

bifie | Bildung

 standards

Stichprobenziehung in Bildungsstandardüberprüfungen und Pilotierungen

Technische Dokumentation – BIST-Ü

*Ann Cathrice George
Ursula Itzlinger-Bruneforth
Matthias Trendtel*

Bundesinstitut
 bifie



Bundesinstitut für Bildungsforschung, Innovation & Entwicklung
des österreichischen Schulwesens
Alpenstraße 121 / 5020 Salzburg
www.bifie.at

**Stichprobenziehung in Bildungsstandardüberprüfungen und Pilotierungen
Technische Dokumentation – BIST-Ü**

BIFIE | Department Bildungsstandards & Internationale Assessments (BISTA),
Salzburg 2015

Der Text sowie die Aufgabenbeispiele dürfen für Zwecke des Unterrichts in österreichischen Schulen sowie von den Pädagogischen Hochschulen und Universitäten im Bereich der Lehreraus-, Lehrerfort- und Lehrerweiterbildung in dem für die jeweilige Lehrveranstaltung erforderlichen Umfang von der Homepage (www.bifie.at) heruntergeladen, kopiert und verbreitet werden. Ebenso ist die Vervielfältigung der Texte und Aufgabenbeispiele auf einem anderen Träger als Papier (z. B. im Rahmen von Power-Point-Präsentationen) für Zwecke des Unterrichts gestattet.

Inhaltsverzeichnis

3 1 Theoretische Eigenschaften von Stichproben

4 2 Zu ziehende Stichproben

- 4 2.1 BIST-Ü-Stichproben
- 5 2.2 Stichproben für Pilotstudien
- 6 2.3 Zielpopulation und Datengrundlage

7 3 Stratifizierung

8 4 Stichprobenziehung

- 8 4.1 Schulenziehung
 - 9 4.1.1 Samplingintervalle
 - 9 4.1.2 Identifikation der gezogenen Schulen
 - 10 4.1.3 Ziehungswahrscheinlichkeiten der Schulen
 - 10 4.1.4 Identifikation von Reserveschulen
 - 11 4.1.5 Overlapcontrol S7 und S3
- 11 4.2 Klassenziehung
 - 12 4.2.1 Auswahl der Klassen an gezogenen Schulen
 - 12 4.2.2 Ziehungswahrscheinlichkeiten der Klassen
- 12 4.3 Schülerziehung für Teilstudien
 - 13 4.3.1 Auswahl der Schüler für Teilstudien
 - 13 4.3.2 Ziehungswahrscheinlichkeiten der Schüler für Teilstudien

14 5 Gewichtung

- 14 5.1 Schülergewichte
- 15 5.2 Non-Response Adjustment

16 Literatur



Im Rahmen der vom BIFIE durchgeführten Schulleistungsstudien gibt es zwei Arten von Stichproben: Zum einen wird in den Bildungsstandardüberprüfungen (BIST-Ü) eine Stichprobenziehung zur Aufteilung der Zielpopulation in Teilstudien durchgeführt, die verschiedenen Zielen dienen (Abschnitt 2). Zum anderen werden Stichproben für Pilotstudien gezogen (Abschnitt 2.2). Um in allen Fällen Stichproben mit „guten“ Eigenschaften zu erhalten (Abschnitt 1), werden theoretisch fundierte Verfahren eingesetzt (vgl. z. B. Cochran, 1977; Kish, 1965), die sich bereits in internationalen Schulleistungsstudien wie PISA, PIRLS und TIMSS etabliert haben (OECD, 2012a; Joncas, 2007, 2008).

Die Stichprobenziehungen erfolgen in einem zweistufigen Verfahren nach der „*probability proportional to size sampling*“-Methode (Abschnitt 4.1) in expliziten und impliziten Strata (Abschnitt 3). Die erste Stufe der Stichprobenziehung beinhaltet das Schulensampling (Abschnitt 4.1). In der zweiten Stufe werden in den gezogenen Schulen zufällig Klassen ausgewählt (Abschnitt 4.2). Abschließend wird durch den Einsatz von Schülergewichten sichergestellt, dass jeder Schüler mit gleicher Wahrscheinlichkeit in die Stichprobe eingeht und somit die gleiche Anzahl von Schülern in der Zielpopulation repräsentiert (Abschnitt 5).

1 Theoretische Eigenschaften von Stichproben

Das Ziel einer Stichprobenziehung ist, dass nicht alle Individuen (Sampling Units, hier: Schüler) einer Population befragt werden müssen, sondern die Eigenschaften der Sampling Units in der gewählten Stichprobe stellvertretend für die Eigenschaften der ganzen Zielpopulation (kurz: Population) stehen. Das heißt, wird ein Wert (z. B. der Mittelwert einer Variable) anhand der Stichprobe geschätzt, so kann man die Güte dieser Schätzung anhand ihres Abstands zum wahren, aber unbekanntem Wert der Variable in der Population bewerten. Dieser Abstand wird Sampling Bias genannt. Im Folgenden sind Eigenschaften von Stichproben genannt, mit denen der Sampling Bias gering gehalten werden kann:

- (1) Zur Schätzung der interessierenden Werte aus der Stichprobe werden unverzerrte Schätzer mit hoher Präzision (niedriger Varianz) eingesetzt.
- (2) Es wird eine repräsentative Stichprobe gezogen, die die Eigenschaften der Population widerspiegelt. Um die Repräsentativität der BIST-Ü-Stichproben sicherzustellen, erfolgen die Stichprobenziehungen in expliziten und impliziten Strata (Abschnitt 3).
- (3) Non-Response Bias, d. h. Antworten von bestimmten Schülern in der Stichprobe fallen aus (z. B. weil der Schüler am Testtag krankheitsbedingt fehlt). Diese Antwortausfälle können zu verzerrten Stichprobenschätzern im Sinne von Punkt (1) führen und werden daher mit Gewichten adjustiert (Abschnitt 5.)

- (4) Eine zu kleine Stichprobe. Dies kann im Falle der Stichprobengrößen bei den BIST-Ü vernachlässigt werden.

2 Zu ziehende Stichproben

In diesem Abschnitt werden die Anforderungen an die BIST-Ü- und Pilot-Stichproben beschrieben. Während die BIST-Ü-Stichprobenziehung auf einer Vollerhebung basiert und die Zielpopulation in Teilstudien aufgesplittet wird, fallen die Stichproben für Pilotstudien deutlich kleiner aus.

2.1 BIST-Ü-Stichproben

Die BIST-Ü sind gesetzlich als Vollerhebungen aller österreichischen Schüler im jeweiligen Zielunterrichtsfach und in der jeweiligen Zielschulstufe verankert. Sie verfolgen drei Ziele: (1) Unterstützung der Schulqualitätsentwicklung durch Rückmeldung auf Lerngruppen- und Schulebene, (2) Rückmeldung der in der BIST-Ü erreichten Kompetenzen an die getesteten Schüler und (3) ein längerfristiges System-Monitoring.

Diese drei Ziele stellen unterschiedliche Anforderungen an den Itempool. Auf der einen Seite erfordert die Rückmeldung auf Individual-, Lerngruppen- und Schulebene einen umfangreichen Itempool, um eine valide Abdeckung der Bildungsstandards zu gewährleisten. Es muss davon ausgegangen werden, dass Items, die flächendeckend einer gesamten Schulstufe in Österreich präsentiert werden, nicht weiter geheim zu halten sind und damit als „verbrannt“ gelten (d. h. in späteren Überprüfungen nicht wieder eingesetzt werden können). Auf der anderen Seite basiert das System-Monitoring allerdings auf einem longitudinalen Link über Items, die zwischen aufeinanderfolgenden Standardüberprüfungen nicht bekannt werden dürfen.

Um beide Anforderungen zu kombinieren, d. h. mit einem großen Itempool testen zu können und gleichzeitig Items für den longitudinalen Link geheim zu halten, wird bei der Testung eines Großteils der Schüler (ca. 93 % aller Klassen, die sog. „S93“) ein eingeschränkter Itempool eingesetzt. Die übrigen ca. 7 % der Klassen (sog. „S7“) werden mit dem gesamten BIST-Ü-Itempool unter stärker kontrollierten Bedingungen (z. B. externe Testleitung) getestet. Durch einen Link über Items, die sowohl in der S93 wie auch in der S7 eingesetzt werden, erfolgt die Rückmeldung (in der S93 und S7) auf Basis eines für die Bildungsstandards erschöpfenden Itempools, während gleichzeitig der Schutz eines Großteils der Items (in der S7)

gewährleistet werden kann. Von der S93 splittet sich weiterhin die S3 (ca. 3 % aller Klassen) ab, welche zur Qualitätsprüfung der BIST-Ü eingesetzt wird (z. B. Aufsicht der Durchführung durch externe Qualitätsprüfer). Die nach Abtrennung der S7 und S3 übrig bleibende Studie wird auch mit „S90“ bezeichnet.

Obwohl die BIST-Ü Vollerhebungen sind, werden dementsprechend Stichprobenziehungen durchgeführt, um Schüler der S7 bzw. der S3 zuzuweisen (vgl. Abschnitt 4.1 und 4.2). Weiterhin werden in BIST-Ü Teilkompetenzen (Sprechen in Englisch und Deutsch) nur an Stichproben getestet (vgl. Abschnitt 4.3), da eine Vollerhebung dieser Teilkompetenzen aus ökonomischen Gründen nicht durchführbar wäre. Alle Studien bzw. Stichproben sollen die Eigenschaften der Population gleich gut widerspiegeln.

2.2 Stichproben für Pilotstudien

Pilotstudien (auch: Pilotierungen oder Feldtests) dienen unter anderem dazu, Items auf deren Eignung für den Einsatz in der BIST-Ü oder der Informellen Kompetenz-Messung (IKM) zu testen. Die Anzahl der Schüler, an denen Items pilotiert werden, muss zwei gegensätzliche Forderungen erfüllen: Zum einen sollen Pilotierungen nur an „wenigen“ Schülern durchgeführt werden, um Schüler nicht zu häufig mit Tests zu konfrontieren und um finanzielle Ausgaben für Tests zu reduzieren. Zum anderen müssen „genügend“ Schüler pilotiert werden, um tragbare Aussagen über das Funktionieren der Items treffen zu können.

Nach Berücksichtigung der Überlegungen von Gonzalez & Foy (1997) zur Aussagekraft von Stichproben wird die Größe der Pilotstichproben von folgenden Faktoren beeinflusst:

- Der Menge an zu pilotierenden Items,
- der Anzahl der Items, die in einem Testheft vorkommen können,
- der Mindestanzahl der Schüler, die ein Item beantworten sollen,
- der Anzahl der Testhefte, die in einer Pilotierung gemeinsam eingesetzt werden können. Dies ergibt sich wiederum einerseits aus dem Testdesign, andererseits aus ökonomischen Begrenzungen.

Für die BIST-Ü wird in den folgenden Fächern bzw. Schulstufen pilotiert:

- Deutsch: 4. und 8. Schulstufe
- Mathematik: 4. und 8. Schulstufe
- Englisch: 8. Schulstufe

Für die IKM werden Pilotstudien in folgenden Fächern bzw. Schulstufen durchgeführt:

- Deutsch; Lesen: 3., 6. und 7. Schulstufe
- Deutsch; Sprachbetrachtung: 3. Schulstufe
- Deutsch; Sprachbewusstsein: 6. und 7. Schulstufe
- Deutsch; Verfassen von Texten (Prozessorientierte Teilkompetenzen): 3. Schulstufe
- Mathematik: 3., 6. und 7. Schulstufe
- Englisch; Listening, Reading, Writing: 6. und 7. Schulstufe
- Naturwissenschaften; Biologie, Physik und Chemie: 7. und 8. Schulstufe

Die Stichproben für Pilotstudien (sowohl für die BIST wie auch für die IKM) sollen die gleichen Qualitätskriterien wie die in Abschnitt 2.1 vorgestellten S3- und S7-Stichproben erfüllen. Da das Vorgehen von Stichprobenziehungen für Pilotstudien dem Vorgehen in BIST-Ü folgt, wird im Weiteren nicht gesondert auf diesen Fall eingegangen.

2.3 Zielpopulation und Datengrundlage

Im Rahmen der BIST-Ü werden in Österreich flächendeckend Schüler in der vierten (Volksschule, VS) oder in der achten Schulstufe (Sekundarstufe I, Sek I) in einem Fach (VS: Deutsch, Mathematik; Sek I: Deutsch, Mathematik, Englisch) getestet. Die Überprüfung betrifft Schüler an allen öffentlichen sowie privaten Schulen mit auf Dauer verliehenem Öffentlichkeitsrecht.

Von der BIST-Ü ausgenommen sind alle außerordentlichen Schüler sowie Schüler mit sonderpädagogischem Förderbedarf (SPF), die in dem jeweiligen Testfach nach dem Lehrplan der Allgemeinen Sonderschule (ASO) oder nach dem Lehrplan einer niedrigeren Schulstufe unterrichtet werden. Schüler mit Körper- oder Sinnesbehinderung nehmen dann nicht an der BIST-Ü teil, wenn sie selbst mit allenfalls im Unterricht zur Verfügung stehenden Unterrichts- oder Hilfsmitteln unter den standardisierten Testbedingungen die gestellten Aufgaben aller Voraussicht nach nicht lösen können. Detaillierte Informationen zu den Teilnahme Kriterien können der Verordnung über Bildungsstandards im Schulwesen (vgl. BGBl. II Nr. 1/2009) sowie deren Novelle (vgl. BGBl. II Nr. 282/2011) entnommen werden.

Im Gegensatz zur BIST-Ü werden die Testinstrumente der IKM allen Schulen zur Verfügung gestellt, die freiwillig an der IKM teilnehmen möchten. Die Zielpopulation bei IKM-Pilotierungen erweitert sich somit um SPF- und ASO-Schüler.

Als Datengrundlage für alle Stichprobenziehungen auf Schulebene dienen vom BMBF (vor 2014: BMUKK) bereitgestellte Daten (Schul- und Klassenliste). Nachgelagerte Ziehungen (Klassen- oder Schülerebene) basieren auf Informationen, die direkt von den Schulen erhoben werden, sofern dies möglich ist.

3 Stratifizierung

Vor der Stichprobenziehung wird die Schulenliste geordnet bzw. stratifiziert. Die Stratifizierung zeichnet sich dadurch aus, dass Schulen in Gruppen (auch Strata oder Schichten) aufgeteilt werden, die anhand der Ausprägungen bestimmter Variablen (der Stratifizierungsvariablen) ähnlich sind. In der BIST-Ü wird eine Stratifizierung aus zwei Gründen eingesetzt:

- (1) Repräsentativität der Stichproben: Es wird sichergestellt, dass Beobachtungen aus allen Strata in die Stichproben einfließen.
- (2) Präzision der Stichprobenschätzer: Ist die Variabilität der Zielvariable (hier: Schülerleistung) innerhalb der Strata kleiner als zwischen den Strata und/oder unterscheiden sich Mittelwerte der Zielvariablen zwischen den Strata deutlich, so kann die Präzision der Stichprobenschätzer in einer stratifizierten Stichprobe im Vergleich zur Präzision der Stichprobenschätzer in einer unstratifizierten Stichprobe gleichen Stichprobenumfangs erhöht werden.

Es werden zwei Arten der Stratifizierung durchgeführt: die explizite und die implizite Stratifizierung. Bei der *expliziten Stratifizierung* wird die Grundgesamtheit anhand von expliziten Stratifizierungsvariablen in mehrere voneinander unabhängige Sampling Frameworks (explizite Strata) aufgeteilt. Aus jedem expliziten Stratum wird dann eine Teilstichprobe gezogen und die einzelnen Stichprobenschätzer am Ende zu einem Gesamtschätzer kombiniert. Die *implizite Stratifizierung* besteht darin, die Liste der Schulen in jedem expliziten Stratum nach weiteren impliziten Stratifizierungsvariablen zu ordnen. Mit diesem Vorgehen wird in sehr einfacher Weise eine gleichverteilte Auswahl der Schulen über die implizite Stratifizierungsvariable hinweg sichergestellt. Unter der Voraussetzung, dass die impliziten Stratifizierungsvariablen mit der Zielvariable (hier: Schülerleistung) korreliert sind, kann die implizite Stratifizierung die Reliabilität der Stichprobenschätzer verbessern (OECD, 2012a).

In den Bildungsstandardüberprüfungen in der Volksschule wird die Variable Bundesland als explizite Stratifizierungsvariable eingesetzt. Innerhalb jedes expliziten Stratums werden die Variablen Schulbezirk und Schulgröße als implizite Stratifizierungsvariablen eingesetzt. Das heißt, die Beobachtungsliste in jedem expliziten Stratum wird zunächst nach Schulbezirk und innerhalb der Schulbezirke nach Schulgröße geordnet. In den Bildungsstandardüberprüfungen in der Sekundarstufe I werden die vier (sechs) Kombinationen aus den beiden Variablen Schulart (AHS/APS bzw. AHS/HS/NMS) und Stadt/Land als explizite Strata gewählt. Innerhalb jedes expliziten Stratums wird die Variable Schulgröße als implizite Stratifizierungsvariable eingesetzt. Bei der Wahl der Stratifizierungsvariablen wurden die Annahmen getroffen, dass die Unterschiede in den Schülerleistungen in den expliziten Strata kleiner als zwischen den expliziten Strata sind und dass die Stratifizierungsvariablen zeitlich stabil über die BIST-Erhebungszyklen bleiben.

4 Stichprobenziehung

In der Stichprobenziehung werden Schüler zufällig für die Stichproben S7 und S3 gezogen. Dabei soll sichergestellt sein, dass jeder Schüler mit gleicher Wahrscheinlichkeit für die S7 oder S3 gezogen werden kann und beide Studien die Eigenschaften der Zielpopulation in der jeweiligen Schulstufe repräsentieren. Um den Bearbeitungsaufwand bei der Administration des Tests zu reduzieren, werden allerdings nicht einzelne Schüler für S7 oder S3, sondern in einem zweistufigen Verfahren zunächst Schulen (Abschnitt 4.1) und aus diesen dann ganze Klassen (vgl. Abschnitt 4.2) gezogen. Die verbleibenden Klassen werden der S90 zugeteilt. Um die bei diesem zweistufigen Vorgehen entstehenden ungleichen Auswahlwahrscheinlichkeiten der Schüler auszugleichen, werden anschließend die Stichproben- und Varianzschätzer unter Einsatz von Schülergewichten berechnet (vgl. Abschnitt 5).

4.1 Schulenziehung

Bevor mit der Ziehung der Schulen begonnen wird, wird zunächst die Schulliste in separate Listen aufgeteilt. Es wird jeweils eine Liste für jedes explizite Stratum gebildet. Innerhalb dieser Listen werden die Schulen nach den impliziten Strata (VS: Schulbezirk und innerhalb der Schulbezirke Schulgröße, SekI: Schulgröße) geordnet. Innerhalb jedes expliziten Stratums werden zur Größe des Stratums in der Gesamtpopulation proportionale Stichproben gezogen (*probability proportional to size*). Nach der Schulenziehung werden die identifizierten Schulen und Reserveschulen in eine Gesamtliste übertragen, die wie die ursprüngliche Liste alphabetisch geordnet ist. Die Information über die Strata spielt für die Klassen- und Schülerziehung keine Rolle mehr.

4.1.1 Samplingintervalle

Um in jedem expliziten Stratum die Schulen gleichverteilt aus der nach den impliziten Strata geordneten Schulliste zu ziehen, werden zunächst Samplingintervalle berechnet.

Sei $N^{(s)}$ die Anzahl der Schulen in Stratum s , $s = 1, \dots, S$, und $J^{(s)}$ die Anzahl der Schulen in Stratum s . Die Größe des Samplingintervalls für die S7 in Stratum s ergibt sich als

$$I_{S7}^{(s)} = \frac{N^{(s)}}{0.07 \cdot J^{(s)}}$$

und mit analogen Bezeichnungen für die S3 in Stratum s als

$$I_{S3}^{(s)} = \frac{N^{(s)}}{0.03 \cdot J^{(s)}}.$$

Die Samplingintervalle $I_{S7}^{(s)}$ und $I_{S3}^{(s)}$, $s = 1, \dots, S$, werden nach den üblichen Rundungsregeln auf ganze Zahlen gerundet.

Im nächsten Schritt wird für jedes Stratum und jede Stichprobe eine ganzzahlige Zufallszahl $R_{S7}^{(s)}$ (bzw. $R_{S3}^{(s)}$) aus einer Gleichverteilung zwischen 0 und $I_{S7}^{(s)}$ (bzw. $I_{S3}^{(s)}$) gezogen. Basierend auf den Werten $R_{S7}^{(s)}$ und $R_{S3}^{(s)}$ werden Zahlen $Z_{S7;h}^{(s)}$ und $Z_{S3;h}^{(s)}$ zur Identifikation der Schulen bestimmt, wobei

$$Z_{h;S7}^{(s)} = R_{S7}^{(s)} + (h - 1) \cdot I_{S7}^{(s)}, \quad h = 1, \dots, J_{S7}^{*(s)}$$

und

$$Z_{h;S3}^{(s)} = R_{S3}^{(s)} + (h - 1) \cdot I_{S3}^{(s)}, \quad h = 1, \dots, J_{S3}^{*(s)}.$$

Hier bezeichnet $J_{S7}^{*(s)}$ bzw. $J_{S3}^{*(s)}$ die Anzahl der in Stratum s für die S7 bzw. S3 zu ziehenden Schulen. Nun wird ermittelt, in welchen Schulen der geordneten Schulliste von Stratum s die Werte $Z_{S7;h}^{(s)}$ und $Z_{S3;h}^{(s)}$ liegen.

4.1.2 Identifikation der gezogenen Schulen

Um die in Abschnitt 4.1.1 gezogenen Schulen zu identifizieren, wird ein kumulatives Größenmaß abgeleitet. Sei

$$N_j^{(s)}, \quad j = 1, \dots, J^{(s)},$$

die Anzahl der Schüler in der j -ten Schule der geordneten Liste von Stratum s und $J^{(s)}$ die Anzahl der Schulen in Stratum s . Es gelte für die Anzahl $N^{(s)}$ aller Schüler in Stratum s

$$N^{(s)} = \sum_{j=1}^{J^{(s)}} N_j^{(s)}.$$

Das kumulative Größenmaß der l -ten geordneten Schule in Stratum s mit $0 \leq l \leq J^{(s)}$ entspricht

$$C_l^{(s)} = \begin{cases} 0 & \text{für } l = 0 \\ \sum_{j=1}^l N_j^{(s)} & \text{für } l \geq 1. \end{cases}$$

Bei der Identifikation der ersten Schule für die S7-Stichprobe wird die Schule l gesucht, für die die erste S7- Identifikationszahl $Z_{1;S7}^{(s)}$ die kumulative Schülerzahl $C_{l-1}^{(s)}$ der Schule $(l-1)$ in der geordneten Liste übersteigt und gleichzeitig kleiner (oder gleich) der kumulativen Schülerzahl der Schule l ist. Allgemein werden alle Schulen l gesucht, für die

$$C_{l-1}^{(s)} < Z_{h;S7}^{(s)} \leq C_l^{(s)}, \quad h = 1, \dots, J^{*(s)} \quad (4.1)$$

gilt. Diese Schulen werden in die Stichprobe für die S7 aufgenommen. Das Verfahren der Identifikation der Schulen für die S3 läuft analog. Siehe zu anschaulichen Zwecken der Abläufe in 4.1.1 und 4.1.2 auch Abbildung 4.1.

4.1.3 Ziehungswahrscheinlichkeiten der Schulen

Bei der oben beschriebenen Vorgehensweise ist die Ziehungswahrscheinlichkeit einer Schule j in Stratum s gegeben durch

$$p_j^{(s)} = \frac{N_j^{(s)}}{N^{(s)}}. \quad (4.2)$$

Damit haben Schulen mit vielen Schülern automatisch höhere Ziehungswahrscheinlichkeiten. Am Ende der Stichprobenziehung soll erreicht werden, dass jeder Schüler mit gleicher Wahrscheinlichkeit gezogen wird.

4.1.4 Identifikation von Reserveschulen

Jeder Schule, die für eine Stichprobe identifiziert wurde, werden jeweils zwei Reserveschulen zugewiesen. Diese Reserveschulen rücken bei Nichteinsetzbarkeit (z. B. Schulschließung) einer ursprünglich identifizierten Schule in die Stichprobe nach. Als Reserveschulen werden jeweils die Schulen direkt vor bzw. direkt nach den identifizierten Schulen in der geordneten

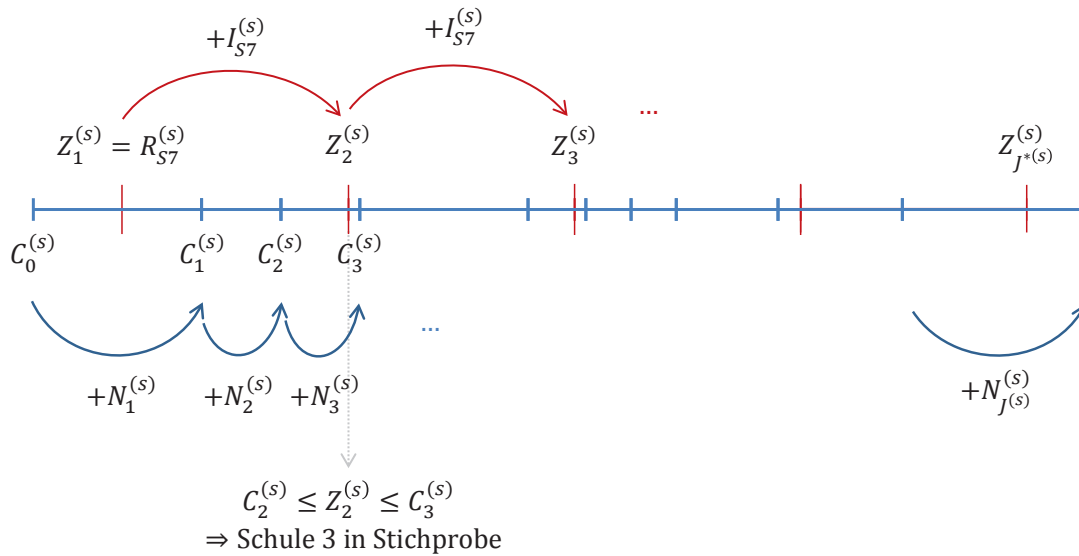


Abbildung 4.1: Geordnete Schulliste in Stratum s mit Samplingintervall $I_{S7}^{(s)}$, Identifikationszahlen $Z_{S7;1}^{(s)}, \dots, Z_{S7;J^{(s)}}^{(s)}$ der Schulen für die $S7$ in Stratum s und kumulativem Größenmaß $C_0^{(s)}, \dots, C_{J^{(s)}}^{(s)}$.

Mit dem Vergleich der Werte $Z_{S7;h}^{(s)}$, $C_{l-1}^{(s)}$ und $C_l^{(s)}$ wird entschieden, welche Schulen l in die Stichprobe aufgenommen werden. In dieser Illustration wird unter anderem die Schulen 3 für die Stichprobe $S7$ identifiziert, da $C_{l-1}^{(s)} < Z_{S7;2}^{(s)} \leq C_l^{(s)}$.

Liste jedes expliziten Stratums gewählt. Die Schule direkt nach der ursprünglich identifizierten wird als erste Reserveschule und die Schule direkt vor der ursprünglich identifizierten als zweite Reserveschule bezeichnet. Die implizite Stratifizierung gewährleistet, dass als Ersatzschule jeweils eine Schule mit möglichst ähnlichen Eigenschaften zu der ausfallenden Schule herangezogen wird.

4.1.5 Overlapcontrol S7 und S3

Sind alle Schulen für die S7- und S3-Stichproben identifiziert, wird abschließend kontrolliert, dass keine Schule mit nur einer Klasse sowohl für die S7 wie auch für die S3 nominiert wurde.

4.2 Klassenziehung

An den in Abschnitt 4.1 gezogenen Schulen werden nun Klassen für die Teilnahme an den Studien gezogen.

4.2.1 Auswahl der Klassen an gezogenen Schulen

Die Klassenziehung an den in Abschnitt 4.1 identifizierten Schulen teilt sich in zwei Fälle auf: Im ersten Fall wurde eine identifizierte Schule j , $j = 1, \dots, J^{*(s)}$, entweder nur für die Teilnahme an einer Studie *S7* oder *S3* identifiziert. In diesem Fall wird aus den $K_j^{(s)}$ -Klassen der j -ten Schule eine Klasse zufällig ausgewählt.

Im zweiten Fall wurde eine Schule (mit mindestens 2 Klassen) für die Teilnahme an der *S3* und an der *S7* gezogen. In diesem Fall werden aus den $K_j^{(s)}$ -Klassen der Schule zufällig zwei ausgewählt.

4.2.2 Ziehungswahrscheinlichkeiten der Klassen

Im ersten Fall, d. h. wenn eine Schule j zur Teilnahme an einer der beiden Studien identifiziert wurde, wird jede der $K_j^{(s)}$ -Klassen an dieser Schule mit gleicher Wahrscheinlichkeit

$$p_k^{(s)} = \frac{1}{K_j^{(s)}} \quad (4.3)$$

gezogen.

Im zweiten Fall, d. h. wenn eine Schule zur Teilnahme an beiden Studien identifiziert wurde, ist die Wahrscheinlichkeit, dass eine Klasse für eine der beiden Studien (entweder *S7* oder *S3*) gezogen wird $\frac{2}{K_j^{(s)}}$. Die Wahrscheinlichkeit, dass diese Klasse für nur die *S7* oder nur die *S3* gezogen wird beträgt weiterhin $p_k^{(s)} = \frac{1}{K_j^{(s)}}$.¹

4.3 Schülerziehung für Teilstudien

In bestimmten Teilstudien (Sprechen in Englisch und Deutsch) der BIST-Ü werden aufgrund des zeitlichen und administrativen Aufwands (v. a. Ausbildung und Einsatz von Assessoren in jeder Schule) nicht alle Schüler getestet. Für diese Teilstudien wird eine getrennte Stichprobenziehung vorgenommen, in der eine vordefinierte Anzahl von Schulen (ca. 120 bis 130) gezogen wird und daraus eine vordefinierte Anzahl von Schülern (VS: ca. 20, SekI: ca. 24).

¹In Studien, wie beispielsweise Pilotstudien, in denen per finanzieller Vorgabe nur eine bestimmte Anzahl von Schulen in die Stichprobe aufzunehmen ist, kann es zur Erhöhung des Stichprobenumfangs sinnvoll sein, größere Klassen mit höherer Wahrscheinlichkeit zu ziehen. In diesem Fall gewichtet man die Ziehungswahrscheinlichkeiten der Klassen in Fall 1 mit den Klassengrößen bzw. der Schüleranzahl. Für Fall 2 bildet man alle möglichen Zweierkombinationen der Klassen und zieht diese entsprechend gewichtet anhand deren Klassengrößen bzw. der Schüleranzahl. Es bleibt zu beachten, dass die Schülergewichte in 5.1 angepasst werden müssen (vgl. z. B. das Verfahren der OECD, 2012b).

4.3.1 Auswahl der Schüler für Teilstudien

In der Sekundarstufe I findet die Schulenziehung proportional in den expliziten vier Strata (AHS/Stadt, AHS/Land, APS/Stadt, APS/Land) statt. In der Volksschule weicht das Vorgehen leicht von dem in der Sekundarstufe I ab. Aufgrund des hohen Anteils (ca. 35 %) kleiner Schulen (mit weniger als 20 testbaren Schülern) und bei einer vorab definierten Anzahl von zu ziehenden Schulen (z. B. 120) besteht die Gefahr, dass die Anzahl der gezogenen Schüler deutlich geringer ausfällt als die angestrebte Stichprobengröße. Um die deutlichen Unterschiede in den Schulgrößen in der Volksschule zu kontrollieren, werden die Schulen zuerst in Kleinschulen (weniger als 20 testbare Schüler) und Großschulen (mindestens 20 testbare Schüler) kategorisiert. Dann wird jedes explizite Stratum (Bundesland) jeweils in zwei Strata (Klein- und Großschulen) aufgesplittet. Die Schulenziehung erfolgt proportional in den so neu gebildeten 18 Strata (Bundesland und Kategorie der Schulgröße).

Anschließend wird für jede gezogene Schule (sowohl in der SekI wie auch in der VS) über alle Klassen der jeweiligen Jahrgangsstufe hinweg eine wiederum fest definierte Anzahl von Schülern (VS: ca. 20, SekI: ca. 24) zufällig aus der Schülerliste gezogen. Die gezogenen Schüler werden zufällig in Paare von jeweils zwei Schülern eingeteilt. Eine Teilmenge der verbleibenden Schüler an den gezogenen Schulen wird als Reserveschüler eingesetzt, die bei Ausfall der gezogenen Schüler getestet wird. Besuchen weniger als die zuvor fest definierte Anzahl von Schülern die jeweilige Jahrgangsstufe einer gezogenen Schule, so wird zufällig die größtmögliche Menge an Paaren gebildet.

4.3.2 Ziehungswahrscheinlichkeiten der Schüler für Teilstudien

Sei n^* die zuvor fest definierte maximale Anzahl von Schülern (VS: ca. 20, SekI: ca. 24), die an einer gezogenen Schule an der Testung teilnimmt. Dann beträgt die Ziehungswahrscheinlichkeit eines Schülers i an Schule j in Stratum s für eine Teilstudie

$$p_{ij}^{*(s)} = \begin{cases} \frac{n^*}{N_j^{(s)}} \cdot \frac{N_j^{(s)}}{N^{(s)}} = \frac{n^*}{N^{(s)}} & \text{für } n^* \leq N_j^{(s)} \\ \frac{\lfloor N_j^{(s)}/2 \rfloor \cdot 2}{N_j^{(s)}} \cdot \frac{N_j^{(s)}}{N^{(s)}} = \frac{\lfloor N_j^{(s)}/2 \rfloor \cdot 2}{N^{(s)}} & \text{für } n^* > N_j^{(s)}. \end{cases}$$

5 Gewichtung

In diesem Abschnitt wird für jeden Schüler i ein Gewicht $w_{ij}^{(s)}$ bestimmt (vgl. Abschnitt 5.1). Dieses Gewicht dient primär dazu, die durch das zweistufige Ziehungsverfahren entstandenen verschiedenen Ziehungswahrscheinlichkeiten der Schüler auszugleichen. In einem weiteren Schritt werden Antwortausfälle (Non-Responses) wiederum durch den Einsatz von Gewichten adjustiert, da diese zu verzerrten Stichprobenschätzern führen können. Unter „Non-Response“ verstehen wir hier den Ausfall der gesamten Antworten eines Schülers zum Test, da dieser beispielsweise krankheitsbedingt nicht teilnehmen konnte. Explizit nicht gemeint sind fehlende Schülerantworten auf bestimmte Items (Item Non-Response). Die berechneten Gewichte werden schließlich so normiert, dass ihre Summe dem Stichprobenumfang im jeweiligen Stratum gleicht (Peaker, 1953).

5.1 Schüलगewichte

Die Schüलगewichte werden, basierend auf den Ziehungswahrscheinlichkeiten der Schüler, separat pro Stratum s und pro Studie, d. h. S7, S3, oder S90, berechnet.

Die Ziehungswahrscheinlichkeit eines Schülers i in Stratum s an Schule j ergibt sich aus der Ziehungswahrscheinlichkeit $p_j^{(s)}$ der Schule j in Stratum s , die Schüler i besucht (Gleichung (4.2)), und der Ziehungswahrscheinlichkeit $p_k^{(s)}$ der Klasse k , in die Schüler i in Schule j geht (Gleichung (4.3)). Die beschriebene Ziehungswahrscheinlichkeit eines Schülers beträgt

$$p_{ij}^{(s)} = p_j^{(s)} \cdot p_k^{(s)}. \quad (5.1)$$

Aus der Ziehungswahrscheinlichkeit in Gleichung (5.1) ergibt sich das Schüलगewicht

$$w_{ij}^{(s)} = \frac{1}{p_{ij}^{(s)}}. \quad (5.2)$$

Die Schüलगewichte sind für jeden Schüler i einer Schule j identisch. Weiterhin sind (bei gleicher Klassenanzahl) die Gewichte für Schüler aus großen Schulen (mit größerer Ziehungswahrscheinlichkeit) kleiner als für Schüler aus kleinen Schulen (mit kleinerer Ziehungswahrscheinlichkeit). Bei gewichteter Bestimmung der Stichprobenschätzer spiegelt jeder Schüler in der Stichprobe eine adäquate Anzahl von Schülern in der Zielpopulation wider, wodurch Repräsentativität gewährleistet ist und die Stichprobenschätzer unverzerrt sind.

5.2 Non-Response Adjustment

Um Verzerrungen in den Stichprobenschätzern auszugleichen, werden pro Schule die in Gleichung (5.1) bestimmten Schülergewichte mit einer schulspezifischen Non-Response-Rate adjustiert.

Sei $X_{ij}^{(s)}$ die dichotome Variable, die anzeigt, ob Schüler i in Schule j in Stratum s am Test teilgenommen hat ($X_{ij}^{(s)} = 1$) oder (beispielsweise krankheitsbedingt) nicht teilgenommen hat ($X_{ij}^{(s)} = 0$). Die Response-Rate für Schule j in Stratum s berechnet sich dann wie folgt:

$$f_j^{(s)} = \frac{\sum_{i=1}^{N_j^{(s)}} X_{ij}^{(s)}}{N_j^{(s)}}.$$

Die Adjustierung der Schülergewichte aus Gleichung (5.2) erfolgt per

$$w_{ij}^{*(s)} = \frac{w_{ij}^{(s)}}{f_j^{(s)}}.$$

Schließlich werden die Gewichte per

$$w_{ij\text{norm}}^{*(s)} = \frac{N^{(s)}}{\sum_{j=1}^{J^{(s)}} \sum_{i=1}^{N_j^{(s)}} w_{ij}^{*(s)} X_{ij}^{(s)}}$$

normiert.

Literatur

- Cochran, W. G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.
- Gonzalez, E., & Foy, P. (1997). Estimation of sampling variability, design effects, and effective sample sizes. In M. O. Martin, & L. Kelly (Eds.), *TIMSS Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.
- Joncas, M. (2007). PIRLS 2006 sample design. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2006 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.
- Joncas, M. (2008). TIMSS 2007 sample design. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.
- OECD (2012a). Sampling. In *PISA 2009 Technical Report*. OECD Publishing.
- OECD (2012b). Survey weighting and the calculation. In *PISA 2009 Technical Report*. OECD Publishing.
- Peaker, G. (1953). A sampling design used by the ministry of education. *Journal of the Royal Statistical Society*, 116, 140–165.

