Nuremberg-Linz-Cracow-Skopje Research-Seminar

Correct Analysis of Complex Samples

Cracow, June, 30th to July, 2nd, 2009

Johann Bacher

Johannes Kepler Universität Linz

Linz 2009

- What are complex samples?
- Why standard statistical software is inappropriate?
- Which software is adequate?

1. Complex samples

= multistage samples with stratification and clustering, assumption of a simple random sample from a large population (i.i.d. assumption) does not fit

Examples

- Mikrozensus (Haslinger/Kytir 2005; Stadler 2005)
- PISA (Programme for International Student Assessment, OECD 2005b; Schreiner u.a. 2007)
- PIRLS (Progress in International Reading Literacy Study; Mullis et al. 2007; Suchan u.a. 2007)
- Öibf-Bildungsstudie (Schlögl/Lachmayr 2004; Bacher/Beham/Lachmayr 2008)

Figure 1: Example of a complex sample

	SchülerInnen	Eltern	
	Übergang in die Sekundarstufe I		
	nicht befragt	15x VS 4. Klasse 15x HS 1. Klasse	
		15x AHS 1. Klasse	15x HS 1. Klasse Nahbereich
	Übergang in die Sekundarstufe II		
	15x HS 4. Klasse	15x HS 4. Klasse	
	15x AHS 4. Klasse	15x AHS 4. Klasse	
	15x AHS 5. Klasse	15x AHS 5. Klasse	
	15x BMS 1. Klasse	15x BMS 1. Klasse	
	15x BS/PT 1. Klasse	15x BPS/PT 1. Kl.	
	15x BHS 1. Klasse	15x BHS 1. Klasse	
	Übergang in den Tertiärbereich		
	15x BHS 5. Klasse	15x BHS 5. Klasse	
	15x AHS 8. Klasse	15x AHS 8. Klasse	

Quelle: Eigendarstellung öibf, entnommen aus Bacher/Beham/Lachmayr (2008: 69)

Characteristics (Sturgis 2004)

- multistage sample with stratification and clustering
- weighting (unequal selection probabilities and unit-nonresponse)

Characteristics (Wolter 1985)

- degree of complexity of sample design
- degree of complexity of sample estimator
- multiple characteristics of variables of interest
- descriptive and analytical uses of the survey data
- the scale or size of survey

2. Analysis of complex samples with standard software?

- Using wrong weights → biased estimations of parameters (mean, proportions, correlation etc.)
- ignoring complex sample → biased estimation of standard errors of parameters
- multistage samples → underestimation of standard errors → overestimation of significance

Figure 2: Examples of incorrect use

	srs	complex samples
unweighted mean in Mathe PISA2006		
Austria (n=4927)	510	510
Poland (n=5547)	501	501
weighted mean in Mathe		
Austria	506	506
Poland	495	495
standard error		
Austria	1,52	3,74
Poland	1,31	2,44
test statistc for 500 for Austria	$t = \frac{506-500}{1,52} = 3,94$	$t = \frac{506-500}{3,74} = 1,60$
p(one sided)	0,000	0,055
p(two sided)	0,000	0,110

3. Measuring design effect of complex samples

Figure 3: Design effect measures of complex samples

design effect: $DEFF(T) = \frac{\sigma(T)^2_{komplex}}{\sigma(T)^2_{einfach}}$	$DEFF(T) = \frac{3{,}74^2}{1{,}52^2} = 6{,}05$
$DEFFSQRT(T) = \sqrt{\frac{\sigma(T)^2_{komplex}}{\sigma(T)^2_{einfach}}} = \frac{\sigma(T)_{komplex}}{\sigma(T)_{einfach}}$	$DEFFSQRT(T) = \frac{3{,}74}{1{,}52} = 2{,}46$
effective sample size $NEFF(T) = \frac{n(T)_{komplex}}{DEFF(T)}$	$NEFF(T) = \frac{4927}{6{,}05} = 814$

4. Correct weighting

The correct weights are for two stage samples: $w_{hij} = \frac{1}{p_{hi}} \cdot \frac{1}{p_{j/hi}}$

Additionally adjustment factors are used: $w_{hij} = \frac{1}{f_{hi} \cdot p_{hi}} \cdot \frac{1}{f_{j/hi} \cdot p_{j/hi}}$

The weights are defined for the target population in analysis of complex samples. They can be rescaled to sample size by: $w^*_{hij} = \frac{n}{N} \cdot w_{hij}$

An example:

Strata	School	Sample frame	p1	Sample frame within school	p2	weights	“wrong” weights (a)
1	1	3 of 150	0,02	5 of 20	0,25	200,0000	200,0000
1	2	3 of 150	0,02	5 of 25	0,20	250,0000	200,0000
1	3	3 of 150	0,02	5 of 15	0,33	150,0000	200,0000
2	1	2 of 25	0,08	5 of 18	0,277	45,0000	50,0000
2	2	2 of 25	0,08	5 of 22	0,227	55,0000	50,0000

(a) Weights compensate for disproportionality of strata, giving the first stratum a higher weight than the second stratum, but ignore differences between schools within a stratum

5. Calculation of standard errors

- linearisation via Taylor series (→ SPSS, STATA …)
- BRR-weights (→ PISA, OECD 2005a; Stata)
- Jack knife-procedure (→ PIRLS, IEA2008, Stata)

→ Lee/Forthofer (2008)

Two stage sampling:

$$V(T) = V_2(T) = \underbrace{\sum_{h=1}^{H} U_h}_{V_1(T)} + \sum_{h=1}^{H}\sum_{i=1}^{n_h} \pi_{hi} \sum_{k=1}^{K_{hi}} U_{hik}$$

$T = \hat{Y}$ = Sum of attribute, e.g. Number of girls in target population

U_h = Sample variance of primary units (PSU, e.g. schools) within strata

π_{hi} = selection probabilities of PSU i of stratum h

U_{hik} = Sample variance of secondary units (SSU, e.g. school children) of PSU i of stratum h

Three stage sampling:

$$V(T) = V_3(T) = V_2(T) + \sum_{h=1}^{H}\sum_{i=1}^{n_h}\pi_{hi}\sum_{k=1}^{K_{hi}} f_{hik}\sum_{j=1}^{n_{hik}}\sum_{l=1}^{L_{hikj}} U_{hikjl}$$

For the mean, the following new variable must be generated

$$z_{hij} = w_{hij}\left(y_{hij} - \hat{\bar{Y}}\right)/\hat{N}$$

The quoted formula can be used.

Special software for sampling

- WesVar (WESTAT 2008)
- SUDAAN (RTI-International 2008; Lee/Forthofer 2006).

Modules of standard statistical software

- ComplexSample (SPSS Inc. 2008)
- PROC SURVEY (SAS Institute 2008)
- SURVEY METHODS (StataCorp 2008)
- SVY in R (Lumley 2003)

Special software, e.g. HLM, LatentGold, …

Figure 4: Comparision of COMPLEX SAMPLE in SPSS and SURVEY METHODS in STATA

	STATA9	SPSS14
drawing a sample	no	yes
variance estimation		
• Taylor-linearization	yes	yes
• Taylor for unequal selection prob. in stage I	no	yes
BBR-weights	yes	no
Jack knife-weights	yes	no
Statistical procedures		
• (relative) frequencies	yes	yes
• means	yes	yes
• ratio estimators	yes	yes
• cross tabulation incl. test of independence	yes	yes

	STATA9	SPSS14
• t-test for theoretical mean	no (a)	no (a)
• t-test for independent samples	no (b)	no (b)
• t-Test for dependent samples	no (c)	no (c)
• correlations coefficients	no (d)	no (d)
• general linear model	no	yes
• linear regression	yes	no (e)
• logistic regression	yes	yes
• regression for ordinal variables	yes	yes
• special regression modules	yes (f)	no

(a) easily to test, generate a new variable $y^* = y - \mu$ and test if y^* is different from zero

(b) can be tested using linear regression or general linear model

(c) easily to test, generate a new variable $d = y_1 – y_2$ with y_1 = first measure and y_2 = second measure and test if d is equal zero

(d) can be computed and tested via linear regression or Pearson's r, Phi, point biseral r.

(e) part of general linear model

(f) e.g. Probit-Regression, Intervall-Regression, Poisson-Regression etc (StataCorp. 2005)

5. Complex Sample in SPSS

two applications must be distinguished:

- The sample is drawn with Complex Samples: Complex Sample can be used for different designs
- The sample is already drawn (secondary analysis): Complex Sample can be used for samples with equal selection probabilities at stage one, but not for unequal selection probabilities, like probabilities proportional to size samples (PPS)

Example:

```
compute p1=-9.
if (schicht = 1) p1=3/150.
if (schicht = 2) p1=2/25.
fre var=p1.

compute p2=-9.
if (schule = 1) p2=5/20.
if (schule = 2) p2=5/25.
if (schule = 3) p2=5/15.
if (schule = 4) p2=5/18.
if (schule = 5) p2=5/22.
fre var=p2.

compute wtot=(1/p1)*(1/p2).


* Analysevorbereitungsassistent.
CSPLAN ANALYSIS
 /PLAN FILE='d:\tarnai\beispiel1roh'
 /PLANVARS ANALYSISWEIGHT=wtot
 /PRINT PLAN
```

```
 /DESIGN  STRATA= Schicht  CLUSTER= Schule
   /ESTIMATOR TYPE=EQUAL_WOR
   /INCLPROB VARIABLE= p1
 /DESIGN CLUSTER= idnr
   /ESTIMATOR TYPE=EQUAL_WOR
   /INCLPROB VARIABLE= p2.

* Deskriptive Statistiken für komplexe Stichproben.
CSDESCRIPTIVES
 /PLAN FILE = 'beispiel1roh'
 /SUMMARY VARIABLES =weiblich Testscore
 /MEAN
 /STATISTICS SE count popsize cin
 /MISSING SCOPE = ANALYSIS CLASSMISSING = EXCLUDE.
```

		Schätzung	Standardfehler	95%-Konfidenzintervall		Umfang der Grundgesamtheit	Ungewichtete Anzahl
		Untere Grenze	Obere Grenze	Untere Grenze	Obere Grenze	Untere Grenze	Obere Grenze
Mittelwert	weiblich	,5429	,05933	,3541	,7317	3500,000	25
	Testscore	458,98	22,443	387,55	530,40	3500,000	25

Figure 5: Example of a multivariate analysis

	b (unstand.)	ß (stand.)	t (einfach)	p (einfach)	t (komplex)	p (komplex)
(Konstante)	0,31		29,75	0,000	15,98	0,000
BUB	0,00	0,00	0,20	0,841	0,19	0,849
WALLEIN	-0,01	-0,01	-0,23	0,816	-0,16	0,871
VERANT	0,03	0,04	1,70	0,089	1,43	0,163
SCHICHT	*0,02*	*0,10*	*3,94*	*0,000*	*2,93*	*0,006*
AHS_Nähe	*0,18*	*0,19*	*8,32*	*0,000*	*3,06*	*0,004*
MATURA	*0,30*	*0,32*	*8,68*	*0,000*	*6,62*	*0,000*
SCHLEIST	*0,19*	*0,35*	*9,37*	*0,000*	*7,68*	*0,000*
Interaktionen						
BUB*WALLEIN	0,06	0,02	1,00	0,316	1,00	0,326
BUB*VERANT	0,01	0,01	0,30	0,763	0,29	0,777
BUB*SCHICHT	0,01	0,04	1,47	0,142	*1,92*	*0,064*
BUB*AHS_Nähe	0,06	0,03	1,34	0,182	1,26	0,216
BUB*MATURA	-0,07	-0,05	-1,35	0,177	-1,62	0,116
BUB*SCHLEIST	-0,03	-0,04	-1,02	0,310	-1,09	0,283

BUB = Geschlecht des Kindes, WALLEIN = weiblicher Alleinerzieherhaushalt, VERANT = väterliche (Mit-)Verantwortung, SCHICHT = soziale Schicht der Eltern, AHS_Nähe = AHS in Wohnortnähe, MATURA = Bildungsaspiration der Eltern (1=Matura oder höher, 0=sonst), SCHLEIST = schulischen Leistungen in der 4. Klasse VS,

7. Analysis of PISA

Generally, PPS-Samples are used, BBR weights are included in the international data set

Table 3.13 ■ The Fay replicates

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R 10	R 11	R 12
1	1	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5
1	2	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5
2	3	1.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5
2	4	0.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5
3	5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5
3	6	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5
4	7	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5
4	8	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5
5	9	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5
5	10	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G(1-k)^2}\sum_{i=1}^{G}(\hat{\theta}_{(i)} - \hat{\theta})^2 = \frac{1}{80(1-0.5)^2}\sum_{i=1}^{80}(\hat{\theta}_{(i)} - \hat{\theta})^2 = \frac{1}{20}\sum_{i=1}^{80}(\hat{\theta}_{(i)} - \hat{\theta})^2$$

OECD (2005a: 50-51)

The data can be analysed with SPSS macros, provided by the OECD:

	without plausibile values	with plausible values
percentages	mcr_se_GrpPct.sps	mcr_se_PctLev.sps for competence levels
univariate statistics	mcr_se_univ.sps	mcr_se_pv.sps
Mittelwertdifferenzen	mcr_se_dif.sps (mcr_se_wleqrt.sps)	mcr_se_pv_dif_PV.sps (mcr_se_pv_wleqrt.sps)
Korrelationen	mcr_se_cor.sps	mcr_se_cor_1PV.sps mcr_se_cor_2PV.sps
Regression	mcr_se_regr.sps	mcr_se_regr_PV.SPS

Examples of macros

```
* DEFINE MACRO.
Include file='d:\PISA2003\makros\mcr_SE_pv.sps'.

pv nrep = 80/
stat = mean/
dep = math /
grp = cnt/
wgt = w_fstuwt/
rwgt = w_fstr/
cons = 0.05/
infile = 'd:\PISA2003\Data\Pisa2003Complex.sav'.
```

```
CNT     stat          SE

AUT   505,611   3,266190
```

Include file='d:\PISA2003\makros\mcr_SE_reg_PV.sps'.

```
reg_PV nrep = 80/
ind=migra weiblich hisei /
dep=math/
grp = cnt /
wgt = w_fstuwt/
rwgt = w_fstr/
cons = 0.05/
infile = 'd:\PISA2003\Data\Pisa2003Complex.sav'/.
```

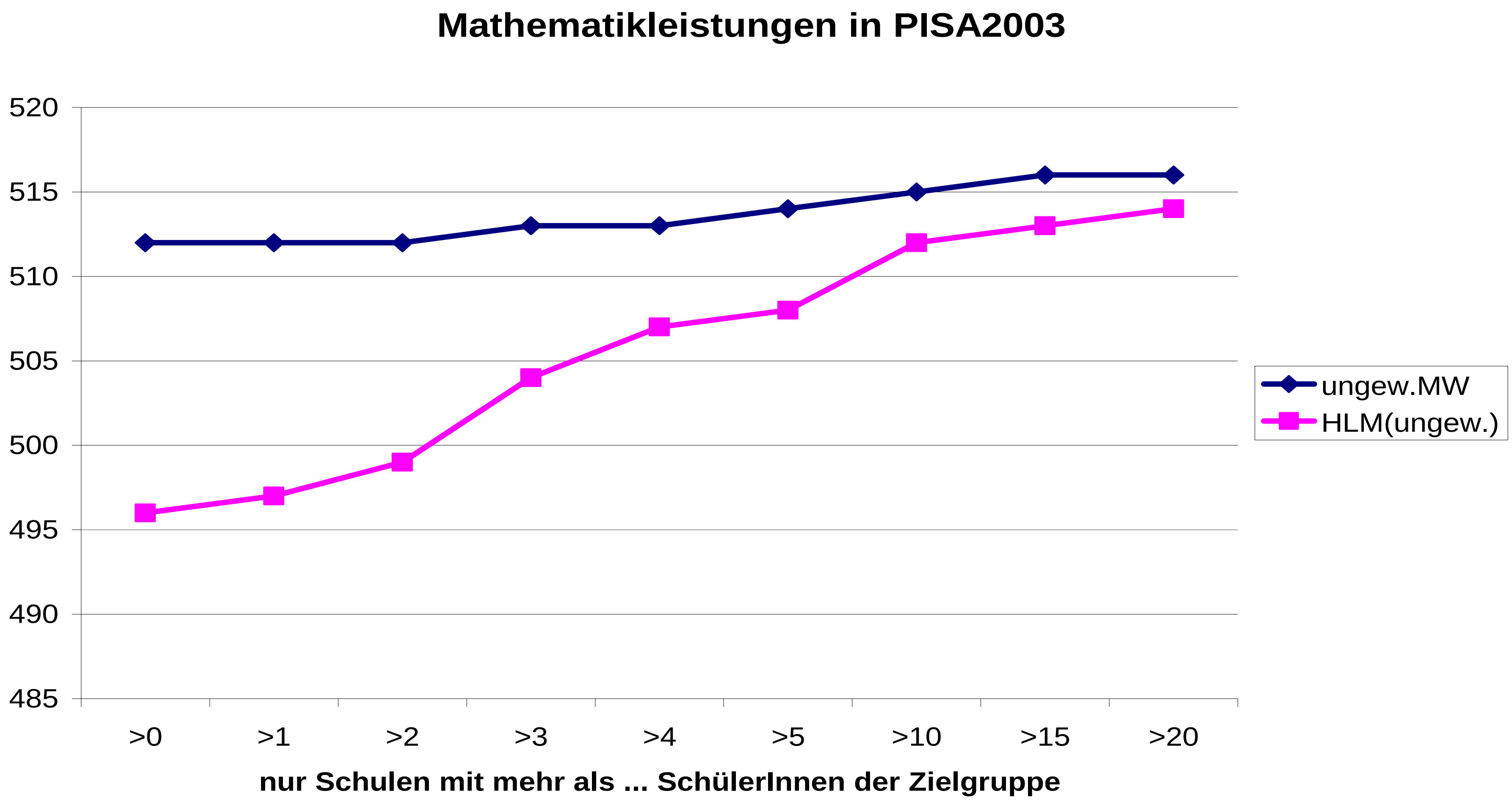
CNT	ind	stat	SE
AUT	b0	440,837	6,933528
AUT	HISEI	1,686	,123306
AUT	migra	-44,496	5,629917
AUT	weiblich	-11,829	3,890988

Using Complex Sample?

	Statistik	SE-BRR	CS SE-WR	CS SE-Equal	CS SE-SRS
Univariate Statistiken					
Weiblich	0,499	0,015634	0,01893	0,01792	0,00772
Mathematik	506	3,266190	3,60970	3,45771	1,54995
Abhängige = CULTPOSS					
AUT b0	-1,011	0,054747	0,05196822	0,050265	0,05052176
AUT HISEI	0,019	0,001114	0,00107307	0,00103822	0,00092326
AUT migra	-0,134	0,045301	0,04495233	0,04335574	0,04228967
AUT weiblich	0,193	0,036737	0,03986684	0,03815616	0,02920098
abh. Variable = math					
AUT b0	440,837	6,933528	8,24874994	7,92719199	4,85484573
AUT HISEI	1,686	0,123306	0,13474943	0,13045336	0,0879692
AUT migra	-44,496	5,629917	6,17232182	5,89585233	4,08844184
AUT weiblich	-11,829	3,890988	4,44537633	4,18574706	2,74470983

CS results in an overestimation of standard erros = not recommended

Using HLM?


Mathematikleistungen in PISA2003
520
515
510
505
500
495
490
485
>0
>1
>2
>3
>4
>5
>10
>15
>20
nur Schulen mit mehr als ... SchülerInnen der Zielgruppe
ungew.MW
HLM(ungew.)

HLM is not recommended for descriptive purposes, but can be used for causal modelling. Before using, test if the assumptions are fulfilled. Strata must be included as context attribute, assumptions should be tested

	SPSS-Makro		HLM	
	Statistik	SE-BBR	Statistik	SE
Univariate Statistiken*				
Weiblich	0,505	0,016200	0,4832	0,017385
Mathematik	505,611	3,266190	500,5037	3,491675
Abhängige = CULTPOSS*				
AUT b0	-1,011	0,054747	-0,6646	0,059228
AUT HISEI	0,019	0,001114	0,0119	0,001161
AUT migra	-0,134	0,045301	-0,13478	0,04420
AUT weiblich	0,193	0,036737	0,1230	0,04256
abh. Variable = math*				
AUT b0	444,159	7,1555336	503,901	5,76599
AUT HISEI	1,644	0,127808	0,2297	0,08772
AUT migra	-44,498	5,644853	-38,167	3,70743
AUT weiblich	-12,529	3,967025	-17,004	2,96231

* unter Ausschluss von Fällen mit mind. einem fehlenden Wert in HISEI, MIGRA, WEIBLICH, CULTPOSS, MATH

Professional users should use stata, esp. the modul PV of MacDonald.

	SPSS-Makro		STATA	
	Statistik	SE-BBR	Statistik	SE-BBR
Univariate Statistiken				
Weiblich	0,499	0,015634	0,499	0,015634
Mathematik	506	3,266190	506	3,266190
Abhängige = CULTPOSS *				
AUT b0	-1,011	0,054747	-1,011	0,054747
AUT HISEI	0,019	0,001114	0,019	0,001114
AUT migra	-0,134	0,045301	-0,134	0,045301
AUT weiblich	0,193	0,036737	0,193	0,036737
abh. Variable = math *				
AUT b0	444,159	7,1555336	444,159	7,1555336
AUT HISEI	1,644	0,127808	1,644	0,127808
AUT migra	-44,498	5,644853	-44,498	5,644853
AUT weiblich	-12,529	3,967025	-12,529	3,967025

* unter Ausschluss von Fällen mit mind. einem fehlenden Wert in HISEI, MIGRA, WEIBLICH, CULTPOSS, MATH

Other Software? will be tested

8. Summary

- Correct analysis of complex sample is important
- SPSS' Complex Sample is appropriate to analyse complex samples, esp. if samples are drawn with Complex Sample
- CS is not appropriate to analyse PISA-Data. However, macros are available for SPSS' users
- HLM also is not appropriate if schools with a small number of students are in the sample
- If the PISA-data should be analysed in a professional way, STATA should be used until SPSS implements BBR weights and plausible values

Thank you for your attention!

johann.bacher@jku.at

References

Bacher, J. 2006: Stichprobendesign, Sozialstruktur und regionale Unterschiede. In E. Neuwirth, E., I. Ponocny & W. Grossmann (Hg.): PISA 2000 und PISA 2003: Vertiefende Analysen und Beiträge zur Methodik (S. 39-51). Graz: Leykam.

Bacher, J., Beham, M., & Lachmayr, N., 2008: Geschlechterunterschiede in der Bildungswahl. Wiesbaden: VS Verlag.

Berger, Y. G., 2004: A Simple Variance Estimator for Unequal Probability Sampling without Replacement. Journal of Applied Statistics, Vol. 31, 305-315

Cohen, J., 1977: Statistical Power Analysis for Behavorial Sciences. Revised Edition. New York ua.

Haslinger, A. & Kytir, J., 2005: Stichprobendesign, Stichprobenziehung und Hochrechnung des Mikrozensus ab 2004. Statistische Nachrichten, 6, 510-518

IEA 2008: IDB-Analyzer. http://pirls.bc.edu/pirls2006/user_guide.html.

Lee, E. S. & Forthofer, R. N., 2006: Analyzing Complex Survey Data. Second Edition. New York: Sage.

Lumley, T., 2003: Analyzing Survey Data in R. R-News, 3(1), 17-20.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M. & Foy, P., 2007: PIRLS 2006 International Report. Boston: IEA (http://pirls.bc.edu/isc/publications.html#p06, 30.6.2008)

OECD (Ed.), 2005a: PISA 2003. Data Analysis Manual. SPSS® Users. Paris: OECD

OECD (Ed.), 2005b: PISA2003. Technical Report. Paris: OECD

Raudenbusch, S. u.a., 2004: HLM6. Scientific Software

RTI-International, 2008: SUDAAN. (http://www.rti.org/sudaan/index.cfm, 30.6.2008)

SAS Institute, 2008: Statistical Analysis with SAS/STAT® Software. (http://www.sas.com/technologies/analytics/statistics/stat/features.html, 30.6.2008)

Schlögl, P., Lachmayr, N., 2004b: Soziale Situation beim Bildungszugang. Motive und Hintergründe von Bildungswegentscheidungen in Österreich. Wien: Eigenverlag.

Schreiner, C., Breit, S., Schwantner, U. & Grafendorfer, A., 2007: PISA2006. Internationaler Vergleich von Schülerleistungen. Die Studie im Überblick. Graz: Leykam.

SPSS Inc., 2008: SPSS Complex Samples™ (http://www.spss.com/complex_samples/data_analysis.htm, 30.6.2008).

Stadler, B., 2005: Daten zum österreichischen Arbeitsmarkt. Österreichische Zeitschrift für Soziologie, 30 (3), 89-100.

StataCorp, 2008: Survey Methods. (http://www.stata.com/capabilities/svy.html, 30.6.2008)

Sturgis, P., 2004: Analysing Complex Survey Data: Clustering, Stratification and Weights. Social research UPDATE, Issue 43, University of Surrey.

Suchan B., Wallner-Paschon Chr., Stöttinger, E. & Bergmüller, S., 2007: PIRLS 2006. Internationaler Vergleich von Schülerleistungen. Graz: Leykam

WESTAT, 2008: WesVar – Software for Analysis of Data form Complex Sample. (http://www.westat.com/wesvar/index.html, 30.6.2008)