J Exp Criminol (2013) 9:91–107
DOI 10.1007/s11292-012-9162-z

Supervision mode effects in computerized delinquency surveys at school: Finnish replication of a Swiss experiment

Janne Kivivuori · Venla Salmi · Simone Walser

Published online: 15 September 2012

Abstract
Objectives This study provides a Finnish replication of a recent Swiss experiment (Walser and Killias: J Exp Criminol 8:17–28, 2012) on the supervision mode effects in computerized delinquency surveys in schools. This study supplements the Swiss study by using individual level randomization and two additional outcome variables: meta-questions of response integrity and incidence-counting heuristics.
Methods A total of 924 ninth grade students (15–16 years old) in southern Finland were randomly assigned (at the level of individuals) to supervision either by their teachers or by an external research assistant. Students then responded to an online self-report delinquency survey. Chi-square and *t* tests were used to compare prevalence levels and means.
Results In both last year and lifetime recall periods, only one offence type (unspecified theft) showed significantly different outcomes, with external supervision yielding a higher prevalence figure. For other offences, no supervision effects were found. When females and males were separately examined, limited evidence of gender-specific supervision effects emerged. Thus, females appear to report more thefts in external supervision while males report more violence in teacher supervision. No statistically significant supervision effects were found in questions probing response integrity and counting heuristics.
Conclusion Using teacher supervision in online self-report delinquency surveys does not appear to compromise the validity of the survey results. The findings thus largely corroborate the results of the earlier Swiss test. How supervision condition interacts with respondent characteristics apart from gender calls for further scrutiny.

J. Kivivuori (✉) · V. Salmi
National Research Institute of Legal Policy, POB 444, 00531 Helsinki, Finland
e-mail: janne.kivivuori@om.fi

S. Walser
Institute of Criminology, University of Zurich, Zurich, Switzerland

Keywords Internet surveys · Research methods · School-based surveys · Self-reported juvenile delinquency · Finland · Switzerland

Introduction

The self-report delinquency survey was invented in the late 1930s and early 1940s and has since become the cornerstone of juvenile delinquency research. Originally motivated by the need to transgress the boundaries of official control-based data in crime measurement, the method is today one of the most widely used criminological methods (Krohn et al. 2010; Kivivuori 2011; Kivivuori and Bernburg 2011). Since its birth, the self-report method has been connected to the extent and manner in which the state has institutional penetration to society (Kivivuori 2011: 163–164). Especially since the 1950s, the method has had a close relationship with schools as an institutional locus of data collection. This has led to legitimate concerns about external validity and other methodological shortcomings (Cernkovich et al. 1985; Hagan and McCarthy 1999).

While extensive data collection has continued in the school framework for a sustained period of time, methodological research on the various contextual aspects of school data collection has only recently been stepped up. Thus, Naplava and Oberwittler (2002) compared school data collection with home contacts and discovered that school-based data collection can have even higher external validity than home contacts, especially in lower social strata. Haen Marshall (2007) examined the various access-related obstacles that threaten to compromise the validity of school-based research. Kivivuori and Salmi (2009) examined the impact of special needs education groups for external validity. They observed that, in countries where special needs education groups exist, their inclusion is important for the external validity of self-report delinquency studies. There are also studies comparing the paper and pencil data collection mode with internet-based data collection, basically supporting the similarity of the results (Lucia et al. 2007).

One of the core issues of school-based data collection has been the supervision condition. This boils down to the simple question, who should supervise the classes where students are anonymously responding to a self-report survey? Does the identity of the supervisor matter? If it matters, why should the person of the supervisor influence the responses? At least two different types of mechanisms can be involved. First, there can be fear-based suppression mechanisms. The students could think that the answers collected by a teacher might be somehow given to the school personnel or to parents. Similarly, the external supervisor could trigger fears of detection, if the students suspect that the data are handed over to the police. If such fear effects exist, they could be alleviated by the use of online responding where digitally submitted responses probably support the feeling of anonymity. The second type of mechanism is related to the altruistic nature of responding. Thus, the person of the supervisor may impact through enhancing rather than suppressing self-reports, if he/she triggers altruistic motivations to respond.

In the Nordic area, the use of teachers as supervisors in data collection has been a popular procedure. In geographically large countries, this decision has great economical benefits as it is not necessary to send research assistants to various parts of the

nation. Since much of the prior methods research supporting teacher supervision was based on paper and pencil responding; there is a need to examine the role of supervision in computerized data collection. A recent Swiss study indicated that there is no difference between external and teacher supervision conditions in computerized data collection (Walser and Killias 2012). However, since methods are always embedded in cultural contexts relating to general trust (Kivivuori 2007), there is no certainty that Swiss results can as such be exported to other contexts. To examine whether the Swiss findings are cross-culturally valid, we decided to replicate the Swiss supervision experiment in Finland.

Prior tests

While there are multiple experimental studies contrasting Web surveys with paper-and-pencil responding, there is a relative shortage of studies comparing supervision effects in either of the two technical solutions. In what follows, key research influencing and inspiring the current study is briefly described.

In 1995, Bjarnason published a methods experiment comparing teacher and external supervision in a school-based drugs survey. Using a split-half random sample of 3,017 Icelandic students aged 16 to 20, he did not find statistically significant differences in self-reports about drug use. There were also no differences in the self-reported willingness to report illicit drug use. Bjarnason concluded that, if the students are given the chance to return the questionnaires in sealed envelopes, teacher supervision does not threaten the validity of self-report survey results (Bjarnason 1995). In the 1990s, the Bjarnason experiment was influential in supporting the decision of Finnish and Swedish researchers to use teacher supervision in national self-report indicator systems. In 2007, Finnish researchers conducted a preliminary small-scale methods test, comparing teacher and external supervision in a self-report delinquency survey (Kivivuori and Salmi 2011). In that test, the respondents were 15–16 year olds (n=482). The findings suggested that teacher and external supervision conditions produce roughly similar results. Some indications were, however, found that external supervision might solicit higher prevalence levels in some offences, most notably in drug use.[1]

Both the Bjarnason study and the Finnish methods pilot were based on paper and pencil responding. As computerized data collection is becoming increasingly popular, the question of whether the supervision condition impacts responding differentially in computerized data collection or not has emerged. Recently, a Swiss study examined the influence of supervision condition in computerized data collection (Walser and Killias 2012). Within a large survey about self-reported delinquency, 80 classes of ninth grade students (15–16 years old), with a total of 1,341 students, were randomly assigned to groups supervised by either a teacher or an external person (i.e. a senior student from the research team). The two groups did not differ significantly with respect to size, gender, age, and school level. Lifetime and last year prevalence of

[1] The study by Lucia et al. (2007) compared computerized and paper-and-pencil responding, but the computer responding mode partially involved teacher supervision (Lucia et al. 2007: 47). The "no difference" finding is thus consistent with other studies.

self-reported delinquency (11 offences and any delinquency), victimization (3 offenses and any victimization), and substance use (alcohol, several drugs, and any substance) were compared between the two supervision groups. Additionally, missing data (i.e. item non-response) was analyzed. In general, answers and missing data rates were comparable under the two conditions. Only 3 out of 57 comparisons showed significantly different outcomes; whenever differences were found, usually when teachers supervised students, prevalence rates were higher and missing data rates were smaller. While the overall basic findings supporting the feasibility of teacher supervision was the same as in the small-scale Finnish pilot, there were some differences because the Swiss study suggested that teacher supervision might solicit higher prevalence levels than external supervision. As discussed by Walser and Killias (2012), this might result if paper and pencil questionnaires offer less anonymity than online questionnaires.

In the Swiss study, no student-class information was collected due to anonymity concerns, and students could not be allocated to classes; cluster effects were therefore not controlled. Indeed, the prior studies discussed here were based on group level randomization, which drains statistical power due to cluster effects. For instance, the existing Finnish self-report delinquency surveys show that class-level clustering leads to design effects in the order of magnitude of 2 or 3. Since the Swiss study did not indicate a difference between experimental groups, a replication benefitted from a setting with more statistical power to discern differences and to avoid the Type II error in statistical inference. The Swiss study nevertheless represents a major advance as it pertains to computerized data collection. In this article, we replicate the Swiss study in a Finnish context. To elaborate on prior findings, we use individual level randomization to experimental conditions. In addition to prevalence, we examined additional outcomes such as variety and incidence of offending. We also examined self-reported nondisclosure propensity and counting heuristics in open-ended incidence reporting.

Current focus and design

The Finnish Self-Report Delinquency Study is a standardized indicator system with repeated nationally representative sweeps since 1995. Targeting ninth grade students (15–16 year olds), the system is based on self-administered paper-and-pencil responding in school classes supervised by teachers who follow a strict data collection procedure. The students respond anonymously in a classroom situation supervised by a teacher.[2] In the future, the FSRD will be based on computer-aided web interviewing, which takes place in the school computer classes. The current replication study was initiated in this context. Previously, a smaller-scale test has been conducted (Kivivuori and Salmi 2011). As that test took place within the paper-and-pencil mode, its role as a guide to future methodological development is limited. The current research adds to prior research by testing supervision effects in computerized data collection.

[2] The corresponding Swedish self-report delinquency survey system also uses teachers as data collectors (see Ring 2010). For an overview of Nordic self-report delinquency surveys, see Kivivuori 2007.

Sample

Ten schools with a total of 60 ninth grade classes (students aged 15–16) were non-randomly selected as pilot research sites. The total number of students in these classes was 1,148. The net sample was 924, yielding a response rate of 80 %. Selection of schools was non-random. The schools were selected from a large urban municipality in the Greater Helsinki area, plus two suburban municipalities.

All respondents filled in the questionnaire by means of computer-aided internet interviewing. The test questionnaire was a variant developed from, and based on, the FSRD questionnaire but including additional questions (victimization, selected individual-level personality measures, and methodological questions). The methodological questions included three questions on nondisclosure propensity (shoplifting, fighting in a public place, and marijuana/hashish use). These were addressed only to those respondents who did not report the relevant offence. For shoplifting and beating up someone, we incorporated conditionally triggered questions for respondents who reported a high number of offences committed in the preceding year (last year incidence). This question tentatively probes the question, how "separate" were the reported incidents.

Design

In each class, students were randomly assigned to teacher supervision and external supervision. Date of birth (even/uneven number) was used in randomization. We selected this procedure because it is independent of students' social relations, which influence sitting arrangements in the classroom. Partially because of this, it is also unrelated to the outcome variable (delinquency). Since the date of birth-based random split of any particular school class often falls short of dividing the class into two groups of equal size, the bigger group was allocated to the condition with fewer previous respondents.[3] Randomization took place in the classroom, when both the external supervisor and the teacher supervisor were present in the class. One group was supervised by a NRILP research assistant (external supervision condition), while the other group was supervised by the school teacher (teacher supervision condition) immediately after randomization. Thus, the respondents knew that the class had been randomized into two groups and that these were supervised by different people; otherwise, the reasons for creating two groups were not explained. Indeed, the field data collectors reported surprisingly few questions from the students concerning why the classes were split into two groups.

In many schools, two computer rooms were available, and the randomized groups could fill in the questionnaire simultaneously. If only one computer class was available, the randomized groups took turns. Finnish computer classes are typically so big that randomized groups fit in. In the class, responding instructions were read to the students before they started responding. In both groups, the instructions were

[3] Randomization was based on the students' own report about the day when they were born. This procedure means that there was no need to collect personal identification data on the respondents. In some classes, the randomization procedure (even/uneven date of birth) placed less than 5 students in the smaller group: in these classes, the students were allocated on the basis of whether their date of birth was at the beginning (days 1–15) or end (days 16–31) of month in order to avoid very small groups.

identical (written and read from the paper by the supervisor). The instructions included assurances that the responses were anonymous.

When the composition of the experimental groups was examined using available control variables,[4] no statistically significant differences emerged. The mean duration of questionnaire completion was also identical in both groups (mean and median: 15 min).

Analysis

We examine the impact of supervision condition by comparing the detected prevalence of offending in the experimental groups. We test the null hypothesis that there is no significant difference between experimental groups using the χ^2 test, with $p<.05$ as the threshold of significance. Because conducting multiple comparisons increases the likelihood of yielding significant differences by chance even if the null hypothesis is true in the population, the analysis strategy is "too sensitive" to findings which are contrary to our expectation of no effect.[5] Following the procedure of the primary study (Walser and Killias 2012), we use the Cramer's V statistic to indicate effect size; the cut-point of 0.1 is regarded as the threshold between no effect and small effect (Cohen 1988). We add to prior research by additionally comparing the behavior of incidence and variety type measures under two supervision conditions. Here, we use the t test while additionally deriving relevant p values from negative binomial regression models when count variables are used as the outcome.

Sample size considerations were based on an a priori power analysis focusing on expected differences in the prevalence of behaviors in experimental groups. Given that we use non-directional tests with 0.05 significance thresholds and measure relatively prevalent behaviors, the failure to reject the null hypothesis should be interpreted with caution (cf. Weisburd et al. 2003). To ensure maximum statistical power, we emphasize non-disaggregated comparisons. However, we additionally present findings separately for males and females because it is substantially relevant to explore whether the supervision context interacts with respondent socio-demographics.

In self-report surveys, the recall period probably influences the manner of responding. For instance, if a lifetime recall period allows more leeway for respondent interpretation, it may be more susceptible to supervision effects. We therefore present the findings separately for lifetime and 12-month recall periods.

Prevalence of delinquency by supervision mode

Table 1 shows the lifetime prevalence of delinquency by supervision condition, separately for each offence item. The main conclusion is that supervision condition appears to have no significant effect on the prevalence of delinquency when lifetime recall period was used. Only one offence, "other theft," yielded a statistically

[4] Gender, age, years lived in the present locality, parental immigration status, family structure, family economic problems, personally available spending money, GPA for mathematics, English, and Finnish.
[5] With 34 item-wise comparisons in Tables 1 and 2, using a Bonferroni correction would require a $p<.0015$ threshold for the falsification of the null hypothesis.

Table 1 Participation in offences, lifetime (%)

	External (n=469)	Teacher (n=455)	χ^2 (df=1)	Cramer's V
Graffiti drawing	31.6	29.0	.71	.028
Vandalism at school	27.9	27.7	.01	.003
Vandalism elsewhere	41.8	40.2	.24	.016
Any vandalism	*55.7*	*52.1*	*1.18*	*.036*
Shoplifting	40.5	38.5	.41	.021
Stealing at school	41.4	37.4	1.55	.041
Motor vehicle theft	1.7	2.6	.95	.032
Other theft	20.9	15.8	3.96*	.065
Breaking and entering	13.0	15.8	1.49	.040
Any theft	*63.5*	*59.3*	*1.72*	*.043*
Bullying at school	50.5	46.6	1.44	.039
Fighting (public place)	23.9	22.4	.28	.017
Beating up someone	17.5	19.1	.52	.021
Robbery	3.0	4.6	1.68	.043
Any violence	*59.3*	*55.8*	*1.13*	*.035*
Use of cannabis drugs	11.3	11.6	.03	.005
Use of other drugs	4.9	5.1	.01	.003
Any drugs use	*14.3*	*13.6*	*.08*	*.010*
Truancy	56.3	55.6	.04	.007
Drunken driving	11.1	12.1	.23	.016
Illegal downloading	80.6	77.8	1.10	.034
Any offence	*93.8*	*92.3*	*.82*	*.030*

Summary values in italics

*p<.05

significant difference, and even there the effect size, as indicated by Cramer's V, is below the small effect size (Cohen 1988) of 0.1. Otherwise, the figures were remarkably similar.

We additionally calculated variables tapping lifetime participation in any vandalism, any theft, any violence, any drugs use, and any offending. In these sum variables, committing at least one of the composite offences means that a respondent has participated in the relevant dimension of offending. There were no supervision effects on results based on these sum variables (Table 1).

When the 12-month recall period was used, the findings again showed striking similarity (Table 2). The supervision condition did not have an effect on reporting offences. Again, the sole exception was the "other theft" category, which yielded a higher prevalence level in the external supervision condition. The effect of supervision on "other theft" is not however very substantial. And again the composite variables tapping offending dimensions show no supervision effects.

Table 2 Participation in offences, last year (%)

	External (n=469)	Teacher (n=455)	χ^2 (df=1)	Cramer's V
Graffiti drawing	15.6	16.3	.08	.010
Vandalism at school	14.9	15.2	.01	.003
Vandalism elsewhere	19.4	19.8	.02	.005
Any vandalism	*29.0*	*31.6*	*.77*	*.029*
Shoplifting	14.3	13.8	.04	.006
Stealing at school	22.0	18.2	1.99	.046
Motor vehicle theft	1.1	1.3	.13	.012
Other theft	7.9	4.6	4.21*	.067
Breaking and entering	3.2	4.0	.39	.020
Any theft	*30.5*	*28.4*	*.51*	*.023*
Bullying at school	18.1	19.6	.31	.018
Fighting (public place)	9.8	10.8	.23	.016
Beating up someone	6.6	6.2	.08	.009
Robbery	1.9	3.1	1.28	.037
Any violence	*24.9*	*28.4*	*1.37*	*.039*
Use of cannabis drugs	9.6	9.7	.00	.001
Use of other drugs	3.8	4.2	.07	.009
Any drugs use	*11.3*	*11.2*	*.00*	*.001*
Truancy	42.2	40.0	.47	.02
Drunken driving	9.4	10.8	.49	.023
Illegal downloading	73.1	71.2	.43	.021
Any offence	*84.2*	*82.9*	*.31*	*.018*

Summary values in italics

*p<.05

It is possible that the supervision condition interacts with some socio-demographic features so that some respondents are more susceptible to supervision condition while others are immune to such influences. We therefore tentatively explored the possibility that respondent gender might be such a factor. To avoid an excessive number of item-specific comparisons, we based these analyses on the five offending dimension sum variables shown in Tables 1 and 2.

The findings of the gender-disaggregated analyses are shown in Tables 3 and 4 for lifetime and last year recall periods. The basic finding remains robust: few if any supervision effects are detected. However, two exceptions (both exceeding the small effect size of 0.1) emerge. First, it seems that females are more prone to report lifetime theft offending when an external supervisor is overseeing the class during data collection. If we look at the individual items (not shown), the difference is similar and at least marginally significant in four of the five constituent offences (shoplifting, stealing at school, other theft, and breaking and entering). The supervision effect on female reporting of lifetime theft participation therefore seems to us

Table 3 Participation in key offending dimensions by gender (%), lifetime

	Males				Females			
	External (n=227)	Teacher (n=234)	χ^2 (df=1)	Cramer's V	External (n=242)	Teacher (n=221)	χ^2 (df=1)	Cramer's V
Vandalism	59.5	58.1	0.09	.014	52.1	45.7	1.87	.064
Theft	64.3	65.8	0.11	.016	62.8	52.5	5.05*	.104
Violence	72.2	72.2	0.00	.000	47.1	38.5	3.52	.087
Drugs use	14.1	15.4	0.15	.018	14.5	11.8	0.74	.040
Any offence	96.5	95.3	0,40	.030	91.3	89.1	0.63	.037

*p<.05

fairly robust. In the 12-month recall period (Table 4), no significant effect was found even though the observed difference was to the same direction.

Secondly, males manifested higher last year violence participation in the presence of their own teacher than in the presence of an outside researcher (Table 4). In contrast to the finding on female theft reporting, none of the constituent violent offending items showed significant effects. However, differences observed in the current male sample were to the same direction in school bullying, fighting, beating someone up, and robbery.

To sum up thus far, it seems that there is no general or pervasive supervision effect while the reporting of specific offence types may be differentially susceptible to supervision effects in specific sub-populations.

The two supervision conditions additionally produced very similar results on the general pattern of delinquency. There was a nearly perfect concordance in how different offences rank in terms of prevalence. Spearman's rank order correlation of prevalence figures for 17 offences was .99 in the lifetime recall period and .98 in the last year recall period. In gender-specific analysis, the lowest rank order correlation was .96 (last year recall period for females; in all cases p<.01).

Table 4 Participation in key offending dimensions by gender (%), last year

	Males				Females			
	External (n=227)	Teacher (n=234)	χ^2 (df=1)	Cramer's V	External (n=242)	Teacher (n=221)	χ^2 (df=1)	Cramer's V
Vandalism	29.5	32.5	0.47	.032	28.5	30.8	0.28	.025
Theft	29.1	31.2	0.25	.023	31.8	25.3	2.37	.072
Violence	32.2	42.3	5.07*	.105	18.2	13.6	1.83	.063
Drugs use	11.9	12.4	0.03	.008	10.7	10.0	0.08	.013
Any offence	88.5	86.8	0.34	.027	80.2	78.7	0.15	.018

*p<.05

Measures of variety and incidence

Variety indicators measure the number of different offence types the respondent reports having committed (instead of counting the number of individual incidents). Due to problems in the counting and definitions of incidents (see below), variety type measures are often preferred in criminology. We calculated variety measures for lifetime and last year recall periods, including all offences shown in Table 1. The range of these variables was thus from zero to 17. Since the likelihood of reporting zero offences versus any other variety score probably reflects the effects of supervision on prevalence, which have already been examined above, we compared means by excluding the value zero.

The supervision mode did not influence overall variety scores (Table 5). Mean lifetime variety was practically identical in teacher and external supervision conditions. The same applies to mean last year variety of offending. Among females, there was a marginally significant ($p=.057$) difference in lifetime variety. Thus, it appears that external supervision may be an environment that promotes self-reports in females. Among males, there was no supervision effect.

The measurement of the number of individual offences (incidence) is more complex than the measurement of variety type constructs. When the number of last year offences is based on an open-ended response, there are always respondents who report very high figures. The problem of how youths really count incidents is examined below. Here, we calculated an overall total last year incidence measure in three steps. First, we recoded all item-specific reports exceeding 50 as 50 annual offences. Second, we added the 16 usable[6] items, resulting in an incidence variable with a potential range from zero to $16 \times 50 = 800$. Third, we defined four respondents with higher scores than 365 as missing data. We then compared incidence reports in supervision conditions. The value zero was excluded in order to bracket out prevalence effects. As can be seen from Table 5, we were unable to detect supervision effects in the number of reported annual incidents. In females, the mean incidence score was slightly higher in external supervision, while the opposite was the case among males (non-significant differences). Since all of the examined variety and incidence distributions were positively skewed, we tested the relevant differences also using a negative binomial model.

We separately inspected the presence of extremely high variety scores in the experimental conditions. The percentage of students who reported 10 or more offence types (lifetime and last year) did not differ in experimental conditions.

Effect of supervision condition on nondisclosure intent

In his 1995 methods experiment, Bjarnason included a "meta-question" about response integrity: "If you had ever tried cannabis, do you think you would say so in a survey like this?" The question was addressed to all respondents and therefore one of the response options was "I have already said so" (see also Hibell et al. 2009: 50). In the current study, we incorporated a similar meta-question for three offences:

[6] Downloading was excluded because its last year incidence question was differentially formulated.

Table 5 Mean variety and incidence of offending, by gender and supervision condition

	All		Females		Males	
	Teacher	External	Teacher	External	Teacher	External
Lifetime variety	5.0	5.1	4.1	4.7[a]	5.8	5.5
Last year variety	3.2	3.2	2.9	3.3	3.6	3.2
Last year incidence	17.7	16.4	12.5	14.2	22.1	18.6

Cell frequencies vary

[a] Marginal significance at $p=.057$ (t test), $p=.051$ (p value from negative binomial regression model)

shoplifting, fighting in a public place, and cannabis use. We utilized the computerized response form by making this question conditional. If the respondent had previously reported that he/she had *not* committed these offences, the corresponding meta-question emerged at the end of the survey. Our question formulation was, "If you had ever used marijuana or hashish, would you have said so it in this questionnaire?" with response options "yes" and "no". The questions for shoplifting and fighting were analogously formulated.

If the supervision condition would affect the students' responding behavior so that in one of the conditions some part of the non-offenders would in fact be offenders, this kind of meta-question addressed to non-offenders could in principle make supervision effects more visible. The results, however, indicate that non-disclosure intention is not influenced by the supervision condition. For all three offence types, detected supervision mode differences were non-significant for all respondents and for both genders separately (Table 6). However, among female respondents, the non-significant differences tend to show lower nondisclosure intent in the external supervision group, an observation which appears consistent with the findings reported above.

Table 6 Percentage students expressing non-disclosure intention, of self-reported non-offenders, by gender and supervision condition

	All			Males			Females		
	External	Teacher	Cramer's V	External	Teacher	Cramer's V	External	Teacher	Cramer's V
Shoplifting	4.7	5.0	.008	6.8	5.6	.024	2.7	4.5	.047
Fighting at a public place	3.9	4.5	.015	5.6	5.9	.006	2.8	3.5	.020
Use of marijuana or hashish	7.9	10.2	.039	10.6	13.5	.045	5.5	6.9	.029

Cell number varies based on the number of non-offenders. All 9 comparisons are non-significant (χ^2 -test)

Counting heuristic in incidence questions

In self-report research, the concept of incidence typically refers to the number of times a person has committed a specific crime. Thus, one respondent may have stolen twice in a year, while another respondent reports having stolen 20 times in a year. These two persons contribute equally to the prevalence of theft among the measured population, but they have a ten-fold difference in the incidence of theft. Generally, the use of open-ended frequency responding has been seen as reflecting methodological progress in the study of self-reported delinquency (Thornberry and Krohn 2000: 35). Open-ended response sets reveal frequent (high incidence) offenders. On the other hand, there may be problems in using open-ended incidence questions. Examination of response distributions indicate that students favor "round numbers" such as 5, 10, and 20, possibly because they cannot remember the exact number of incidents (Kivivuori 2007: 40–42). Students may use the frequency reporting option as a means of communicating simultaneously about frequency and severity of action (Andersson 2011: 154–155). Furthermore, it has been suggested that different respondents may use different counting units in incidence reports (Ring 1999: 77). Thus, if a student steals 10 candy bars at a single time and place, he/she can report 10 thefts or 1 theft, depending on how he/she understands or uses concepts such as "how many times" or "offence."

For two offences (shoplifting and beating up someone), we added a conditional follow-up question. If the respondent reported a high incidence figure, a further question was asked: "Were all these thefts different incidents?" The response options were "yes, they were all different incidents," "no, all thefts took place at a single time from a single shop," and "no, some of them were different but some took place at single distinct place/time." This follow-up was asked for shoplifting and (with an analogous question) beating up someone. The condition was at least 10 offences in shoplifting and at least 5 offences in beating up someone. Respondents indicating that number or more triggered the follow-up. Fourteen respondents had shoplifted more than 10 times in the year, and 12 youths reported at least 5 beatings during the year. Thus, we could not compare the experimental conditions due to small number of observations.

In shoplifting, 1 respondent out of 14 (7 %) said that all reported incidents clustered to a single occasion or target. That person thus reported a single incident of theft as multiple incidents, probably because he/she had stolen multiple items. Two persons (14 %) reported that their theft reports referred to a mix of separate incidents and items stolen at a single time. If these three respondents are regarded as one group, 21 % of youths claiming to have shoplifted at least 10 times count incidents in a manner that diverges from the standard interpretation of incidence scores. In beating up someone, there were 12 people who reported at least 5 beatings; of these, 3 said that some of the incidents took place at a single occasion. Thus, 25 % of multiple offenders may count incidents in a way that diverges from the standard interpretation of incidence counts.

With hindsight, the problem with our question was that we set the threshold of the follow-up question too high. There were few people who reported a sufficient number of offences. There is no evidence that supervision mode impacts the operational definitions of how incidents are counted by the responders. On the other hand, the findings suggest that incidence-counting heuristics are a relevant methodological question in survey based delinquency research.

Conclusion

Main findings

In this research, we replicated an earlier Swiss experiment on supervision effects in computerized delinquency surveys (Walser and Killias 2012). That study did not find supervision effects. The main result of our replication was that the prior findings were corroborated. We could find no *pervasive* supervision effects on delinquency survey responding in computerized data collection. The general thrust of the analysis was that there were few effects on prevalence, irrespective of the duration of the recall period. The same applies to variety and incidence of self-reported delinquency. There was no evidence that supervision condition affects the number of high frequency respondents. The relative ranking of offences in terms of prevalence remains the same irrespective of the supervision mode. There was no statistically significant supervision effect on non-response intention. The data quantity was too small to assess the impact of supervision on unit counting heuristics, but there is no reason to suggest that supervision could influence how respondents interpret the identity of an "incident."

At least one prior study based on paper-and-pencil responding has suggested that students may be more open to self-reports if the supervisor is not from the school (Kivivuori and Salmi 2011). The main thrust of the current findings contradicts this: few supervision effects were found. The difference may be due to methodological problems in prior group randomized studies, or in the paper-and-pencil response technique. It is possible that computer-based administration neutralizes problems that teacher supervision may have in paper-and-pencil responding. After all, in paper-and-pencil responding, the supervisor acts as an intermediary between the respondent and the researchers; in computer-based data collection, the data bypasses the supervisor in that it is directly submitted to the researchers via the internet.

In the prior Swiss study, no differences in lifetime and last year prevalence rates of self-reported delinquency, victimization, and substance use could be found for students who were supervised by a teacher versus an external person during the online interviews (Walser and Killias 2012). The present study expands these findings by showing that other aspects of delinquency (variety and incidence) are not affected by supervision mode either. Also, it appears that meta-questions about response integrity are resistant to differences in supervision condition. Given the current findings and prior research, there is no evidence that teacher presence would somehow seriously compromise the validity of anonymous self-reports.

While the lack of major supervision effects remains a core finding of the current research, some differences were observed. Interestingly, limited evidence of supervision effects emerged from the gender disaggregated analysis. Females were more prone to report thefts if the supervisor was from outside the school. This effect was found with the composite theft sum variable, and with its constituent specific items. There were also some indications that external supervisors may make females more at ease responding to self-report delinquency questions. Among females, the non-significant differences in the current sample tended to show higher prevalence levels in external supervision. This was also reflected in the higher lifetime variety score among females in outside supervision. Thus, the supervision effect among females may be general instead of theft-specific. For males, the only supervision effect now

detected was to the opposite direction: males apparently felt more at ease in reporting violence when the class was supervised by a teacher.

Social–psychological mechanisms of influence

It is one thing to observe supervision effects, and another to explain and interpret such effects. What social–psychological mechanisms might explain, for instance, gender-specific supervision effects? While this cannot be definitely answered by the current study, possible interpretations can be suggested. Maybe (theft) offending is particularly tarnishing for female reputation within the school community, making them wary of honest self-reports in the presence of their teacher. Apart from suppression effects resulting from the presence of the teacher, the finding may result if the external supervisor enhances reporting or memory retrieval. From the point of view of the "conversation" paradigm of survey responding, responding is an act of altruism towards the researchers (Kivivuori et al. 2012). The identity of the supervisor is part of the frame in which the act of responding takes place. The external supervisor may trigger altruistic reactions towards the supervisor, and perhaps more so in females than in males. In contrast, among the males, the only supervision effect involved less violence reporting in the presence of an external supervisor. The males were thus less, not more, "altruistic" towards the external supervisor. Conceivably, male responding behavior could be explained by lack of trust towards the external data collector. In extreme cases, some respondents might suspect that the external supervisor is trying to detect specific offenders.

Strengths, limitations and future research

As an improvement to prior research on supervision effects, we used individual level randomization which gave us more statistical power than group randomization. While our sample was sufficiently large to detect relatively small differences, it did not allow us to disaggregate beyond gender. With hindsight, the decision to calibrate sample size without consideration to disaggregation was a limitation. Since we detected some gender-specific effects, future studies might benefit from an option to control for what other respondent features might interact with supervision effects. The responding behavior of immigrant populations is a case in point. There is evidence that immigrant youths may under-report offending in self-report surveys (Batenburg-Eddes et al. 2012). Immigration status might therefore also interact with supervision condition.

Regarding the manner of randomization, birth date functions well in the overall study population, and it is strong in being independent of seating arrangements in the class (which in turn are influenced by assortative friendship formation). However, it can result in uneven-sized and small response groups in some of the classes. Since a very small responding group may itself introduce methods effects, other randomization means could be used instead, as long as they are not susceptible to seating arrangements or other sources of bias.

Apart from such technical limitations and challenges, some more general shortcomings warrant consideration. The notion that responding behavior can reflect altruism points to a limitation of the kind of study that compares observed prevalence levels in different methodological circumstances. If we detect supervision effects on

responding, the current study design cannot directly ascertain how this difference pertains to the truthfulness or accuracy of self-reports. Are females under-reporting in teacher supervision, or over-reporting in external supervision? Furthermore, the concepts of under-reporting and over-reporting may be misleading in this context. It may be rare that students knowingly invent offences that do not exist. Supervision effects may reflect situational factors influencing memory retrieval or subtle definitional processes regarding what behaviors are "reportable" as crimes, just like other contextual factors in survey research (Kivivuori et al. 2012). In any case, higher prevalence figures cannot be automatically equated with more truthful responding. In delinquency research, more is not necessarily better in terms of validity. In the future, experimental research might benefit from supplementary qualitative interviews with study subjects, who could explain how they perceived the actual data collecting situation.

An additional caveat is that the relevance of possible supervision effects depends on the research interest. If the purpose of the study is to study the patterns of youth crime, various administration modes produce quite similar results. Similarly, if the main purpose of data collection is to ascertain delinquency *trends*, administration mode may matter less. In Finland, the decreasing prevalence of shoplifting since the early 1990s was observed in independent measurements using teacher and external supervision (Kivivuori 2009: 84–85).

Cultural factors influence how people respond to survey questions (Johnson et al. 2005). Self-report delinquency surveys are hardly exceptions to this rule. There is a need for further international studies on the methodology of self-report delinquency surveys. Such studies could examine the cultural and social embeddedness of the validity of the self-report delinquency survey (Haen Marshall 2007; Kivivuori 2007). In this research, we contributed to internationally comparative methods research by replicating an earlier Swiss study. Had we detected differential supervision effects, these could have been mediated by cultural factors. As the findings were roughly similar, one reason for this may be the similarity of the two countries. Indeed, Finland and Switzerland are probably culturally relatively close to one another when compared with the full spectrum of countries using the self-report delinquency survey. There is reason to believe that responses to sensitive topic surveys manifest wider variation if more countries were compared. For instance, findings from the European School Survey Project on Alcohol and Other Drugs (ESPAD) indicate that specific countries manifest an above-average proportion of students expressing nondisclosure intent. For instance, Lithuania, Latvia, Croatia, Bulgaria, Greece, Ukraine, and Romania had high "unwillingness to report" scores (Hibell et al. 2009: 50). As international studies using self-report methods are likely to proliferate, there is clearly an urgent need to study the cultural variability of its applicability.

Acknowledgements We thank researcher Mikko Aaltonen and senior research analyst Reino Sirén for their assistance in various phases of the research process.

References

Andersson, L. (2011). *Mått på brott. Självdeklaration som metod att mäta brottslighet*. Kriminologiska institutionens avhandlingsserie nr 29. (Stockholm: Stockholm University)

Batenburg-Eddes, T., Butte, D., van de Looij-Jansen, P., Schietart, W., Raat, H., & de Waart, F. (2012). Measuring juvenile delinquency: how do self-reports compare with official police statistics? *European Journal of Criminology, 9*, 23–37.

Bjarnason, T. (1995). Administration mode bias in school survey on alcohol, tobacco and illicit drug use. *Addiction, 90*, 555–559.

Cernkovich, S. A., Giordano, P. G., & Pugh, M. D. (1985). Chronic offenders: the missing cases in self-report delinquency research. *Journal of Criminal Law & Criminology, 76*, 705–732.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Erlbaum.

Haen Marshall, I. (2007). "Pourquoi pas?" Versus "Absolutely not!" Cross-national differences in access to schools and pupils for survey research. *European Journal of Criminal Policy and Research, 16*, 89–109.

Hagan, J., & McCarthy, J. (1999). *Mean streets: youth crime and homelessness*. Cambridge: Cambridge University Press.

Hibell, B., Guttormsson, U., Ahlström, S., Balakireva, O., Bjarnason, T., Kokkevi, A., & Kraus, L. (2009). *The 2007 ESPAD Report. Substance use among students in 35 European countries*. Stockholm: The Swedish Council for Information on Alcohol and other Drugs (CAN).

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles. Evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36*, 264–277.

Kivivuori, J. (2007) *Delinquent behaviour in Nordic capital cities*. Scandinavian Research Council for Criminology and National Research Institute of Legal Policy, Publication 227. (Helsinki: Scandinavian Research Council for Criminology)

Kivivuori, J. (2009). Self-reported delinquency studies in Finland. In R. Zauberman (Ed.), *Self-reported crime and deviance studies in Europe. Current state of knowledge and review of use* (pp. 77–100). Brussels: Vubpress.

Kivivuori, J. (2011). *Discovery of hidden crime. Self-report delinquency surveys in criminal policy context*. Oxford: Oxford University Press.

Kivivuori, J. & Bernburg, J. G. (2011). Delinquency research in the Nordic countries. In M. Tonry & T. Lappi-Seppälä (Eds.), *Crime and justice in Scandinavia*. Crime and justice: A Review of Research, Vol. 40 (pp. 405–477). (Chicago: University of Chicago Press)

Kivivuori, J., & Salmi, V. (2009). The challenge of special needs education in school-based delinquency research. *Journal of Scandinavian Studies in Criminology and Crime Prevention, 10*, 2–17.

Kivivuori, J. & Salmi, V. (2011). Supervision mode effects in school-based delinquency and victimization survey: preliminary test. NRILP Research Brief 20/2011. Helsinki: National Research Institute of Legal Policy

Kivivuori, J., Sirén, R., & Danielsson, P. (2012). Gender framing effects in victim surveys. *European Journal of Criminology, 9*, 142–158.

Krohn, M. D., Thornberry, T. P., Gibson, C. L., & Baldwin, J. M. (2010). The development and impact of self-report measures of crime and delinquency. *Journal of Quantitative Criminology, 26*, 509–525.

Lucia, S., Herrmann, L., & Killias, M. (2007). How important are interview methods and questionnaire designs on research on self-reported juvenile delinquency? An experimental comparison of Internet vs. paper-and-pencil questionnaires and different definitions of the reference period. *Journal of Experimental Criminology, 3*, 39–64.

Naplava, T., & Oberwittler, D. (2002). Methodeneffekte bei der Messung selbstberichteter Delinquenz von männlichen Jugendlichen. *Monatsschrift für Kriminologie und Strafrechtsreform, 85*, 401–423.

Ring, J. (1999). *Hem och skola, kamrater och brott.* Stockhom: Kriminologiska institutionens avhandlingsserie nr 2. (Stockholm: Stockholms University).

Ring, J. (2010). *Brott bland ungdomar i årskurs nio. Resultat från skolundersökningen om brott åren 1995–2008. Brottsförebyggande rådet 2010:6*. Stockholm: Brottsförebyggande rådet.

Thornberry, T. P. & Krohn, M. D. (2000). The self-report method of measuring delinquency and crime. In *Measurement and analysis of criminal justice*. Criminal Justice 2000, Vol. 4. National Institute of Justice.

Walser, S., & Killias, M. (2012). Who should supervise students during self-report interviews? A controlled experiment on response behaviour in online questionnaires. *Journal of Experimental Criminology, 8*, 17–28.

Weisburd, D., Lum, C. M., & Yang, S.-M. (2003). When can we conclude that treatments or programs "don't work"? *Annals of the American Academy of Political and Social Science, 581*, 31–48.

Janne Kivivuori is Research Director and Professor at the National Research Institute ofLegal Policy, Helsinki, Finland, and Adjunct Professor of Sociology at the University ofHelsinki. His research has centred on studies of juvenile delinquency, homicide, and the methodology of crime measurement. His recent monograph Discovery of Hidden Crime presents a history of the self-report delinquency survey (Oxford University Press 2011).He has published articles in journals such as *Acta Sociologica*, *British Journal of Criminology*,*European Journal of Criminology*, and *Crime and Justice – A Review of Research*.

Venla Salmi is Researcher at the National Research Institute of Legal Policy. She has been centrally involved in the development of the Finnish Self-Report Delinquency Survey.Her other research interests include business victim surveys and domestic violence research. Her work has appeared in journals such as *European Journal of Criminology*,*Journal of Scandinavian Studies in Criminology and Crime Prevention* and *Journal of Substance Use*.

Simone Walser is a PhD student at the University of Zurich, Institute of Criminology. Her research interests include juvenile delinquency and particularly the effects of situational factors and routine activities in youth behaviour. Her recent work on the methodology of self-report delinquency questionnaires has appeared in the *Journal of Experimental Criminology*.