Plagiarism in student papers

Prevalence estimation using special techniques for sensitive questions

Ben Jann

ETH Zurich, jannb@ethz.ch

Venice International University
November 30, 2009

Outline

- Introduction
  - Plagiarism
  - Approaches to Estimate the Prevalence of Plagiarism
- Using Dejeopardizing Techniques to Measure Plagiarism
  - Study A: Randomized Response Technique
  - Study B: Item Count Technique
  - Study C: The Crosswise Model
- Conclusions

Plagiarism

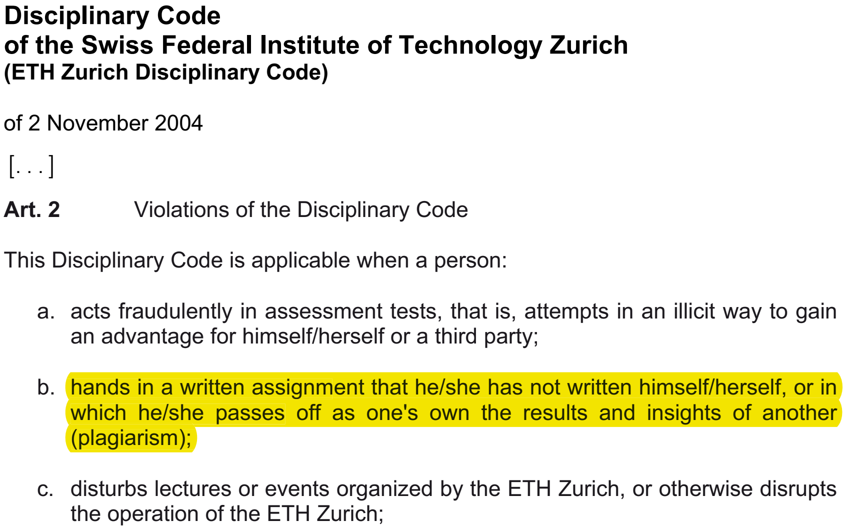
- What is plagiarism?

Definition of the U.S. National Academy of Sciences

"Appropriation of another person's ideas, processes, results, or words without giving appropriate credit, including those obtained through confidential review of others' research proposals and manuscripts"

- In the age of the Internet, Wikipedia, etc. Universities increasingly begin to worry about plagiarism in student papers and homework assignments.

Plagiarism



Disciplinary Code
of the Swiss Federal Institute of Technology Zurich
(ETH Zurich Disciplinary Code)

of 2 November 2004

[. . .]

Art. 2 Violations of the Disciplinary Code

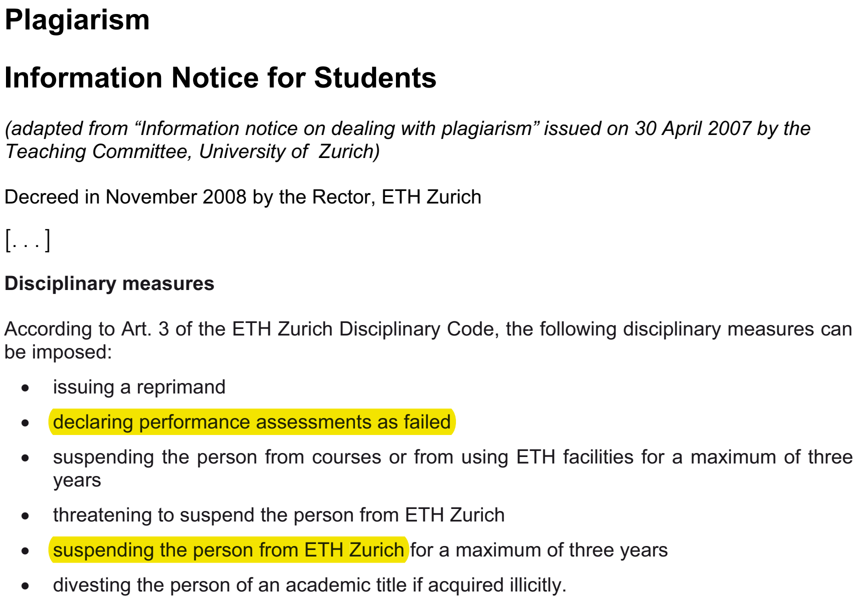
This Disciplinary Code is applicable when a person:

a. acts fraudulently in assessment tests, that is, attempts in an illicit way to gain an advantage for himself/herself or a third party;

b. hands in a written assignment that he/she has not written himself/herself, or in which he/she passes off as one's own the results and insights of another (plagiarism);

c. disturbs lectures or events organized by the ETH Zurich, or otherwise disrupts the operation of the ETH Zurich;

Plagiarism



Plagiarism

Information Notice for Students

(adapted from “Information notice on dealing with plagiarism” issued on 30 April 2007 by the Teaching Committee, University of Zurich)

Decreed in November 2008 by the Rector, ETH Zurich

[. . .]

Disciplinary measures

According to Art. 3 of the ETH Zurich Disciplinary Code, the following disciplinary measures can be imposed:

- issuing a reprimand
- declaring performance assessments as failed
- suspending the person from courses or from using ETH facilities for a maximum of three years
- threatening to suspend the person from ETH Zurich
- suspending the person from ETH Zurich for a maximum of three years
- divesting the person of an academic title if acquired illicitly.

Plagiarism

- It might be important for Universities to know something about how frequent different forms of plagiarism occur.
- Asking students directly may yield biased estimates because plagiarism is a sensitive topic. Will Students be willing to tell the truth?
- *"A question is sensitive when it asks for a socially undesirable answer, when it asks, in effect, that the respondent admits he or she has violated a social norm"* (Tourangeau/Yan 2007: 860).

Approaches to Estimate the Prevalence of Plagiarism

- Data collection without asking questions
  - Official number of students found guilty
  - Systematic inspection of student papers using special software
    - ★ Krohn/Schlombs/Taubert (2003): 10 out of 39 group seminar papers at the University of Bielefeld were identified as partial or severe plagiarism (using Google).
    - ★ Sattler (2007): 19.5% of papers from 159 students of the University of Leipzig were identified as partial plagiarism (using Plagiarism-Finder software).
- Direct questions
  - Self-reports (past behavior; intentions)
  - Other-reports (plagiarism of other students)
    - ★ Knoop (2006): 32.3% of 192 interviewed students at University of Münster reported to know at least one plagiarizing fellow student.
- Dejeopardizing question techniques
  - Randomized Response, Item Count Technique, etc.

Using Dejeopardizing Techniques to Measure Plagiarism

- I will now present results from three studies in which dejeopardizing techniques were used to estimate the prevalence of plagiarism.
  - Study A: Randomized Response Technique
  - Study B: Item Count Technique
  - Study C: The Crosswise Model
- The three studies were implemented as methodological experiments using direct questioning as control condition.

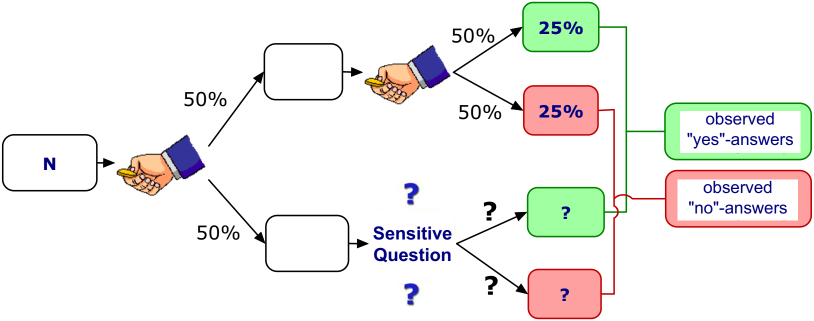
The Randomized Response Technique (RRT)

(Warner 1965; also see, e.g., Fox and Tracy 1986)

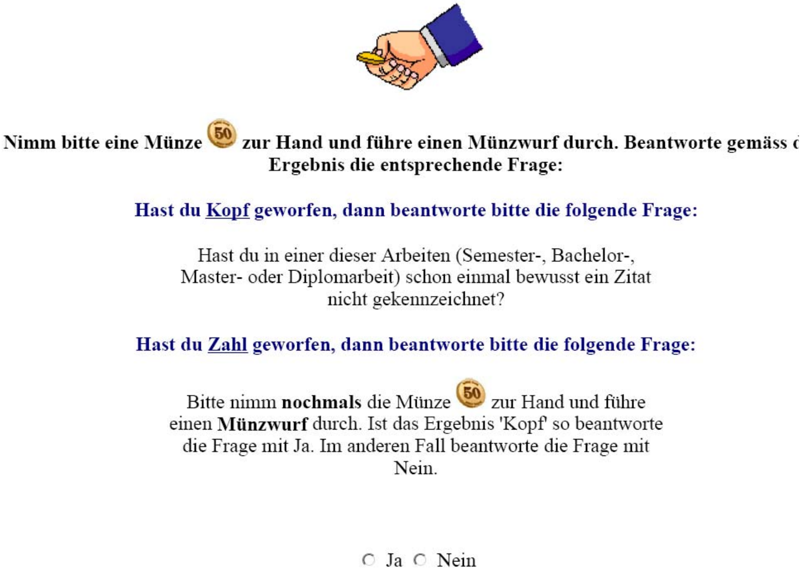
- Basic idea: anonymity through randomization.
- Depending on the outcome of a randomization device (e.g. roll a dice), the respondent has to answer the sensitive question or give an automatic “yes” or “no” answer (or answer an unthreatening question of which the distribution is known).
- Since only the respondent knows the outcome of the randomization device, a “yes” answer cannot be interpreted as an admission of guilt.
- However, the proportion of the sample that has engaged in the behavior of interest can be calculated with knowledge of the properties of the randomizing device.

Using RRT to Measure Plagiarism

- Web-Survey among ETH students in 2005
- Response rate: 33 Percent
- Research team: Elisabeth Coutts, Andreas Diekmann, Georg Böcherer, Stefan Senn, Philipp Stadelmann, Diego Stutzer
- Used RRT-design:



Using RRT to Measure Plagiarism



Nimm bitte eine Münze zur Hand und führe einen Münzwurf durch. Beantworte gemäss d
Ergebnis die entsprechende Frage:

Hast du Kopf geworfen, dann beantworte bitte die folgende Frage:

Hast du in einer dieser Arbeiten (Semester-, Bachelor-, Master- oder Diplomarbeit) schon einmal bewusst ein Zitat nicht gekennzeichnet?

Hast du Zahl geworfen, dann beantworte bitte die folgende Frage:

Bitte nimm **nochmals** die Münze zur Hand und führe einen **Münzwurf** durch. Ist das Ergebnis 'Kopf' so beantworte die Frage mit Ja. Im anderen Fall beantworte die Frage mit Nein.

○ Ja ○ Nein

Using RRT to Measure Plagiarism

- Results: plagiarism prevalence estimates (in percent)

	direct questions	RRT	difference
seminar/term paper,	12.0	3.7	−8.3
diploma thesis	(2.0)	(4.0)	(4.4)
	N = 266	N = 495	
other written	19.4	17.6	−1.8
assignments	(1.4)	(2.4)	(2.8)
	N = 826	N = 1521	

(standard errors in parentheses)

Using RRT to Measure Plagiarism

- Explanations for the unexpected results:
  - difficulties understanding RRT, no trust in RRT
  - Web-surveys already anonymous enough?
  - “Self-protective no” bias: Respondents who did not commit plagiarism are reluctant to give a “yes” answer to the non-sensitive question.
- Approaches to deal with the “self-protective no” bias
  - directly approach the problem using specific instructions
  - apply methods to detect cheaters and correct the RRT estimates
  - use alternative methods that are not (or less) affected by the “self-protective no” bias

The Item Count Technique (ICT)

(see, e.g., Dalton et al. 1994, Raghavarao and Federer 1979)

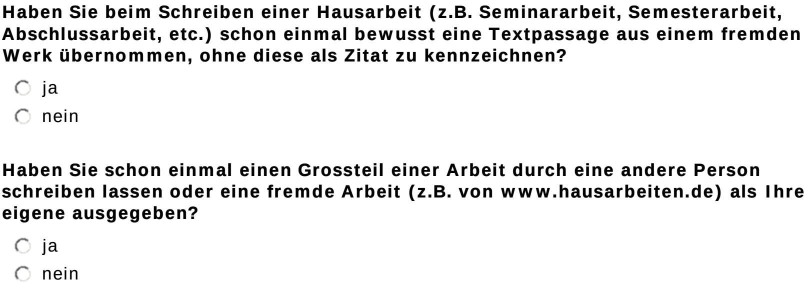
- Given a list of statements, respondents report how many of them are true, but not which ones. For some respondents the list contains the sensitive item, for others not (randomized).
- Example: "How many of the following statements apply to you?"

Group A (short list)	Group B (long list)
I have a cat.	I have a cat.
I have blue eyes.	I have blue eyes.
I like country music.	I like country music.
	I use drugs.

- Prevalence estimate = mean difference
- Advantage: Requires no randomization device.

Using ICT to Measure Plagiarism

- Web-Survey among students of the University of Konstanz, Summer 2009
- Response rate: 23.7 Percent
- Research team: Ben Jann and Philipp Stirnemann (thanks to Thomas Hinz, Katrin Auspurg, and Pascal Gienger from the University of Konstanz for supporting the project)
- Questions:



Haben Sie beim Schreiben einer Hausarbeit (z.B. Seminararbeit, Semesterarbeit, Abschlussarbeit, etc.) schon einmal bewusst eine Textpassage aus einem fremden Werk übernommen, ohne diese als Zitat zu kennzeichnen?

○ ja

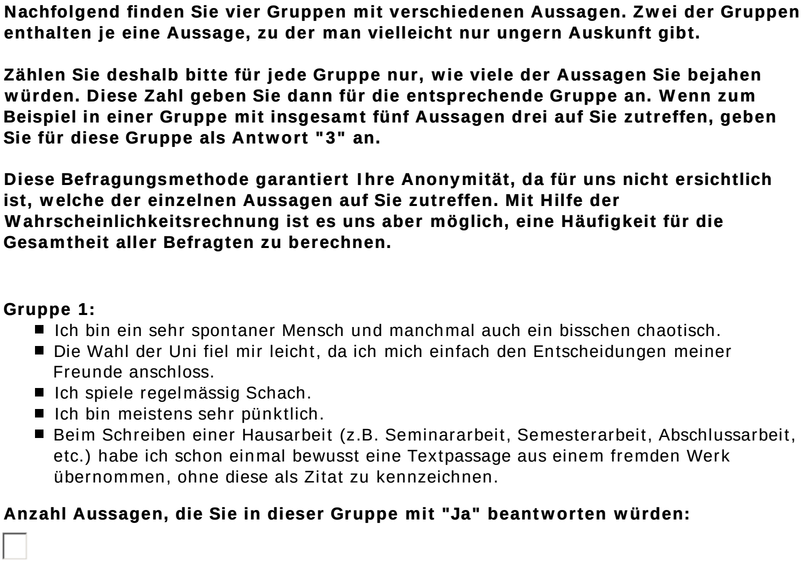
○ nein

Haben Sie schon einmal einen Grossteil einer Arbeit durch eine andere Person schreiben lassen oder eine fremde Arbeit (z.B. von www.hausarbeiten.de) als Ihre eigene ausgegeben?

○ ja

○ nein

Using ICT to Measure Plagiarism



Nachfolgend finden Sie vier Gruppen mit verschiedenen Aussagen. Zwei der Gruppen enthalten je eine Aussage, zu der man vielleicht nur ungern Auskunft gibt.

Zählen Sie deshalb bitte für jede Gruppe nur, wie viele der Aussagen Sie bejahen würden. Diese Zahl geben Sie dann für die entsprechende Gruppe an. Wenn zum Beispiel in einer Gruppe mit insgesamt fünf Aussagen drei auf Sie zutreffen, geben Sie für diese Gruppe als Antwort "3" an.

Diese Befragungsmethode garantiert Ihre Anonymität, da für uns nicht ersichtlich ist, welche der einzelnen Aussagen auf Sie zutreffen. Mit Hilfe der Wahrscheinlichkeitsrechnung ist es uns aber möglich, eine Häufigkeit für die Gesamtheit aller Befragten zu berechnen.

Gruppe 1:

- Ich bin ein sehr spontaner Mensch und manchmal auch ein bisschen chaotisch.
- Die Wahl der Uni fiel mir leicht, da ich mich einfach den Entscheidungen meiner Freunde anschloss.
- Ich spiele regelmässig Schach.
- Ich bin meistens sehr pünktlich.
- Beim Schreiben einer Hausarbeit (z.B. Seminararbeit, Semesterarbeit, Abschlussarbeit, etc.) habe ich schon einmal bewusst eine Textpassage aus einem fremden Werk übernommen, ohne diese als Zitat zu kennzeichnen.

Anzahl Aussagen, die Sie in dieser Gruppe mit "Ja" beantworten würden:

□

Using ICT to Measure Plagiarism

- Results: plagiarism prevalence estimates

	direct question	ICT 1	ICT 2
	(400)	(858)	(855)
partial plagiarism	8.0%	9.1%	10.4%
	(1.4%)	(5.3%)	(6.4%)
full plagiarism	2.0%	−6.8%	−0.2%
	(0.7%)	(5.9%)	(6.8%)

(standard errors in parentheses)

The Crosswise Model

(Yu, Tian, and Tang 2007)

- Very simply idea: Ask a sensitive question and a non-sensitive question and let the respondent indicate . . .
  - A: whether the answer is “yes” to both questions or “no” to both questions
  - B: whether the answer is “yes” to one questions and “no” to the other

		non-sensitive question	
		no	yes
sensitive question	no	**A**	**B**
	yes	**B**	**A**

- In either case, the researcher does not know whether the answer to the sensitive question is “yes” or “no” for a specific respondent.
- The prevalence of the non-sensitive item must be unequal 0.5 and known (furthermore, the non-sensitive item must be independent of the sensitive item).

The Crosswise Model

(Yu, Tian, and Tang 2007)

- Let
  - X be the observed answer ("A" or "B")
  - Y be the sensitive question with $\pi_Y = \Pr(Y = \text{yes})$
  - Z be the non-sensitive question with $\pi_Z = \Pr(Z = \text{yes}) \neq 0.5$
  - $\text{Cov}(Y, Z) = 0$
- Then: $\pi_A = \Pr(X = \text{A}) = (1 - \pi_Y)(1 - \pi_Z) + \pi_Y \pi_Z$
- Hence: A natural estimator for π_Y is

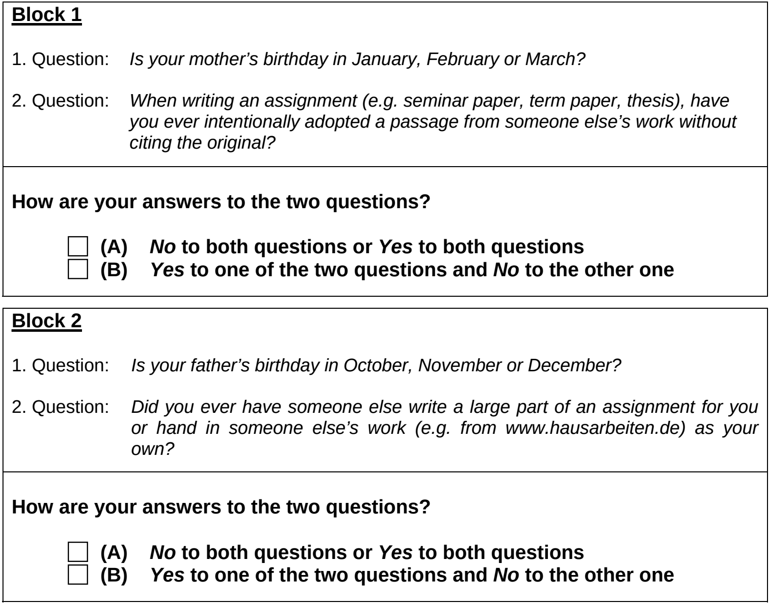
$$\hat{\pi}_Y = \frac{\hat{\pi}_A + \pi_Z - 1}{2\pi_Z - 1} \qquad \text{Var}(\hat{\pi}_Y) = \frac{\text{Var}(\hat{\pi}_A)}{(2\pi_Z - 1)^2}$$

- Note that formally the crosswise model is identical to Warner's model.

Using the Crosswise Model to Measure Plagiarism

- Classroom survey (written questionnaire) at different Universities (ETH Zurich, University Leipzig, LMU Munich), Spring/Summer 2009
- Total sample size approx. 500.
- 3/4 crosswise model, 1/4 direct questions
- Research team: Ben Jann, Julia Jerke, Ivar Krumpal (thanks to Norman Braun and Jochen Groß from LMU Munich for their support).

Using the Crosswise Model to Measure Plagiarism



Block 1

1. Question: *Is your mother's birthday in January, February or March?*

2. Question: *When writing an assignment (e.g. seminar paper, term paper, thesis), have you ever intentionally adopted a passage from someone else's work without citing the original?*

How are your answers to the two questions?

□ **(A)** ***No* to both questions or *Yes* to both questions**
□ **(B)** ***Yes* to one of the two questions and *No* to the other one**

Block 2

1. Question: *Is your father's birthday in October, November or December?*

2. Question: *Did you ever have someone else write a large part of an assignment for you or hand in someone else's work (e.g. from www.hausarbeiten.de) as your own?*

How are your answers to the two questions?

□ **(A)** ***No* to both questions or *Yes* to both questions**
□ **(B)** ***Yes* to one of the two questions and *No* to the other one**

Using the Crosswise Model to Measure Plagiarism

- Results: plagiarism prevalence estimates (in percent)

	direct questions (N = 96)	crosswise (N = 310)	difference
partial plagiarism	7.3 (2.7)	22.3 (5.5)	15.0 (6.1)
full plagiarism	1.0 (1.0)	1.6 (5.0)	0.6 (5.1)

(standard errors in parentheses)

Using the Crosswise Model to Measure Plagiarism

```
. cwlogit plagiat1 crosswise zurich munich female bachelor semester ///
>     journals internet students proofread goodgrades, pyes(pyes) nolog
Crosswise model logistic regression              Number of obs     =        379
                                                 Nonzero outcomes  =        189
P(surrogate "yes") =  pyes                       Zero outcomes     =        190
                                                 LR chi2(11)       =      20.83
                                                 Prob > chi2       =     0.0352
Log likelihood =  -202.9246                      Pseudo R2         =     0.0488

------------------------------------------------------------------------------
    plagiat1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   crosswise |    1.90966   .5951165     3.21   0.001     .7432529    3.076067
      zurich |   1.205714   .8628404     1.40   0.162    -.4854224     2.89685
      munich |  -.2935347   .9339085    -0.31   0.753    -2.123962    1.536892
      female |   .1310311   .6306861     0.21   0.835    -1.105091    1.367153
    bachelor |   .0719657   .7070102     0.10   0.919    -1.313749     1.45768
    semester |  -.1511776   .1316926    -1.15   0.251    -.4092904    .1069352
    journals |  -.0420907   .7151018    -0.06   0.953    -1.443665    1.359483
    internet |    1.34571   2.364382     0.57   0.569    -3.288394    5.979814
    students |    1.35031   .6117542     2.21   0.027     .1512942    2.549326
   proofread |   .0769544   .7458451     0.10   0.918    -1.384875    1.538784
  goodgrades |  -.8288506   .8247797    -1.00   0.315    -2.445389    .7876879
       _cons |  -3.575383   2.508581    -1.43   0.154    -8.492112    1.341346
------------------------------------------------------------------------------
```

Conclusions

- Validity of estimates obtained using Randomized Response Technique (RRT) is questionable ("self-protective no" bias).
- Two other techniques were tested: the Item Count Technique (ICT) and the Crosswise Model. For the ICT the results are mixed. The Crosswise Model, however, worked well.
- Compared to the RRT, the Crosswise Model has several advantages:
    - A randomizing device (e.g. coins, cards, dice) is not required.
    - Lower complexity of instructions.
    - Lower cognitive burden for the respondent.
    - Overall the Crosswise Model seems better suited for application in self-administered questionnaires than RRT.
    - Most importantly, the Crosswise Model appears to generate a higher sense of protection and better evades self-protective respondent behavior (no obvious self-protective answering strategy).

Thank you for your attention!

References

- Dalton, Dan R., James C. Wimbush, and Catherine M. Daily. 1994. Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. *Personnel Psychology* 47:817–828.
- Fox, James Alan, and Paul E. Tracy. 1986. *Randomized response: A method for sensitive surveys.* London: Sage
- Raghavarao, Damaraju, and Walter T. Federer. 1979. Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 41:40–45
- Knoop, S. 2006. Plagiat per Mausklick – Das Plagiieren von Internettexten in wissenschaftlichen Hausarbeiten. Eine explorative Befragung von Studierenden und Dozenten an der WWU Münster. Magisterarbeit, Universität Münster.
- Krohn, W., C. Schlombs, and N.-C. Taubert. 2003. Plagiierte Hausarbeiten. Problemlage an der Universität Bielefeld. In: http://www.uni-bielefeld.de/Benutzer/MitarbeiterInnen/Plagiate/iug2001.html
- Sattler, S. 2007. Plagiate in Hausarbeiten: Erklärungsmodelle mit Hilfe der Rational Choice Theorie. Hamburg: Verlag Dr. Kovac.
- Tourangeau, R., and T. Yan. 2007. Sensitive questions in surveys. *Psychological Bulletin* 133: 859-883.
- Warner, S. L. 1965. Randomized-response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* 60:63–69.
- Yu, J.-W., G.-L. Tian, and M.-L. Tang. 2008. Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika* 67:251-263.