	4	5	0.3	1.7	98.1
JU	5	0	0	0	0	0	0	0	0	7	12	12	1.5	1.2	97.3
LU	7	4	0	0	0	0	0	0	0	1	12	38	0.9	2.4	96.7
NE	10	12	1	3	1	1	0	0	2	5	35	73	3.9	3.0	93.2
NW	1	0	0	0	0	0	0	0	0	1	2	2	0.5	1.6	97.9
OW	1	1	2	0	0	0	0	0	0	1	5	5	1.1	1.3	97.5
SG	3	1	0	0	0	0	0	0	0	0	4	10	0.2	2.4	97.4
SH	1	0	1	0	0	0	1	0	1	0	4	4	0.6	2.4	97.1
SO	3	2	1	1	0	0	0	0	0	2	9	15	0.6	2.5	96.9
SZ	0	0	0	0	0	0	0	0	0	0	0	0	0.0	1.0	99.0
TG	1	2	0	0	0	0	0	0	1	1	5	13	0.5	2.4	97.1
TI	7	4	0	5	0	0	0	0	0	6	22	79	2.5	2.2	95.4
UR	1	0	0	0	0	0	0	0	0	1	2	2	0.5	0.9	98.6
VD	2	11	1	2	1	2	0	1	0	7	27	170	2.1	2.2	95.7
VS_d	20	1	0	0	0	0	0	0	0	0	21	21	2.6	1.3	96.1
VS_f	13	6	0	0	0	0	0	0	2	3	24	61	2.4	1.0	96.6
ZG	1	7	3	0	0	0	0	0	0	1	12	12	1.0	3.5	95.6
ZH	1	3	2	0	1	0	0	0	0	10	17	99	0.7	2.4	96.9
CH	**126**	**67**	**31**	**18**	**4**	**3**	**2**	**2**	**13**	**69**	**345**	**1'056**	**1.3**	**2.1**	**96.6**

Ausschlussgründe: a) Geringe Kenntnisse der Testsprache; *b)* Lernbehinderung; *c)* Kognitive Beeinträchtigung; *d)* Verhaltensbehinderung; *e)* Sprachbehinderung; *f)* Sehbehinderung; *g)* Hörbehinderung; *h)* Körperbehinderung; *i)* Mehrfachbehinderung; *j)* Andere Gründe.

Anmerkungen: In Sonderschulen unterrichtete Schülerinnen und Schüler können in den meisten Fällen nicht einem bestimmten Schuljahr zugeordnet werden. Die hier dargestellten Schätzungen beruhen auf Anteilen von SuS eines bestimmten Jahrgangs.

8.2 Rücklaufquoten auf Schulebene

Ein kleiner Teil der gezogenen bzw. zur Erhebung aufgebotenen Schulen hat die Teilnahme an der ÜGK 2016 verweigert. In *Kantonen mit zweistufigen Stichprobenverfahren* wurden deshalb Ersatzschulen gezogen (vgl. 3.4), die teilweise für verweigernde Schulen eingesprungen sind. Entsprechende Kennzahlen sowie Rücklaufquoten sind getrennt nach Kanton in Tabelle A.2 dargestellt. Zusätzlich zu ungewichteten Rücklaufquoten – die sich aus dem Verhältnis zwischen den Summen teilnehmender und aufgebotener Schulen ergeben – wurden die Schulen mit dem Basisgewicht w_{1i} (vgl. 1.4) sowie dem erwarteten Schülerbestand gewichtet.

Tabelle A.2: Anzahl aufgebotener, erhobener und ersetzter Schulen sowie Rücklaufquoten getrennt nach Kanton

	Anzahl Schulen			Mit Schülerbestand gewichtete Anzahl Schulen		Rücklaufquoten in %		
Kanton	aufgeboten	erhoben (ohne Ersatzschulen)	ersetzt	aufgeboten	erhoben	ungew. (ohne Ersatzschulen)	ungew. (mit Ersatzschulen)	gew. (mit Ersatzschulen)
AG	57	45	9	6'643	6'506	78.9	94.7	97.9
AI	4	4	-	203	203	100.0	-	100.0
AR	13	13	-	527	527	100.0	-	100.0
BE_d	65	61	3	9'017	8'884	93.8	98.5	98.5
BE_f	17	14	-	838	829	82.4	-	98.9
BL	29	24	-	2'684	2'585	82.8	-	96.3
BS	16	15	-	1'475	1'420	93.8	-	96.3
FR_d	9	9	-	914	914	100.0	-	100.0
FR_f	15	15	-	2'849	2'849	100.0	-	100.0
GE	28	24	-	4'731	4'561	85.7	-	96.4
GL	9	9	-	394	394	100.0	-	100.0
GR	61	61	-	1'849	1'849	100.0	-	100.0
JU	12	11	-	848	842	91.7	-	99.3
LU	55	55	0	4'119	4'119	100.0	100.0	100.0
NE	17	16	-	2'003	1'999	94.1	-	99.8
NW	10	10	-	430	430	100.0	-	100.0
OW	11	11	-	450	450	100.0	-	100.0
SG	55	55	0	5'493	5'493	100.0	100.0	100.0
SH	21	21	-	725	725	100.0	-	100.0
SO	15	15	-	1'978	1'978	100.0	-	100.0
SZ	23	22	-	1'647	1'646	95.7	-	99.9
TG	22	20	-	2'752	2'735	90.9	-	99.4
TI	42	41	-	3'314	3'295	97.6	-	99.4
UR	11	11	-	354	354	100.0	-	100.0
VD	55	49	2	7'547	7'392	89.1	92.7	97.9
VS_d	17	17	-	840	840	100.0	-	100.0
VS_f	29	26	-	2'674	2659	89.7	-	99.4
ZG	16	16	-	1'212	1212	100.0	-	100.0
ZH	89	83	2	13'336	12'850	93.3	95.5	96.4
CH	**823**	**773**	**16**	**81'846**	**80'540**	**93.9**	**95.9**	**98.4**

8.3 Rücklaufquoten auf Schülerebene

Schülerinnen und Schüler, die nicht mindestens eine gültige Antwort in den Mathematikaufgaben oder im Fragebogen gegeben haben und gleichzeitig nicht von der Erhebung ausgeschlossen wurden (vgl. die Tabellen A.1 und A.2), galten als abwesend. Für die grosse Mehrheit der abwesenden Schülerinnen und Schüler wurde von Seiten der Schulen kein Abwesenheitsgrund kommuniziert. Abwesenheiten aufgrund von technischen Problemen oder Verweigerung der Eltern sind äusserst selten vorgekommen. Um Rücklaufquoten auf Schülerebene zu berechnen, wurden die Summen von erhobenen sowie von allen zur Erhebung aufgebotenen Schülerinnen und Schülern (ohne Ausschlüsse) ins Verhältnis gesetzt. Entsprechende ungewichtete und gewichtete (Gewichtung mit w_{1i}, w_{2ij} sowie f_{1i}; vgl. Kapitel 5) Rücklaufquoten können Tabelle A.3 entnommen werden.

Tabelle A.3: Ungewichtete und gewichtete Anzahl erhobener, abwesender Schülerinnen und Schüler sowie entsprechende Rücklaufquoten

	Ungewichtete Anzahl Schülerinnen und Schüler		Gewichtete Anzahl Schülerinnen und Schüler		Rücklaufquoten in %	
Kanton	erhoben	abwesend	erhoben	abwesend	ungewichtet	gewichtet
AG	1'112	92	6'417	486	92.4	93.0
AI	202	2	202	2	99.0	99.0
AR	482	32	482	32	93.8	93.8
BE_d	1'093	134	7'771	1'035	89.1	88.3
BE_f	749	56	757	57	93.0	93.0
BL	703	62	2'386	197	91.9	92.4
BS	628	118	1'229	201	84.2	86.0
FR_d	717	27	861	33	96.4	96.3
FR_f	743	32	2'646	104	95.9	96.2
GE	665	92	4'068	463	87.8	89.8
GL	376	15	376	15	96.2	96.2
GR	925	52	1'740	89	94.7	95.1
JU	683	126	688	127	84.4	84.4
LU	1'093	95	3'692	310	92.0	92.2
NE	648	58	1'751	132	91.8	93.0
NW	410	15	410	15	96.5	96.5
OW	437	6	437	6	98.6	98.6
SG	1'137	61	4'555	250	94.9	94.8
SH	666	42	666	42	94.1	94.1
SO	746	64	2'200	189	92.1	92.1
SZ	755	46	1'545	89	94.3	94.5
TG	996	45	2'591	109	95.7	96.0
TI	695	49	3'022	182	93.4	94.3
UR	357	9	357	9	97.5	97.5
VD	1'014	73	7'451	510	93.3	93.6
VS_d	784	28	784	28	96.6	96.6
VS_f	755	35	2'434	126	95.6	95.1
ZG	1'142	55	1'142	55	95.4	95.4
ZH	1'710	173	12'131	1'178	90.8	91.1
CH	**22423**	**1694**	**74788**	**6'070**	**93.0**	**92.5**

9 Anhang B: Zusatzinformationen zu Schulstichproben

9.1 Für Stratifizierung verwendete Schulmerkmale

In den Kantonen AG, BE_d, LU, SG, VD und ZH wurden zweistufige Stichprobenverfahren eingesetzt: Vor der Ziehung einzelner Schülerinnen und Schüler wurden zunächst Schulhäuser mithilfe eines *PSS*-Verfahrens ausgewählt (vgl. 3.3).

Die Schulen wurden vor der Stichprobenziehung geschichtet und sortiert (vgl. 3.1). Die Schulmerkmale, die zur Bildung entsprechender expliziter und impliziter Strata verwendet wurden, die Anzahl gezogener und erhobener Schulen sowie die geschätzten Schülerbestände pro Stratum, sind in Tabelle B.1 dargestellt. Dabei sind folgende Punkte zu beachten:

- Aufgrund von Teilnahmeverweigerungen, Schulzusammenschlüssen oder Schulausschlüssen wegen einer sehr geringen Zahl im 11. Schuljahr unterrichteter Schülerinnen und Schüler (vgl. 3.5) bestehen teilweise Differenzen zwischen der Anzahl gezogener und der Anzahl erhobener Schulen.
- In expliziten Strata, in welchen ausschliesslich ein bestimmtes Schulprogramm unterrichtet wird (z.B. Kantonsschulen in LU, SG oder ZH), wurde nur die geschätzte Anzahl unterrichteter Schülerinnen und Schüler (Schulgrösse) zur impliziten Sortierung herangezogen. In den anderen Strata diente häufig der Anteil bestimmter Schulprogramme als Sortierungsvariable.
- In ZH und VD wurden auch explizite Strata – beruhend auf Anteilen unterrichteter Schulprogramme – gebildet. So wurden beispielsweise in VD die öffentlichen Schulen auf zwei Strata mit einer vergleichbaren Anzahl Schulen aufgeteilt, wobei der Anteil im Programm «voie prégymnasiale» unterrichteter Schülerinnen und Schüler als Unterscheidungsmerkmal diente.
- Im Stratum für private Schulen des Kantons AG haben sowohl beide erstgezogenen Schulen als auch sämtliche Ersatzschulen die Teilnahme an der ÜGK 2016 verweigert.
- Im Kanton SG wurden sämtliche Kantonsschulen in die Stichprobe aufgenommen. Im entsprechenden Stratum wurde also kein Stichprobenverfahren angewendet. Dieses Vorgehen ist hauptsächlich auf die verhältnismässig hohen Schülerbestände in Kantonsschulen zurückzuführen.

Tabelle B.1: Explizite Schulstrata, dazugehörige implizite Stratifizierungsvariablen für *Kantone mit zweistufigen Stichprobenverfahren* sowie Schul- und Schülerbestände

Kanton	*Explizite Strata*	*Implizite Stratifizierungsvariablen*	*a)*	*b)*	*c)*	*d)*
AG	Bezirksschulen	Schulgrösse	15	15	41	2'616
	Sekundarschulen	Schulgrösse	7	7	24	701
	Sekundar- und Realschulen	Anteile Schulprogramme, Schulgrösse	32	32	73	3'162
	Private Schulen	Schulgrösse	2	0	9	76
BE_d	Gymnasien	Kant. Schulprogramm, Schulmodell, Schulgrösse	8	8	22	1'658
	Sekundar- und Realschulen	Anteil Schulprogramme, Schulmodell, Schulgrösse	40	39	128	6'039
	Realschulen	Kant. Schulprogramm, Schulmodell, Schulgrösse	15	12	108	779
	Private Schulen	Kant. Schulprogramm, Subventionierungsgrad, Schulgrösse	6	5	35	726
LU	Kantonsschulen	Schulgrösse	7	7	8	783
	Gemischte öffentliche Schulen	Schulmodell, Anteile Leistungsniveaus, Schulgrösse	46	46	56	3'153
	Private Schulen	Subventionierungsgrad, Schulgrösse	2	2	9	161
SG	Kantonsschulen	*Alle Kantonsschulen fielen in die Stichprobe*	5	5	5	927
	Gemischte öffentliche Schulen	Kant. Schulprogramm, Schulgrösse	47	46	84	4'389
	Private Schulen	Kant. Schulprogramm, Subventionierungsgrad, Schulgrösse	4	4	17	252
VD	Hoher Anteil «voie prégymnasiale»	Anteil Schulprogramme, Schulgrösse	25	24	42	4'043
	Hoher Anteil «voie générale»	Anteil Schulprogramme, Schulgrösse	27	25	39	3'455
	Private Schulen	Subventionierungsgrad, Schulgrösse	2	2	21	459
ZH	Kantonsschulen	Schulgrösse	14	14	19	3'050
	Sekundarschulen (tiefer Anteil B & C)	Anteile Leistungsniveaus, Schulgrösse	34	31	92	4'989
	Sekundarschulen (hoher Anteil B & C)	Anteile Leistungsniveaus, Schulgrösse	35	34	70	4'517
	Private Schulen	Anteil Sonderklassen, Kant. Schulprogramm, Schulgrösse	9	6	61	978

Legende: a) Anzahl gezogener Schulen; *b)* Anzahl erhobener Schulen; *c)* Anzahl Schulen in Population; *d)* Geschätzte Anzahl Schülerinnen und Schüler in Population (Statistik der Lernenden 2014/15).

Anmerkungen: Aufgrund von Schulverweigerungen, Schulzusammenschlüssen oder Schulausschlüssen wegen einer sehr geringen Zahl im 11. Schuljahr unterrichteter Schülerinnen und Schüler kann es Differenzen zwischen der Anzahl gezogener und der Anzahl erhobener Schulen geben.

Wurden in Schulen mehrere kantonale Schulprogramme angeboten, wurden die Schulen nach den prozentualen Anteilen bestimmter Programme (z.B. Anteil Realschüler) sortiert.

9.2 Umgang mit kleinen und sehr kleinen Schulen

Enthielten Strata Schulen mit weniger als 20 Schülerinnen und Schülern, bestand das Risiko, dass die Anzahl gezogener Schülerinnen und Schüler dem erwünschten Stichprobenumfang nicht genügt. Aus diesem Grund wurden vor der Schulziehung die Listen wählbarer Schulen mithilfe der folgenden Schritte analysiert (vgl. auch 3.5 sowie OECD, 2017, S. 77):

- Innerhalb jedes expliziten Stratums wurde der gesamthafte Schülerbestand prozentual auf *sehr kleine* Schulen (K1; Schülerbestand < 3), *kleine Schulen* (K2; Schülerbestand ≥ 3 und < 10), *mittelgrosse* Schulen (M; Schülerbestand ≥ 10 und < 20) und *grosse* Schulen (G; Schülerbestand ≥ 20) aufgeteilt, sodass K1 + K2 + M + G = 1 galt.
- Wenn K1 + K2 ≤ 0.01 war und somit über 1 Prozent des Schülerbestands im Stratum ausmachte, wurden *kleine* sowie *sehr kleine* Schulen unterrepräsentiert (*TCS* angepasst) und die Anzahl zu ziehender Schulen angehoben.
- Wenn K1 + K2 < 0.01 und M ≥ 0.04 galt, dann wurde lediglich die Anzahl zu ziehender Schulen angehoben.
- Wenn K1 + K2 < 0.01 und M < 0.04 galt, dann wurden keine Anpassungen vorgenommen.

Falls die Anzahl zu ziehender Schulen angehoben werden musste, wurden neue Schulstichprobenumfänge beruhend auf den folgenden Formeln berechnet:

- Berechnung des Faktors L = 1 + 3 (K1) / 4 + (K2) / 2.
- Berechnung der mittleren Schulgrösse für *mittelgrosse* Schulen (MENR), *kleine* Schulen (K2ENR) und *sehr kleine* Schulen (K1ENR).
- Der minimale Schulstichprobenumfang für *grosse* Schulen entsprach der ursprünglichen Anzahl zu ziehender Schulen multipliziert mit G und L.
- Der minimale Schulstichprobenumfang für *mittelgrosse* Schulen entsprach (N/2 · M · L) / K2ENR, wobei N den Schülerbestand in *mittelgrossen* Schulen repräsentiert.
- Der minimale Schulstichprobenumfang für *sehr kleine* Schulen entsprach (N/4 · K1 · L) / K1ENR, wobei N den Schülerbestand in *sehr kleinen* Schulen repräsentiert.

10 Anhang C: Auswertungshinweise

In den folgenden Abschnitten wird am Beispiel der Berechnung kantonaler Anteile an Schülerinnen und Schülern, welche die Grundkompetenzen erreichen, der adäquate Einbezug von Schülergewichten, *Plausible Values* und *Replicate Weights* mit den Analyseprogrammen SPSS und R kurz illustriert.

10.1 SPSS

Ohne Zusatzsoftware sind Auswertungen unter der Berücksichtigung von *Plausible Values* und *Replicate Weights* in SPSS nicht möglich. Ein praktisches Zusatzpaket, das entsprechende SPSS-Makros inklusive Benutzeroberfläche enthält, wird in Form des *IDB Analyzers* von der International Association for the Evaluation of Educational Achievement (IEA) angeboten. Dieses Softwarepaket wurde entwickelt, um die Auswertung von Daten, die im Rahmen von Large-Scale-Assessments (z.B. PISA oder TIMMS) erhoben wurden, mit SPSS zu erleichtern. Eine speziell für die Auswertung der ÜGK 2016 angepasste Version des *IDB Analyzers* liegt der Datenauslieferung bei.

Zur Analyse von Daten aus der ÜGK 2016 mit SPSS muss nach dem Start des *IDB Analyzers* die entsprechende Software und das *Analysis Module* ausgewählt werden (die Optionen *SAS* sowie *Merge Module* wurden für die Analyse von ÜGK-Daten nicht angepasst). Anschliessend wird die Benutzeroberfläche dargestellt, die in die Schritte 1 bis 5 gegliedert ist:

1. Im ersten Schritt wird der zu untersuchende Datensatz definiert. Dabei ist darauf zu achten, dass der entsprechende Dateiname bestimmte Sonderzeichen nicht enthält.
2. In Schritt 2 wird *Analysis Type* ÜGK/COFO/VECOF ausgewählt und die Analysemethode (*Statistic Type*) festgelegt. Da hier kantonale Anteile an Schülerinnen und Schülern, welche die Grundkompetenzen erreichen, berechnet werden sollen, wird die Option *Benchmarks* ausgewählt. Abhängig **von** der hier ausgewählten Analysemethode wird die hiermit generierte SPSS Syntax auf ein bestimmtes Makro verwiesen, das mit dem *IDB Analyzer* installiert wurde.
3. In Schritt 3 werden die Analysevariablen angegeben. Im Fenster auf der linken Seite werden sämtliche im Datensatz enthaltenen Variablen (Namen und Beschreibung inklusive Suchfunktion) dargestellt. Im rechten Fenster werden abhängig von der ausgewählten Analysemethode diverse Gruppen von Analysevariablen dargestellt. Variablen in der Liste links können markiert und mithilfe der Pfeile in die Analysevariablen auf der rechten Seite verschoben werden. Bei der Berechnung von *Benchmarks* werden auf der rechten Seite die

Gruppen *Grouping Variables*, *Plausible Values*, *Achievement Benchmarks* und *Weight Variable* angezeigt. Um Resultate getrennt nach Kanton zu erhalten, ist es notwendig, die Variable *id_canton* den *Grouping Variables* hinzuzufügen. Die hier bereits vorhandene Variable *id_country* ist für eine erfolgreiche Auswertung notwendig, im Resultat aber irrelevant. Wird das Fenster mit den *Plausible Values* markiert, werden im linken Fenster die bereits gruppierten (20 Werte pro Skala) *Plausible Values* dargestellt. Für die hier beschriebene Analyse wird *PL_PVM_01-20* (*PL* steht für *Proficiency Level*, *M* steht für die Gesamtskala Mathematik) ausgewählt und in die *Plausible Values* auf die rechte Seite verschoben. Im Fenster *Achievement Benchmarks* wird schliesslich das Kriterium, auf dessen Basis die Anteile berechnet werden, eingetragen. Da *PL_PVM_01-20* die Werte 0 (Grundkompetenzen nicht erreicht) oder 1 (Grundkompetenzen erreicht) annehmen kann, wird eine 1 in die *Achievement Benchmarks* eingetragen. Das in Kapitel 5 beschriebene Schülergewicht *smp_w_nrastubw* wurde bereits automatisch dem Fenster *Weight Variable* hinzugefügt. Die zur Schätzung des Stichprobenfehlers benötigten *Replicate Weights* werden vom *IDB Analyzer* automatisch im Hintergrund berücksichtigt.

4. Im vierten Schritt wird der Dateipfad festgelegt, unter welchem die auf Basis der bisher getätigten Schritte erstellte SPSS Syntax und die entsprechenden Ergebnisse (Output) gespeichert werden.

5. Abschliessend wird mit dem Button *Start SPSS* die SPSS Software gestartet und die entsprechende Syntax in einem neuen Fenster generiert. Es besteht die Möglichkeit, die einzelnen Optionen, die in den bisherigen Schritten getätigt wurden, direkt in der SPSS Syntax anzupassen. Das Ausführen der SPSS Syntax öffnet ein neues Fenster (Output), in welchem die Ergebnisse dargestellt werden.

Unten rechts auf der Benutzeroberfläche des *IDB Analyzers* befindet sich ein *Help Button* mit einer Verknüpfung zu ei§ner ausführlichen Anleitung zur Software. Dort werden beispielsweise Korrelationen, Regressionen oder die Berechnung von Perzentilen mit dem *IDB Analyzer* genauer beschrieben.

10.2 *BIFIE-Survey* (*R*-Paket)

Für die Programmierumgebung *R* wurden zahlreiche Pakete entwickelt, mit Hilfe welcher sich Analysen auf Grundlage komplexer Stichprobendesigns durchführen lassen. Ein speziell für *Large-Scale-Assessments* im Bildungsbereich entwickeltes *R*-Paket ist *BIFIE-Survey* (Robitzsch & Oberwimmer, 2019). Im Vergleich zu SPSS und dem *IDB Analyzer* ermöglich *R* flexiblere und schnellere Auswertungen, es werden jedoch Grundkenntnisse in *R* vorausgesetzt.

Die Grundidee der Funktionsweise von BIFIE-Survey wird folglich am Beispiel der Berechnung kantonaler Anteile an Schülerinnen und Schülern, welche die Grundkompetenzen erreichen, vorgestellt.

Nach dem Start von *R*, ist es in einem ersten Schritt notwendig das Paket *BIFIE-Survey* zu installieren und zu laden.

```
install.packages("BIFIEsurvey")
library(BIFIEsurvey)
```

Liegt der Datensatz im *sav*-Format (SPSS) vor, lässt sich dieser direkt mit dem *R*-Paket *foreign* in die R-Umgebung einlesen.

```
install.packages("foreign")
library(foreign)
uegk16 <- read.spss("Datensatz.sav", to.data.frame=TRUE)
```

Anschliessend ist es notwendig die *Replicate Weights* festzulegen. Dazu bietet sich der *R*-Befehl *grep* an, mit welchem sich mehrere Variablen mit demselben Präfix ansteuern lassen (Präfix für *Replicate Weights*: *smp_w_nrasturw*).

```
reps.col <- grep("smp_w_nrasturw", names(uegk16))
```

Sämtliche Berechnungen in *BIFIE-Survey* greifen auf *BIFIE.dat*-Objekte zu, die zunächst erstellt werden müssen. Dabei werden die Datensätze eingelesen und Schülergewichte sowie *Replicate Weights* definiert.

```
uegk16.dat <- BIFIE.data(data=uegk16,
wgt=uegk16$smp_w_nrastubw,
wgtrep=uegk16[,reps.col],
fayfac=(1/120*4))
```

Schliesslich können mithilfe der *Plausible Values* für die Gesamtskala Mathematik und dem Befehl *BIFIE.univar* die kantonalen Anteile an Schülerinnen und Schüler, welche die Grundkompetenzen erreicht haben, geschätzt werden.

```
desc1 <- BIFIEsurvey::BIFIE.univar(uegk16.dat,
vars=c("PL_PVM_"),
group="id_canton_n")
summary(desc1)
desc1$stat
```

Die Dokumentation (Robitzsch & Oberwimmer, 2019) bzw. Hilfefunktion in *R* zum *BIFIE-Survey*-Paket enthält wertvolle Hinweise zu zahlreichen weiteren Analysenmöglichkeiten.