Universität Zürich
Institut für Bildungsevaluation

Ehemaligenbefragung 2006

Expertise zuhanden der Bildungsdirektion des Kantons Zürich, Bildungsplanung

Urs Moser & Domenico Angelone
Zürich, 11. März 2009

Inhalt

1 Ausgangslage

Das Institut für Bildungsevaluation wurde von der Bildungsdirektion des Kantons Zürich beauftragt, das methodische Vorgehen bei der Datenanalyse und Ergebnisdarstellung der Befragung ehemaliger Zürcher Mittelschülerinnen und Mittelschüler zu überprüfen und zu beurteilen.

Zur Beurteilung der angewandten Methoden wurden einerseits der Bericht «Befragung ehemaliger Zürcher Mittelschülerinnen und Mittelschüler» des Statistischen Amts des Kantons Zürich sowie verschiedene Dokumente der Bildungsdirektion genutzt. Andererseits wurden statistische Analysen mit den Daten aus dem Jahr 2006 durchgeführt. Dokumente und Daten wurden dem Institut für Bildungsevaluation für die vorliegende Expertise zur Verfügung gestellt.

Der Auftrag war zeitlich beschränkt. Es konnten deshalb folgende drei Aspekte des methodischen Vorgehens bei der Analyse und Ergebnisdarstellung – zum Teil etwas differenzierter, zum Teil etwas weniger differenziert – beurteilt werden:

1. Rücklaufquote
2. Vergleich der Schulen
3. Berechnung der Schulmittelwerte

Die Rücklaufquote ist ein Indikator für die *Repräsentativität* der Ergebnisse. Der Vergleich der Schulen entspricht einem Ranking, das entweder auf *signifikante oder zufällige* Unterschiede zwischen den Schulmittelwerten hinweist. Die Berechnung der Schulmittelwerte wirkt sich auf die *Fairness des Vergleichs* aus.

Die folgenden Ausführungen sind weder vollständig noch beruhen sie auf einer spezifischen Literaturrecherche. Das methodische Vorgehen bei der Datenanalyse und Ergebnisdarstellung der Befragung ehemaliger Zürcher Mittelschülerinnen und Mittelschüler wird vielmehr aufgrund der Ausführungen in Lehrbüchern über sozialwissenschaftliche Methoden und Schuleffektivitätsforschung beurteilt.

2 Rücklaufquote

Unter der Rücklaufquote beziehungsweise der Ausschöpfung wird das Verhältnis der verteilten Fragebögen zu den ausgewerteten Fragebögen einer Schule verstanden. Richtlinien über Rücklaufquoten sind uns keine bekannt. Es gibt Umfragebereiche, da wird eine Rücklaufquote von 15 Prozent bereits als hoch eingestuft und andere, da wird ein Rücklauf von 65 Prozent als inakzeptabel beurteilt (PISA-Studie).

Die Rücklaufquote bei der Befragung ehemaliger Zürcher Mittelschülerinnen und Mittelschüler im Jahr 2006 variiert zwischen 48 und 70 Prozent pro Schule; die Anzahl auswertbarer Fragebögen variiert zwischen 18 und 122 Stück pro Schule.

Die relativ stark variierenden und zum Teil eher geringen Rücklaufquoten bilden ein erstes Problem für die Darstellung der Ergebnisse nach Schulen. Eine geringe Rücklaufquote kann ein Hinweis dafür sein, dass die Fragebögen als eher langweilig interpretiert werden und für die Befragten nicht von hoher Bedeutung sind. Für die Rücklaufquote spielen auch andere Faktoren eine Rolle, beispielsweise der Absender. Rücklaufquoten sind erfahrungsgemäss am höchsten, wenn die Umfrage regional begrenzt ist und von einer Universität im Einzugsgebiet durchgeführt wird. Rücklaufquoten lassen sich beispielsweise durch schriftliches oder telefonisches Nachfragen, aber auch durch monetäre Anreize steigern (incentives)[1].

Die grosse Variation der Rücklaufquote bei der Ehemaligenbefragung führt zur Vermutung, dass der Rücklauf systematisch verzerrt sein könnte. Es ist daher angebracht, den Rücklauf – zumindest bei sehr grossen Schulen – zu analysieren und eine Repräsentativitätskontrolle durchzuführen. Für eine faire Diskussion sollte unseres Erachtens ausgeschlossen werden können, dass der Rücklauf durch ein bestimmtes Merkmal der antwortenden Personen, das mit den Erfolgskriterien – beispielsweise Zufriedenheit – korreliert, verzerrt wird und dadurch ein aussagekräftiger Vergleich zwischen den Schulen nicht zulässig wäre.

Für eine Repräsentativitätskontrolle werden die statistischen Daten (beispielsweise Maturitätsprofil oder Geschlecht) der antwortenden Personen mit den statistischen Daten der Zielpopulation verglichen. Stellt sich heraus, dass in der Stichprobe der antwortenden Personen einzelne Merkmale über- oder unterrepräsentiert sind, dann muss überprüft werden, ob die Beantwortung der Fragen von diesen Merkmalen – in unserem Beispiel vom Maturitätsprofil oder vom Geschlecht – abhängt. Falls Verzerrungen vorhanden sind, könnten Gewichtungsmethoden angewendet werden. Bei relativ kleinen Gruppen lässt sich eine gewichtete Hochrechnung allerdings statistisch nicht rechtfertigen. Trotzdem wäre es sinnvoll, in Zukunft aufgrund einer Repräsentativitätskontrolle gewisse Verzerrungen entweder nachzuweisen oder auszuschliessen.

3 Vergleich der Schulen

3.1 Statistische Signifikanz

Das zweite Problem der Ergebnisdarstellung nach Schulen liegt darin, dass die Schätzungen pro Schule fehlerbehaftet sind. Die Mittelwerte liegen innerhalb eines bestimmten Vertrauensintervalls. Sie können vor allem bei geringen Fallzahlen ($n<30$) beim nächsten Mal deutlich tiefer oder deutlich höher ausfallen, auch wenn sich alle Schülerinnen und Schüler einer Schule beteiligen. Damit die fehlerbehafteten Schätzungen beurteilt werden können, wird jeweils pro Schule ein Stichprobenfehler berechnet. Zudem wird überprüft, ob sich die Mittelwerte zweier Schulen statistisch signifikant unterscheiden. Dazu wird Formel (1) für unabhängige Stichproben benutzt:

---

[1] Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation.* Berlin: Springer.

(1) $$z_{ij} = \frac{|\bar{x}i - \bar{x}j|}{\sqrt{SE^2 i + SE^2 j}}$$

wobei $\bar{x}i$ und $\bar{x}j$ die Mittelwerte der Schulen i und j sowie $SE^2 i$ und $SE^2 j$ die entsprechenden Stichprobenfehler der beiden Schulen sind[2].

Ist zij > 1.96, dann ist die Differenz zwischen zwei Schulen bei einer Irrtumswahrscheinlichkeit von α= 0.05 statistisch signifikant, sofern die Hypothese ungerichtet ist (zweiseitiger Test). Ist zij > 2.58, dann ist die Differenz zwischen zwei Schulen bei einer Irrtumswahrscheinlichkeit von α= 0.01 statistisch signifikant, sofern die Hypothese ungerichtet ist. Wird die Hypothese gerichtet formuliert (einseitiger Test), dann beträgt der kritische Wert zij = 1.658 (positiv gerichtete Hypothese) oder zij = –1.658 (negativ gerichtete Hypothese). Anstelle der Normalverteilung sollte bei kleinen Fallzahlen (n < 30) beziehungsweise wenigen Freiheitsgraden die t- Verteilungsfunktion genutzt werden. Ab 30 Freiheitsgraden kann die t-Verteilungsfunktion allerdings durch die Normalverteilung approximiert werden.

3.2 Vergleich der Einzelschule mit allen Schulen

Als Problem wird von der Bildungsdirektion erwähnt, dass die Einzelschulen jeweils mit allen Schulen verglichen werden. Aus statistischer Sicht ist es in der Tat nicht ganz korrekt, die Einzelschulen jeweils mit allen Schulen zu vergleichen, ohne die Kovarianz zwischen den beiden Stichproben zu berücksichtigen. Jede Einzelschule ist in der Stichprobe aller Schulen enthalten, weshalb eine Einzelschule nicht von allen Schulen unabhängig ist. Bei der Berechnung der Differenzen zwischen einer Einzelschule und allen Schulen muss deshalb die Kovarianz berücksichtigt werden. Dazu wird Formel (2) für abhängige Stichproben benutzt:

(2) $$z_{ij} = \frac{|\bar{x}i - \bar{x}j|}{\sqrt{SE^2 i + SE^2 j - 2\,\mathrm{cov}(\bar{x}i, \bar{x}j)}}$$

wobei $\bar{x}i$ und $\bar{x}j$ die Mittelwerte der Schulen i und j sowie $SE^2 i$ und $SE^2 j$ die entsprechenden Stichprobenfehler der beiden Schulen sind. Wenn zij > 1.96 ist, dann ist die Differenz zwischen zwei Schulen statistisch signifikant bei einer Irrtumswahrscheinlichkeit von α = 0.05, wenn zij > 2.58 ist, dann ist die Differenz zwischen zwei Schulen statistisch signifikant bei einer Irrtumswahrscheinlichkeit von α = 0.01. Wenn die beiden Schulen voneinander unabhängig sind, dann beträgt die Kovarianz «0».

Weil die beiden Stichproben (Einzelschule, alle Schulen) nicht vollständig voneinander unabhängig sind und die Ergebnisse vermutlich leicht positiv korrelieren, fällt die Kovarianz positiv und die Varianz der Differenz $\bar{x}i - \bar{x}j$ dementsprechend etwas geringer aus. Das heisst, dass zij unter Einbezug der Kovarianz grösser wird als ohne Einbezug

---

[2] Henkel, R.E. (1976). *Tests of Significance*. Newbury Park: Sage University Paper.

der Kovarianz. Ohne Berücksichtigung der Kovarianz wird die Differenz bei einem Vergleich einer Einzelschule mit allen Schulen unterschätzt und das Ergebnis eher als nicht signifikant erklärt, obwohl es eigentlich signifikant wäre (Fehler 2. Art, Beta-Fehler).

3.3 Multipler Vergleich

Bei einem Schulranking werden meistens mehr als zwei Schulen miteinander verglichen. Dabei steigt die Wahrscheinlichkeit für das fälschliche Entdecken eines signifikanten Effekts (Fehler 1. Art, Alpha-Fehler) massiv an[3]. Aus diesem Grund wird häufig eine Bonferroni-Korrektur angewendet. Damit kann die Signifikanz bei Mehrfachvergleichen korrekt bestimmt werden. Die Modifikation über die Bonferroni-Korrektur hält die Irrtumswahrscheinlichkeit für das fälschliche Entdecken eines signifikanten Effekts bei $\alpha = 0.05$ konstant. Dazu wird die Irrtumswahrscheinlichkeit durch die Anzahl der Vergleiche dividiert. Bei 24 Vergleichen in einem Schulranking wird die Irrtumswahrscheinlichkeit α entsprechend dem Quotienten (3) angepasst.

(3) $$\alpha_{n=24} \frac{0.05}{n-1} = \frac{0.05}{23} = 0.0022 .$$

Weil bei einem Schulranking alle Schulen gegeneinander verglichen werden, also ungerichtete Hypothesen geprüft werden, ist eine Bonferroni-Korrektur bei Signifikanztests dringend zu empfehlen. Ansonsten besteht die Gefahr, Unterschiede nachzuweisen, obwohl gar keine vorhanden sind (Fehler 1. Art, Alpha-Fehler).

3.4 Konfidenzintervalle für Schulmittelwerte

Für den Schulvergleich aufgrund der Befragung ehemaliger Zürcher Mittelschülerinnen und Mittelschüler wird die Methode mit den Vertrauensintervallen gewählt. Konkret heisst dies, dass sich die Ergebnisse zwischen zwei Schulen statistisch signifikant unterscheiden, wenn sich die Vertrauensintervalle nicht überschneiden. Die Berechnung der Konfidenzintervalle ist technisch gesehen eher ein Schätzverfahren als ein Signifikanztest, aber es besteht eine Beziehung zwischen den beiden Methoden. Anstelle der Schätzung einer Statistik wird ein Intervall festgelegt, in dem mit hoher Wahrscheinlichkeit der wahre Parameter liegt. Die Formel zur Bestimmung der Konfidenzintervalle (5) erhält man, wenn man die z-Wert-Formel (4) für den Signifikanztest für einen Stichprobenmittelwert auflöst[4].

(4) $$z_{ij} = \frac{|\bar{x} - \mu|}{\sigma}$$

(5) $$\mu_{U/L} = \bar{x} \pm z\sigma_{\bar{x}}$$

---

[3] Toothaker, L.E. (1993). *Multiple Comparison Procedures.* Newbury Park: Sage University Paper.

[4] Henkel, R.E. (1976). *Tests of Significance*. Newbury Park: Sage University Paper. [Seite 73]

Für jede Schule wird das Vertrauensintervall berechnet beziehungsweise die obere und untere Grenze (lower and upper confidence limits), indem zum Mittelwert der z-Wert (1.96 bei einer Irrtumswahrscheinlichkeit von $\alpha = 0.05$ und 2.58 bei einer Irrtumswahrscheinlichkeit von $\alpha = 0.01$) multipliziert mit dem Stichprobenfehler addiert wird. Weil die Formel des Signifikanztests für Mittelwertsdifferenzen für die Berechnung der Konfidenzintervalle genutzt wird, ist es offensichtlich, dass zwischen Signifikanztest und Konfidenzintervall eine Beziehung besteht. Ein Konfidenzintervall kann für die Testung jeglicher Nullhypothesen gegen nicht gerichtete Alternativhypothesen genutzt werden. Überschneiden sich die Konfidenzintervalle zweier Schulen, dann unterscheiden sich die Mittelwerte nicht signifikant; überschneiden sich die Konfidenzintervalle nicht, dann unterscheiden sich die Mittelwerte statistisch signifikant.

Bei der Darstellung der Mittelwerte mit den Konfidenzintervallen handelt es sich um eine konservative Signifikanzprüfung. Wenn sich zwei Konfidenzintervalle nicht überschneiden, unterscheiden sich zwei Schulen in der Regel statistisch signifikant. Diese Aussage ist allerdings mit Vorsicht zu geniessen. Zwei Schulen können sich auch dann statistisch signifikant unterscheiden, wenn sich die Konfidenzintervalle überschneiden. Welches Verfahren angewendet werden muss, ist vor allem durch die Fragestellung bestimmt (gerichtet oder ungerichtete Hypothese, Vergleich von zwei oder von allen Schulen).

Die Frage, *was wie* miteinander verglichen werden soll, lässt sich je nach Interesse unterschiedlich beantworten. Bei einem Schulranking anhand von Selbsteinschätzungen sind aufgrund des kargen Forschungsstandes ungerichtete Hypothesen sowie multiple Vergleiche sinnvoll. Das heisst, es kann nicht gesagt werden, welche Schulen aufgrund verschiedener Merkmale (Maturitätsprofil, Schwerpunktprogramme, Grösse, Stadt/Land, Kurz- oder Langgymnasium, Frauenanteil etc.) eigentlich die höchsten Mittelwerte erreichen müssten. Das fehlende Wissen über die Ursachen der Ergebnisse von retrospektiven Schulbeurteilungen ist auch ein Grund dafür, weshalb für die Schulen zurzeit keine «Value Added Performance» berechnet werden kann. Die Darstellung der Schulmittelwerte mit den zugehörigen Konfidenzintervallen im Sinne eines konservativen Signifikanztests ist deshalb angemessen. Ob aufgrund der Schulmittelwerte hingegen auf die Qualität einer Schule geschlossen werden kann, ist zu bezweifeln.

3.5 Konfidenzintervalle für Populationsanteile

Die Nutzung der Konfidenzintervalle ist auch für die Berechnung von signifikanten Unterschieden zwischen relativen Häufigkeiten beziehungsweise Prozentwerten sinnvoll. Die Probleme bei Schulvergleichen stellen sich unabhängig davon, ob Intervallschätzungen (beispielsweise Mittelwerte [μ]) oder Punktschätzungen (beispielsweise relative Häufigkeiten [π]) vorgenommen werden. Das Vorgehen von Sachs (2002) ist für kleine Gruppen geeignet und nutzt die F-Verteilung[5]. Benutzt wird die F-Verteilung, wenn die Stichproben relativ klein sind, wobei die Grenze zwischen klein und gross durch folgende Formel (6) bestimmt wird.

---

[5] Sachs, L. (2002). *Angewandte Statistik. Anwendung statistischer Methoden.* Berling: Springer.

(6) $$n < \frac{9}{p(1-p)}$$

Für einen Teil der Schulen, die von der Ehemaligenbefragung betroffen sind, ist die F-Verteilung unabhängig von p angemessen. In der Regel liegt die Grenze für die Nutzung der F-Verteilung jeweils zwischen 36 und 100 Individuen pro Stichprobe.

(7) $$\pi_{U/L} = p \pm z \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

Für grössere Sichproben wird die Formel (7) auf der Basis der Binomialverteilung beziehungsweise der Normalverteilung zur Berechnung der Konfidenzintervalle für Populationsanteile (π) angewendet[6].

3.6 Praktische Bedeutsamkeit

Unseres Erachtens ist es wesentlich informativer, anstelle der statistischen Signifikanz mit der praktischen Bedeutsamkeit zu operieren. Die statistische Signifikanz lässt sich sehr stark beeinflussen. Die praktische Bedeutsamkeit entspricht hingegen in den meisten Fällen einer Standardisierung von Differenzen. Sie ist deshalb für den Vergleich von Schulen ein informativer und wenig kritisierbarer Wert.

Wenn die standardisierte Differenz zwischen zwei Schulmittelwerten grösser als 0.5 ist (Effektgrösse ES > 0.5), dann ist der Unterschied zwischen zwei Schulen bedeutsam. Ist die Differenz sogar grösser als 0.8 (Effektgrösse ES > 0.8), dann ist der Unterschied zwischen den Schulen als gross zu beurteilen. Allerdings ist die zuverlässige Interpretation der Effektgrössen auch daran gebunden, dass sämtliche Schulmittelwerte zuverlässig gemessen werden und repräsentativ sind.

Die Effektgrössen lassen sich auch für relative Häufigkeiten angeben. Die Differenzen sind aber in Abhängigkeit von den Verteilungen zu beurteilen und benötigen zudem eine kleine Transformation, die bei Cohen (1988, S. 179ff.) nachgelesen werden kann[7].

Die Probleme einer tiefen und verzerrten Ausschöpfung sowie die geringe Anzahl auswertbarer Fragebögen pro Schule werden durch die Darstellung von Effektgrössen allerdings nicht gelöst. Von daher ist eine Kombination der Darstellungen mit Konfidenzintervallen und Effektgrössen wünschenswert.

---

[6] Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation.* Berlin: Springer. Seite 393.
«Bei grösseren Stichproben kann man von der Tatsache Gebrauch machen, dass die Binomialverteilung für n·p·(1-p)>9 hinreichend gut durch die Normalverteilung approximiert werden kann, was die Bestimmung von Konfidenzintervallen erheblich erleichert.»

[7] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Earlbaum Associates.

4 Berechnung von Schulmittelwerten

4.1 Schätzer

Schulmittelwerte werden häufig als einfaches arithmetisches Mittel dargestellt. Für die Berechnung des arithmetischen Mittels ist die Grösse einer Schule nicht relevant. Ob ein Mittelwert aufgrund von 20 oder aufgrund von 200 Schülerinnen und Schülern zustande kommt, bleibt unberücksichtigt.

Eine sinnvolle Alternative zum arithmetischen Mittel sind «empirical Bayes estimates» (EB), die sich mit Mehrebenenanalysen berechnen lassen. Dadurch können unterschiedliche Fallzahlen berücksichtigt werden, denn es ist offensichtlich, dass die Schätzung von Schulmittelwerten aufgrund von 20 Schülerinnen und Schüler weniger zuverlässig (reliabel) ausfällt als aufgrund von 200 Schülerinnen und Schülern.

Bei einer Berechnung des Mittelwerts aufgrund einer einfachen OLS-Regression entspricht das Intercept einer Schule dem Mittelwert der Schule. Der Mittelwert wird aus dem Gesamtmittelwert (grand mean) und dem Zufallseffekt für diese Schule berechnet, wie Formel (8) zeigt. Jede Schule wird als Dummy-Variable in die Gleichung eingeführt, weshalb die Schulmittelwerte ebenso gut mit dem arithmetischen Mittel geschätzt werden können.

(8) $$\beta_{0j} = \beta_0 + u_{0j}$$

Wird der Schulmittelwert mit dem empirical Bayes-estimates geschätzt, dann wird zusätzlich die Reliabilität der Schätzung berücksichtigt. Der Mittelwert der Schule ergibt sich aus dem geschätzten ordinary least square-Schätzer und der Reliabilität der Schätzung, wie das Formel (9) aufzeigt.

(9) $$\beta_{0j}^{EB} = \lambda \beta_{0j}^{OLS}$$

Die Reliabilität λ ergibt sich aus der Intraklassenkorrelation (Rho = Anteil der Varianz zwischen den Schulen [π] an der Gesamtvarianz [$\pi+\sigma^2$]) und der Anzahl Schülerinnen und Schüler der Schule, wie in Formel (10) dargestellt. Die Reliabilität steigt, je mehr Schülerinnen und Schüler einer Schule beteiligt sind und je grösser die Varianz zwischen den Schulen im interessierenden Merkmal ist.

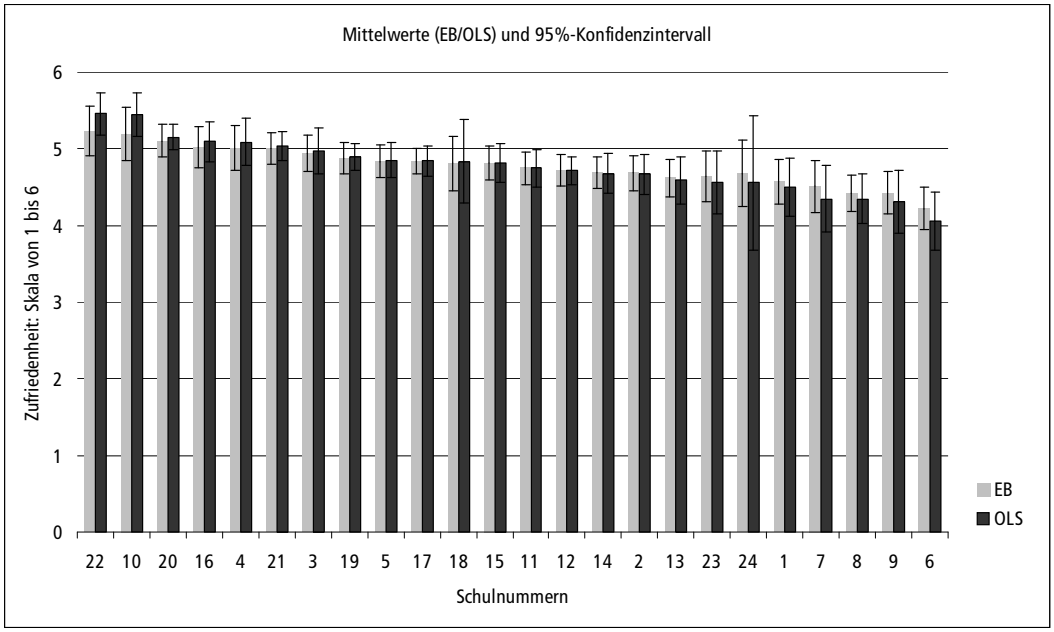
(10) $$\lambda = \pi_{00} / (\pi_{00} + \sigma^2 / n_j)$$

Angenommen, die Reliabilität in einer Schule ist sehr hoch, dann entspricht der Mittelwert nahezu dem durch die OLS Regression geschätzten Mittelwert. Ist die Reliabilität hingegen sehr klein, dann wird der Schulmittelwert nicht nur durch den ordinary least squares-Schätzer, sondern auch noch durch den Gesamtmittelwert bestimmt. Das

heisst, dass sich die Schätzung der Schulmittelwerte mit abnehmender Reliabilität dem Gesamtmittelwert annähert[8].

Dieses Phänomen ist in der Abbildung 1 dargestellt. Die Mittelwerte (arithmetisches Mittel beziehungsweise OLS-Schätzer und EB-Schätzer) der 24 Schulen sind als Säulen mit dem 95%-Konfidenzintervall dargestellt. Durch die Nutzung der empirical Bayes estimates werden die Differenzen zwischen den Schulen mit den höchsten und tiefsten Werten deutlich geringer. Zum einen kommt es zu einer Annäherung der Mittelwerte aufgrund der zum Teil geringen Fallzahlen, zum andern ist auch die Intraklassenkorrelation (Rho = 0.07) eher gering. Die Schülerinnen und Schüler innerhalb einer Schule sind sich zwar ähnlicher als die Schülerinnen und Schüler aller Schulen. Mit einer Intraklassen-Korrelation von 7 Prozent ist der Anteil der Varianz zwischen den Schulen an der Gesamtvarianz jedoch nicht besonders gross.

Abbildung 1: EB-Schätzer versus OLS-Schätzer



Der empirical Bayes estimates wird auch als «shrinkage estimates» bezeichnet, der den OLS-Schätzer proportional zur Reliabilität schrumpfen lässt[9]. Je zuverlässiger der Mittelwert geschätzt wird, desto geringer ist die «Schrumpfung zum Mittelwert». Mit abnehmender Reliabilität verlässt man sich aber zunehmend auf den Gesamtmittelwert und weniger auf den Schulmittelwert.

---

[8] Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models for social and behavioural research: Applications and data analysis methods.* Newbury Park: Sage Publications. [Seite 124f.]

[9] Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models for social and behavioural research: Applications and data analysis methods.* Newbury Park: Sage Publications. [Seite 78ff.]

Die Nutzung der empirical Bayes estimation hat zur Folge, dass die Extremmittelwerte in der Randverteilung näher dem Gesamtmittelwert liegen und die Varianz der Mittelwerte etwas geringer ist. Auch die Stichprobenfehler sind etwas geringer. Das bedeutet, dass sowohl die Varianz der Schulmittelwerte als auch die Konfidenzintervalle bei Anwendung der empirical Bayes estimation etwas geringer ausfallen als bei der einzelnen Berechnung des Stichprobenfehlers einer Schule.

4.2 Value Added Performance

Werden bei der Schätzung von Mittelwerten zugleich Kontextinformationen (beispielsweise Maturitätsprofil, soziale Herkunft, Geschlecht) berücksichtigt, dann spricht man auch von Value Added Performance. Dazu werden Zusammenhänge zwischen Kontextinformationen und den Erfolgskriterien (beispielsweise Zufriedenheit) berechnet.

Zur Berechnung dieser Zusammenhänge, beispielsweise zwischen dem Maturitätsprofil und der Zufriedenheit, wurde die Varianzanalyse korrekt durchgeführt. Die Gruppenunterschiede werden aber im Bericht nicht expliziert (Ergebnisse der Mittelwertsvergleiche)[10]. So wird beispielsweise geschrieben, dass Personen mit musischem Profil zufriedener seien. Dies stimmt in dieser allgemeinen Aussage allerdings nicht. Personen mit musischem Profil sind nicht einfach zufriedener als alle anderen, sondern nur zufriedener als jene der Profile «Wirtschaft» und «Neusprachlich».

Insgesamt wird im Bericht die Darstellung der Ergebnisse über die Zusammenhänge etwas vermisst. Vertiefende Analysen müssten für die künftigen Schulvergleiche von Interesse sein. Schulmittelwerte werden für öffentliche Rankings mit Vorteil statistisch kontrolliert als Value Added Performance ausgewiesen. Während es eine Selbstverständlichkeit ist, dass bei der Berechung von Schulmittelwerten anhand der Schulleistungen immer auch die Zusammensetzung der Schulen berücksichtigt wird, scheint dies bei Selbsteinschätzungen kein Thema zu sein. Es könnte aber durchaus sein, dass Frauen die Zufriedenheit höher einschätzen als Männer, dass sich die Einschätzungen nach Maturitätsprofil, gewählter Hochschule, Dauer des Gymnasiums u.a. systematisch unterscheiden.

5 Fazit

Die Schulvergleiche im Rahmen der Befragung ehemaliger Zürcher Mittelschülerinnen und Mittelschüler werden methodisch korrekt durchgeführt. Aus einer «statistischen» Perspektive lässt sich kaum etwas bemängeln. Wird allerdings die aktuelle Diskussion über angemessene Methoden der Schuleffektivitätsforschung bei der Beurteilung des methodischen Vorgehens der Ehemaligenbefragung berücksichtigt, dann fällt das Urteil weniger positiv aus. Es ist durchaus Optimierungsbedarf vorhanden.

---

[10] Dazu wäre ein Gruppenvergleich mit der Korrektur für multiple Vergleiche (Scheffé-Test) notwendig.

Die *Rücklaufquote* ist sowohl *ein Mass für die Repräsentativität* der Ergebnisse insgesamt als auch für die Ergebnisse der einzelnen Schulen. Die Rücklaufquoten der Schulen sind unterschiedlich hoch. Zum Teil sind sie kleiner als 50 Prozent, weshalb unseres Erachtens Repräsentativitätskontrollen angebracht sind.

Der *Vergleich der Schulen entspricht einem Ranking*. Der Vergleich erfolgt nicht nur anhand der Schulmittelwerte, sondern auch anhand der Konfidenzintervalle. Diese Methode ermöglicht es den Leserinnen und Lesern auf einfache Art und Weise zu beurteilen, ob sich zwei Schulen statistisch signifikant unterscheiden. Dabei handelt es sich um eine eher konservative Methode, was für einen multiplen Vergleich angemessen ist. Zu beachten ist allerdings, dass die Frage nach der Signifikanz nicht nur eine statistische Frage ist, sondern vor allem auch eine theoretische.

Für die Leserschaft könnte es zudem von Interesse sein, ein Mass für die praktische Bedeutsamkeit der Differenzen zu haben. Dafür spricht, dass aus den Grafiken und den Mittelwertsvergleichen bei der verwendeten Skala (1 bis 6) keine zuverlässige Beurteilung der Differenzen möglich ist. Allerdings gilt es zu beachten, dass die Angabe der praktischen Bedeutsamkeit nur dann wirklich einer Optimierung entspricht, wenn die Rücklaufquote hoch ist und die Schule von der Stichprobe repräsentiert wird, wenn die Grösse der Schule rechnerisch berücksichtigt wird und wenn die «Value Added Performance» ausgewiesen wird. Die Darstellung der praktischen Bedeutsamkeit führt deshalb ohne weitere methodische Anpassungen zu keiner überzeugenden Verbesserung.

Die Wahl der Methode der *Mittelwertsberechnung bestimmt die Fairness des Vergleichs*. Die Nutzung des arithmetischen Mittelwertes oder des OLS-Schätzers als Mass für die Schulmittelwerte ist nicht falsch. Die Parameter spiegeln die Datenlage sozusagen ungeschminkt. Unseres Erachtens ist ein Vergleich aufgrund dieser Mittelwerte – trotz Angaben von Konfidenzintervallen – allerdings unfair. Es wird weder die Grösse der Schule berücksichtigt noch die Tatsache, dass es zwischen den Schulen unter Umständen kaum Unterschiede gibt beziehungsweise dass die Varianz, die auf Schulmerkmale zurückzuführen ist, zu vernachlässigen ist. Dadurch besteht die Gefahr, dass in der Öffentlichkeit und in den Medien Differenzen diskutiert werden, die weder praktisch bedeutsam noch statistisch gesichert sind.

Ein weiteres Problem ist, dass die *Mittelwerte ungeachtet von relevanten Einflussgrössen* berechnet werden. Bei der Berechnung von Schulmittelwerten aufgrund der fachlichen Leistungen gehört es längst zum Standard, Schulmittelwerte als sogenannte Value Added Performance auszuweisen. Es gibt unseres Erachtens keinen Grund, weshalb Erkenntnisse über systematische Zusammenhänge zwischen individuellen Merkmalen der Befragten und retrospektiven Einschätzungen für die Schätzung von Schulmittelwerten nicht genutzt werden sollen. Zugleich muss festgehalten werden, dass momentan Erkenntnisse über die Bedeutung von schulischen Rahmenbedingungen und spezifischen Merkmalen der Schülerschaft für retrospektive Selbsteinschätzungen fehlen. Dies soll keine Kritik an der aktuellen Ergebnisdarstellung der Ehemaligenbefragung sein, sondern ein Hinweis darauf, wie problematisch die Rankings sind.

Aus den erwähnten Gründen lässt sich streng genommen auch nicht beurteilen, wer die beste oder die schlechteste Schule ist beziehungsweise welche Schule am besten oder am schlechtesten aufs Studium vorbereitet. Dazu müssten nach unserer Meinung ohnehin weitere, harte Kriterien einbezogen werden. Das vorliegende Ranking ist deshalb nicht nur aus methodischen Gründen, sondern vor allem auch aufgrund der einseitigen Datenlage wenig überzeugend: Ist das Liceo Artistico tatsächlich die beste Schule, nur weil die Abgängerinnen und Abgänger im Nachhinein die Fragen zur Zufriedenheit sehr positiv beantworten?

Die Gefahr für Simplifizierungen ist vorhanden. Es ist zu hoffen, dass die Schulen damit umgehen können. Trotzdem scheint es angebracht, in Zukunft über die methodischen Probleme von Schulvergleichen in einfacher Form zu informieren und die Schulen nicht aufgrund von unbedarften Publikationen verständlicherweise zu verärgern.