



Universidade do Minho
Escola de Psicologia

Hugo Miguel dos Santos Gomes

**Measuring offending: Field experiments
and improving the accuracy of self-reports
of delinquent behavior**

FCT Fundação
para a Ciência
e a Tecnologia

 **REPÚBLICA
PORTUGUESA**
CIÊNCIA, TECNOLOGIA
E ENSINO SUPERIOR



 **UNIVERSITY OF
CAMBRIDGE**
Institute of Criminology

 **FULBRIGHT**
Portugal

UF **UNIVERSITY of
FLORIDA**



Universidade do Minho
Escola de Psicologia

Hugo Miguel dos Santos Gomes

**Measuring offending: Field experiments
and improving the accuracy of self-reports
of delinquent behavior**

Doctoral Thesis
PhD in Applied Psychology

Work supervised by
Professor Ângela Maia
and
Professor David P. Farrington

July 2021

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.



Atribuição-NãoComercial-SemDerivações
CC BY-NC-ND

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

ACKNOWLEDGMENTS

"If I have seen further it is by standing on the shoulders of Giants"

Isaac Newton (1675)

This doctoral dissertation would not have been possible without the great assistance and support from the kind people I have had the pleasure to deal with during this journey.

I would like to thank my supervisors, Professor Ângela Maia, Professor David Farrington, and Professor Marvin Krohn. I could not have asked for better scientific and personal advisers. Through our time working together I was exposed to your brilliancy, work ethic, and caring for others. You have become role models that I aspire to live up to during my career.

To all my friends and colleagues from the University of Minho, the University of Cambridge, and the University of Florida, that I have had the pleasure to meet along this doctoral journey.

This doctoral dissertation was supported by the Fundação para a Ciência e a Tecnologia (FCT - SFRH/BD/122919/2016) and by the Fulbright Commission Portugal, which allowed me to become a Ph.D. visiting student at the Institute of Criminology, University of Cambridge, and a Fulbright Scholar at the Department of Sociology and Criminology & Law, University of Florida.

Above all, I would like to thank my wife Joana for her tireless support and patience during my doctoral work.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

MEASURING OFFENDING: FIELD EXPERIMENTS AND IMPROVING THE ACCURACY OF SELF-REPORTS
OF DELINQUENT BEHAVIOR

ABSTRACT

The body of knowledge on the causes and correlates of offending behavior is completely reliant on the quality of crime measures. However, methodological research on the assessment of offending behavior is very scarce. This doctoral dissertation aimed to assess the state of the art of crime measurement and to improve the accuracy of self-reports of offending (SRO). Chapter I describes a review of the advantages and limitations of the three main methodological techniques, i.e. official records, observation, and SRO. Considering the advantages of observation methods presented in this chapter, especially when applied within field experimental designs, we have carried out the systematic review presented in Chapter II. In this review, we have discussed the benefits of field experiments in the study of the etiology of offending. However, field experiments are very rarely used in the study of offending behavior, where SRO are the most widely used measurement method. In Chapter III, we have carried out a systematic review of methodological experiments testing potential sources of bias in SRO, providing relevant information to improve the accuracy of SRO. Taking into consideration the inconsistent results from methodological studies using SRO and other sensitive topics regarding the benefits of self-administration, we set out to assess the sensitivity of questions about offending behavior. In Chapter IV, we have developed a multi-dimensional assessment of question sensitivity and asked a total of 249 students to rate the sensitivity of several behavioral variables, which included offending behaviors. Results demonstrated that questions about offending behavior are perceived as highly sensitive. Further, we have included an experimental manipulation that allowed us to show that questions about offending behavior occurring over a distant time period are perceived as less sensitive than questions about recent offending. Chapter V presents two methodological experiments with a 2 (interviewer-administered vs. self-administered) × 2 (paper-and-pencil vs. computer interviews) factorial design. The first experiment was carried out in Portugal ($N = 181$), and the second was a replication study with students from a University in Florida ($N = 154$). Findings showed an increased odds of reporting offending behavior in self-administered surveys, suggesting that SRO provide more accurate estimates of offending behavior using self-administered surveys. Finally, we have included a general discussion on the main findings from this dissertation, highlighting the major contributions and implications on behavioral assessment.

Keywords: Field experiments; Modes of administration; Offending; Self-report; Sensitive questions

MEDIDAS DE CRIME: UM CONTRIBUTO PARA AS EXPERIÊNCIAS DE CAMPO E PARA A OTIMIZAÇÃO DOS AUTORRELATOS DE COMPORTAMENTO DELINQUENTE

RESUMO

O conhecimento acerca das causas do comportamento delincente está totalmente dependente da qualidade das medidas de crime. No entanto, a investigação metodológica sobre as medidas do comportamento delincente é muito limitada. A presente dissertação teve como objetivo avaliar o estado da arte da avaliação de crimes, bem como otimizar a precisão dos autorrelatos de comportamento delincente (ACD). O Capítulo I apresenta uma revisão da literatura sobre as vantagens e desvantagens das três principais técnicas de medida de crime, i.e. registos oficiais, observação e ACD. Tendo em conta as vantagens dos métodos de observação, especialmente quando aplicados em experiências de campo, realizámos a revisão sistemática apresentada no Capítulo II. Nesta revisão, discutimos os benefícios das experiências de campo no estudo da etiologia da delinquência. No entanto, estas experiências apenas raramente são utilizadas no estudo do comportamento delincente, onde os ACD são o método mais utilizado. No Capítulo III, realizámos uma revisão sistemática da literatura sobre as experiências metodológicas que testam potenciais fontes de enviesamento nos ACD, fornecendo informações relevantes para a otimização dos ACD. Tendo em conta a inconsistência entre os estudos metodológicos usando ACD e relatos de tópicos sensíveis em relação aos benefícios da autoadministração, no Capítulo IV, criámos uma avaliação da sensibilidade das questões e recrutámos 249 estudantes universitários para realizarem uma avaliação da sensibilidade dos ACD. Os resultados demonstraram que questões sobre crimes são tópicos altamente sensíveis. Adicionalmente, incluímos uma manipulação experimental que nos permitiu demonstrar que questões sobre crimes ocorridos há mais tempo são percebidas como menos sensíveis do que questões sobre crimes recentes. O Capítulo V apresenta duas experiências metodológicas com um *design* fatorial de 2 (entrevista cara-a-cara vs. autoadministração) x 2 (papel-e-lápis vs. computador). A primeira experiência foi realizada em Portugal ($N = 181$) e a segunda consiste numa replicação com estudantes de uma universidade da Flórida ($N = 154$). Os resultados destas experiências revelaram um aumento no relato de comportamentos delinquentes no formato de questionários autoadministrados, sugerindo que os ACD fornecem estimativas de crime com maior precisão em condições de autoadministração. Por fim, incluímos uma discussão geral sobre as principais conclusões desta dissertação, destacando os seus principais contributos e implicações.

Palavras-chave: Autorrelatos; Crime; Experiências de campo; Modos de administração; Questões sensíveis

TABLE OF CONTENTS

INTRODUCTION.....	1
Measures of offending behavior.....	2
Observation methods within field experiments.....	4
Self-report methodology.....	6
Sensitive questions.....	6
Modes of administration.....	8
Measurement bias in self-reports of offending.....	10
The present dissertation.....	12
References.....	15
CHAPTER I. MEASURING OFFENDING: SELF-REPORTS, OFFICIAL RECORDS, SYSTEMATIC OBSERVATION AND EXPERIMENTATION	26
Abstract.....	27
Introduction.....	28
Definition and units of measurement.....	28
Measure of crime.....	29
Comparing official records and self-reports of offending.....	32
Scaling-up factor.....	33
Criminal career research.....	34
Self-reports of offending.....	37
Alternative methods for measuring crime.....	40
Conclusions.....	42
References.....	45
CHAPTER II. FIELD EXPERIMENTS ON DISHONESTY AND STEALING: WHAT HAVE WE LEARNED IN THE LAST 40 YEARS?.....	51
Abstract.....	52
Introduction.....	53
Experimental approach.....	53
Laboratory versus field experiments.....	54
Field experiments in the study of deviance.....	55
Theoretical framework: factors influencing deviance.....	56
The present study.....	57

Methods	58
Search strategy	58
Inclusion criteria	58
Search for eligible studies.....	59
Results	62
Fraudulent/dishonest behavior.....	64
Stealing	69
Keeping money.....	74
Shoplifting.....	76
Discussion	79
Fraudulent/dishonest behavior.....	79
Stealing	80
Keeping money.....	80
Shoplifting.....	81
Benefits versus costs for the other versus costs for the self	82
Past, present, and future	82
Conclusion	83
References.....	108

CHAPTER III. MEASUREMENT BIAS IN SELF-REPORTS OF OFFENDING: A SYSTEMATIC REVIEW OF EXPERIMENTS 117

Abstract.....	118
Introduction.....	119
Methods	121
Search strategy	121
Inclusion criteria	122
Search for eligible studies.....	123
Analysis	125
Results	125
Modes of administration	132
Procedures of data collection	134
Questionnaire design	136
Discussion	137
Modes of administration	138
Procedures of data collection	139

Questionnaire design	140
Limitations	141
General conclusion	142
References	144
CHAPTER IV. HOW SENSITIVE ARE SELF-REPORTS OF OFFENDING?: THE IMPACT OF RECALL PERIODS ON QUESTION SENSITIVITY.....	151
Abstract	152
Introduction.....	153
Self-reports of sensitive questions	153
Definition of sensitive questions.....	154
Present study	155
Methods	155
Sample and study design	155
Measures.....	156
Procedures.....	156
Experimental design	157
Data analysis.....	157
Results	157
Recall periods.....	157
Discussion	158
References.....	163
CHAPTER V. THE IMPACT OF MODES OF ADMINISTRATION ON SELF-REPORTS OF OFFENDING: EVIDENCE FROM TWO METHODOLOGICAL EXPERIMENTS	165
Abstract.....	166
Introduction.....	167
Sensitive questions.....	168
Modes of administration.....	169
The impact of modes of administration on self-reports of offending.....	170
The present study	171
Experiment 1: Method	172
Experiment 1: Participants.....	172
Experiment 1: Design.....	172
Experiment 1: Instruments	172

Experiment 1: Procedure.....	173
Experiment 1: Results.....	175
Experiment 1: Descriptive analysis.....	175
Experiment 1: Modes of administration (Interview vs. Survey)	175
Experiment 1: Modes of data collection (Paper-and-pencil vs. Computer-assisted).....	176
Experiment 1: Interaction effects	179
Experiment 1: Discussion.....	179
Experiment 1: Modes of administration (Interview vs. Survey)	179
Experiment 1: Modes of data collection (Paper-and-pencil vs. Computer-assisted).....	180
Experiment 1: Interaction effects	181
Experiment 2: Method	182
Experiment 2: Participants.....	182
Experiment 2: Design, questionnaire, and procedure.....	182
Experiment 2: Results.....	182
Experiment 2: Descriptive analysis.....	182
Experiment 2: Modes of administration (Interview vs. Survey)	183
Experiment 2: Modes of data collection (Paper-and-pencil vs. Computer-assisted).....	186
Experiment 2: Interaction effects	186
Experiment 2: Discussion.....	186
Experiment 2: Modes of administration (Interview vs. Survey)	187
Experiment 2: Modes of data collection (Paper-and-pencil vs. Computer-assisted).....	187
Experiment 2: Interaction effects	188
General discussion	188
Limitations	192
Conclusions	192
References.....	194
INTEGRATIVE DISCUSSION.....	200
Major contributions	201
Measures of offending behavior.....	201
Observation methods within field experiments	203
Self-reports of offending behavior.....	204
Implications	208
Strengths and limitations.....	211
Future studies	213

Conclusions	216
References.....	217
APPENDIX.....	224

LIST OF ABBREVIATIONS

ACASI - Audio Computer-Assisted Self-Interview
ANOVA - Analysis of Variance
BPL - Bogus Pipeline
CAPI - Computer-Assisted Personal Interview
CASI - Computer-Assisted Self-Interview
CCTV - Closed-Circuit Television
CMA - Comprehensive Meta-Analysis
CSDD - Cambridge Study in Delinquent Development
CSR - Corporate Social Responsibility
EAS - Electronic Article Surveillance
IRR - Incidence Rate Ratio
ISRD - International Self-Report Delinquency
OJJDP - Office of Juvenile Justice and Delinquency Prevention
OR - Odds Ratio
PAPI - Paper-And-Pencil Personal Interview
PGS - Pittsburgh Girls Study
PI - Personal Interview
PVMs - Public View Monitors
PYS - Pittsburgh Youth Study
SAQ – Self-Administered Questionnaires
SDRS - Socially Desirable Response Set
SEU - Subjective Expected Utility
SRO - Self-Reports of Offending
SSDP - Seattle Social Development Project
TACASI - Telephone Audio Computer-Assisted Self-Interview

LIST OF FIGURES

Figure 1. The prevalence of crimes in England and Wales according to different sources	31
Figure 2. Prevalence and frequency of offending according to different sources	37
Figure 3. Flowchart of the systematic search processes	61
Figure 4. Flowchart of the process of systematic search	124
Figure 5. Average scores of sensitivity for offending items by recall period	160
Figure 6. Interaction effects of modes of administration and modes of data collection on lifetime offending variety	181
Figure 7. The effect of modes of administration (left) and modes of data collection (right) on overall offending variety	190

LIST OF TABLES

Table 1. Descriptive information on 60 studies in the systematic review	63
Table 2. Summary of field experiments in the Fraudulent/ dishonest behavior category.	85
Table 3. Summary of field experiments in the Stealing category	93
Table 4. Summary of field experiments in the Keeping money category.....	100
Table 5. Summary of field experiments in the Shoplifting category	104
Table 6. Descriptive information on studies in the systematic review	127
Table 7. Main findings of experiments in the systematic review	130
Table 8. Average question sensitivity of behavioral items	158
Table 9. Mean comparisons of question sensitivity by recall period	159
Table 10. Average question sensitivity of behavioral items for the American pilot study	162
Table 11. Demographic characteristics by experimental manipulations (experiment 1).....	177
Table 12. Experiment 1: Prevalence of offending and variety by modes of administration (left) and by modes of data collection (right)	178
Table 13. Demographic characteristics by experimental manipulations (experiment 2).....	184
Table 14. Experiment 2: Prevalence of offending and variety by modes of administration (left) and by modes of data collection (right)	185

INTRODUCTION

The study of the causes and correlates of offending has generated a large body of knowledge about the etiology of criminal behavior. From a developmental and life-course perspective, the acquired knowledge about the patterns of offending, risk and protective factors, as well as the effect of life events, allowed a comprehensive theoretical understanding of the development of offending (e.g., Farrington, 2005; Farrington et al., 2019; Gibson & Krohn, 2012; Moffitt, 1993). This knowledge allows the prediction of future offending and plays a major role in the development of early prevention strategies and effective interventions (e.g., Fagan et al., 2019; Farrington, 2021; Farrington & Coid, 2003; Rijo et al., 2020; Zara & Farrington, 2016).

However, knowledge about the development of offending behavior is completely reliant on the quality of crime measures. Inaccurate or biased measures of offending behavior will inevitably result in misleading conclusions about the predictors and patterns of offending and, in turn, result in poor policies and interventions (Livingston, 2013; Pepper & Petrie, 2003). This makes it very important for researchers to use the best possible practices for measuring offending behavior. Nevertheless, the assessment of criminal behavior is particularly demanding and there is a ceiling to the accuracy of crime estimates (Krohn et al., 2012; Sullivan & McGloin, 2014). Offending behavior is not only a sensitive and socially undesirable matter, it involves illegal practices that are punishable by law and people naturally try to conceal it. All these aspects of offending add to the already challenging task of assessing human behavior, making crime measurement an inherently difficult task (Osgood et al., 2002).

Measures of offending behavior

In the present dissertation, we started by asking a fundamental research question. “What are the main measures of offending behavior?” In order to provide an answer to this question, we have carried out a review of the literature on the major crime measurement methodologies, reviewing their advantages and limitations (Gomes et al., 2018). In this review of the literature, presented in Chapter I, we concluded that there are three main methodological techniques of crime assessment. First, official records, which consist of the consultation of officially recorded information by the police, prisons, and/or the courts regarding the practice of crimes. Second, researchers may use direct and indirect observation techniques to assess offending behavior. Third, self-reports of offending (SRO), where people are asked whether they have practiced several types of offenses (Maxfield & Babbie, 2009). Because observation techniques are very difficult to implement, official records and SRO are the two most widely used measurement methods in the study of criminal behavior (Piquero et al., 2014). However, there is considerable controversy about the best measures of crime, as well as the best conditions in which to collect such data.

For many years, research on criminal behavior relied mostly on data obtained from official records (Thornberry & Krohn, 2000). However, many researchers criticized this methodology, mainly because official records seriously underreported the true amount of offending behavior (e.g., Murphy et al., 1946) and because the obtained criminal data varied depending on whether the officially recorded information was provided by the police, courts, and/or prisons, which could lead to completely different conclusions (Sellin, 1931). Farrington and Jolliffe (2004) made the similar observation that only part of the total crimes committed are reported to the police, from which only a part are recorded by the police, out of which only a fraction result in convictions, and so on in a successive funneling process. In this discussion regarding the accuracy of crime measurements provided by different records (i.e., police, judicial, or penal statistics), Sellin (1931, p. 346) made a very important observation that “the value of a crime rate for index purposes decreases as the distance from the crime itself in terms of procedure increases”.

If we apply the ‘Sellin’s dictum’ onto the broader aspects of crime measurement, observation of offending behavior may be regarded as the most valuable assessment, where the behavior is assessed directly without any funneling or other biasing aspects described above. In our review (Gomes et al., 2018), we identified some studies using direct field observations to assess offending behavior, such as shoplifting (e.g., Buckle & Farrington, 1984, 1994). Others used indirect field observation methods to assess offending by creating opportunities for people to steal coins left in telephone booths (e.g., Bickman, 1971; Franklin, 1973) or money from apparently ‘lost letters/wallets’ (e.g., Farrington & Knight, 1979, 1980; Hornstein et al., 1968; Merritt & Fowler, 1948). However, field observation of offending behavior is a challenging task, mainly because offending is unpredictable and offenders actively try for their offenses not to be observed (Buckle & Farrington, 1984; Gomes et al., 2018). For all these reasons, studies using observation methods to test hypotheses relating to the causes of offending behavior are very scarce (Farrington et al., 2020).

At the same time, the limitations of the criminal data provided by official records led researchers to apply the self-report technique to assess offending behavior. In 1943, Porterfield published the first study using the self-report methodology to measure delinquent behavior. But it was Short and Nye’s work on SRO across socio-economic status that fully displayed the potential of the self-report technique in etiological studies (Nye et al., 1958; Short & Nye, 1958), and which revolutionized researchers’ opinion on the utility and feasibility of SRO (Thornberry & Krohn, 2000). Following Short and Nye’s work, SRO became more and more used and, in the next decade, Hirschi (1969) developed a highly influential study on the etiology of delinquent behavior solely based on the self-report methodology.

Still, many researchers continued to cast doubt about the ability of respondents to provide useful information regarding their own criminal behavior through self-reports (e.g., Gibbons, 1979). This motivated a large body of research on the psychometric qualities of SRO that still stands until today (e.g., Ahonen et al., 2020; Auty et al., 2015; Farrington, 1973; Farrington et al., 2014; Gold, 1966; Hindelang et al., 1981; Huizinga & Elliott, 1986; Jolliffe et al., 2003; Kazemian & Farrington, 2005; Piquero et al., 2014; Yan & Cantor, 2019). These studies repeatedly showed SRO as a valid and reliable measure of delinquent behavior, making self-reports one of the most used measurement methods in the contemporary study of criminal behavior (Jolliffe et al., 2003). The gradual improvement and the widespread application of SRO completely revolutionized our knowledge about delinquent behavior (Thornberry & Krohn, 2000). From being regarded as a taboo topic by early scholars, self-reports came a long way into being considered “the most significant methodological innovation to date in our pursuit of understanding criminal behavior” (Krohn et al., 2012, p. 23).

The literature reviewed in Chapter I (Gomes et al., 2018) shows that the validity of crime measures is bounded by a definite ceiling, and that perfect assessment of offending behavior is beyond the reach of contemporary measurement methods (e.g., Krohn et al., 2012; Sullivan & McGloin, 2014). Each methodology presents its own set of advantages and limitations, whereby a mixed-methods approach might result in the best assessment of the offending phenomenon. Nevertheless, researchers and practitioners must consider the specific qualities of each measurement technique and select the method(s) that best fit their research questions (for a discussion see Gomes et al., 2018).

Observation methods within field experiments

According to the literature included in our review of offending measures (Gomes et al., 2018), observation techniques provide the most valid information. Observation is the data source closest to the actual offending behavior, which eliminates many potential biasing factors. Through observations, researchers are able to assess the behavior of participants in the real world without them being aware that their behavior is being assessed. These characteristics are very important because they make it possible to test cause-and-effect relationships within naturalistic field experiments (Farrington, 1979).

Field experiments combine the benefits of the experimental design and the external validity of testing hypotheses in the real world. Contrary to the cross-sectional and longitudinal studies, the experimental design makes it possible to test cause-and-effect relationships through the manipulation of variables under strictly controlled conditions (Christensen, 1985; Zimny, 1961). This makes experiments crucial for the development of scientific knowledge because they provide unambiguous conclusions about

the variables affecting human behavior. On the other hand, field experiments overcome the limitations of the artificiality of laboratory experiments where participants are aware that their behavior is being scrutinized (Farrington, 1980; Harrison & List, 2004). In the laboratory, the research setting may influence participants' behavior in multiple ways (e.g., social desirability), which compromises its internal validity (Levitt & List, 2007). In the particular case of offending and deviant behaviors, this concern is especially relevant because people naturally try to conceal undesirable behaviors (Gomes et al., 2018). Considering these limitations, naturalistic field experiments provide the greatest internal and external validity (Farrington, 1979).

In 1979, Farrington carried out a review of field experiments on deviance with special reference to dishonesty. This review included field experiments using multiple techniques to observe unaware participants acting dishonestly. For example, researchers left apparently 'lost' coins and observe whether or not members of the public dishonestly claimed them (e.g., Farrington & Kidd, 1977; Feldman, 1968; Korte & Kerr, 1975). Some experiments included in this review were able to actually observe offending behavior, such as theft (e.g., Diener et al., 1976; Steinberg et al., 1977). Faced with the scarcity of this robust design, Farrington (1979, p. 242) concluded by expressing his hope "that psychologists will have the ingenuity, determination, and social responsibility to meet the challenge of experiments on deviance".

Despite the benefits of observation of offending in real-life settings, especially when applied in experimental designs, most research on the causes of offending behavior is nonexperimental and field experiments are rare in social science (Franzen & Pointner, 2013; Gomes et al., 2018). However, multiple naturalistic field experiments have been conducted by behavioral economists (e.g., Harrison & List, 2004; Levitt & List, 2009). Several of these real-world experiments use field observations that are very relevant to the study of offending (Farrington et al., 2020), such as stealing and monetary dishonesty (for a review see Rosenbaum et al., 2014). Kerschbamer et al. (2016), for example, used computers with prearranged defects to study fraud in the computer repair price. Cohn et al. (2019) studied civic honesty in 40 countries by using apparently 'lost' wallets, providing the opportunity to members of the public to steal. Balafoutas et al. (2013) resorted to GPS data to test the dishonest behavior of taxi drivers by comparing the chosen route to the estimated correct fare.

In order to provide a review of the field methods used to assess participants' deviant and dishonest behavior in the real world, we have carried out a systematic review of field experiments seeking to study the causes of offending or monetary dishonesty that have been reported since the review of Farrington (1979). This systematic review, presented in Chapter II (Gomes et al., 2021a), illustrates the potential of field experiments to study the causes of offending and dishonest behavior in the real world,

which we hope will inform and motivate more researchers to apply such methods in the study of the causes of offending. However, the field experimental design is still very rarely used in the study of offending behavior, which is dominated by the self-report methodology.

Self-report methodology

SRO are the most widely used method of measuring criminal and deviant behavior (Gomes et al., 2018). However, despite the large effort to establish the validity of SRO, especially comparing data obtained using self-reports to official records (e.g., Clark & Tifft, 1966; Hardt & Peterson-Hardt, 1977; Kulik et al., 1968; Schore et al., 1979), much less attention has been given to the study of measurement biases and cognitive processes associated with the disclosure of offending behavior. Survey researchers, on the other hand, have developed a large body of knowledge on the processes underlying survey responses and how questions shape participants' answers (e.g., Schwarz, 1999). Multiple cognitive processes are involved in providing information about one's own behavior. Prior to providing an accurate estimation, survey respondents have to comprehend the question, recall relevant information, and compute a judgment through adding, averaging, and combining behavioral information (Schwarz, 1999; Tourangeau et al., 2000). Measurement error may occur in all of these processes. Asking questions about sensitive behaviors adds a further layer of potential bias because respondents may deliberately edit their answers in order to avoid disclosing socially undesirable information (Bradburn et al., 1979; Sudman & Bradburn, 1974; Tourangeau & Yan, 2007).

Sensitive questions

Over the past decades, researchers have used self-report questionnaires to study increasingly sensitive topics (Tourangeau & Yan, 2007). Tourangeau and colleagues (Tourangeau et al., 2000; Tourangeau & Yan, 2007) provided a three-dimensional definition of question sensitivity (i.e., intrusiveness, threat of disclosure, and social desirability). First, intrusiveness refers to questions that are themselves an invasion of privacy. Respondents may feel that these questions are inappropriate and none of the researcher's business, whether or not the respondents have themselves engaged in such behavior. For example, respondents may feel that a question about stealing is an invasion of privacy, regardless if they have ever stolen something. Threat of disclosure, on the other hand, refers to the respondent's concern about their truthful answers becoming known to a third party. In this case, the question's sensitivity is dependent on the respondent's previous behavior. A question about stealing, for example,

poses no threat of disclosure for someone who has never engaged in such illegal practice. However, respondents who have stolen in the past may fear potential consequences if their honest answers become known by their employer, their parents, etc. Third, social desirability reflects the extent to which a question elicits socially desirable answers. Considering that stealing is a socially undesirable behavior, a question about stealing may be regarded as sensitive because the socially desirable answer would be to deny this practice.

These specific features of sensitive questions may compromise response accuracy by decreasing the likelihood of participants providing truthful answers to questions about sensitive behaviors (Tourangeau & Yan, 2007). In fact, evidence suggests that much of the misreporting found in self-reports of sensitive topics is a consequence of a motivated process of respondents editing their answers (Tourangeau & Yan, 2007). According to the motivated misreporting hypothesis, respondents who have engaged in socially undesirable behaviors will deliberately edit their responses in a socially desirable way in order to provide a positive image of themselves (Sudman & Bradburn, 1974; Tourangeau et al., 2000). Further, as the topics of the questions become more sensitive, the respondents' motivation to edit their answers increases, progressively compromising response quality (Tourangeau & Yan, 2007).

One of the most replicated effects of asking sensitive questions is the tendency of respondents to systematically underreport socially undesirable behaviors (Krumpal, 2013; Tourangeau et al., 2000). Methodological experiments have provided evidence that survey respondents underreport sensitive behaviors such as food intake (e.g., Wehling & Lusher, 2019), risky sexual behaviors (e.g., Giguère et al., 2019), substance use such as cigarettes (Liber & Warner, 2018), alcohol (e.g., Kabashi et al., 2019; Littlefield et al., 2017; Vinikoor et al., 2018), and other drugs (e.g., Druckman et al., 2015; Gerdtz et al., 2020; Kirtadze et al., 2018; Palamar et al., 2021), as well as deviant and criminal behaviors (e.g., Clark & Tiffit, 1966; Wolter & Laier, 2014). Further, and in accordance with the motivated misreporting hypothesis, Hser (1997) found that underreporting is more evident for highly sensitive drugs (e.g., cocaine and opiates) than for less sensitive drugs (e.g., marijuana).

In trying to circumvent the tendency to underreport socially undesirable behaviors, survey researchers have implemented data collection strategies to improve participants' willingness to report sensitive information. The bogus pipeline, for example, consists of attaching a device to the participants that they believe can detect false reports. This technique results in an increased rate of self-reported sensitive behavior (e.g., Strang & Peterson, 2020). Similarly, randomized response techniques such as the item count technique (e.g., Wolter & Laier, 2014) or the unmatched count technique (e.g., Dalton et al., 1994), where participants' reports of behavior are indirectly estimated without asking them to explicitly

reveal their sensitive behavior, consistently result in higher rates of disclosure than traditional direct self-reports (Druckman et al., 2015; Kirtadze et al., 2018).

The systematic tendency of respondents to underestimate the true prevalence of sensitive behaviors, as well as the consistently higher rates of sensitive behavior obtained in conditions where the threat of disclosure is reduced (i.e., randomized response techniques) and honesty requests are heightened (i.e., bogus pipeline) cannot be explained by chance. Further, if these effects resulted from comprehension or memory faults, the response errors would be expected to be found in both directions (i.e., over and underreports). However, inaccurate responding occurs systematically in the socially desirable direction. These findings are solid evidence that respondents to sensitive questions deliberately edit their answers (Bradburn et al., 1979; Tourangeau et al., 2000).

Taking into account the tendency of respondents to underreport the true amount of sensitive behaviors, survey researchers often use the 'more is better' assumption to determine which research method provides the most accurate reports. Even though this is just an assumption and researchers should use an external criterion for self-reported information whenever possible (e.g., biomarkers of drug use), the 'more is better' assumption is very useful in the study of behaviors where no gold standard can be applied, such as offending behavior. Using this assumption, survey researchers are able to experimentally compare different methods, such as different modes of administration. The modalities that result in higher reporting rates of socially undesirable behavior are assumed to be the most likely to yield accurate results (Tourangeau & Yan, 2007).

Modes of administration

Modes of administration are key fundamental features of the self-report methodology that can have a substantial impact on the quality of behavioral reports (Richman et al., 1999; Tourangeau & Yan, 2007). Survey information may be collected using very different types of modes of administration. Two of the most relevant variables in administration modalities are 1. whether or not respondents provide their answers to an interviewer (i.e., self-administration); and 2. whether the questions are presented on a piece of paper or on a computer. The combination of these variables provides four modes of administration that are the most typically used in behavioral assessment, i.e., paper-and-pencil personal interviews (PAPI), computer-assisted personal interviews (CAPI), paper-and-pencil self-administered questionnaires (SAQ), and computer-assisted self-administered interviews (CASI) (Thornberry & Krohn, 2000).

Methodological research shows that the self-administration of surveys significantly affects participants' responses to sensitive questions (Sudman & Bradburn, 1974). Experimental studies comparing interviewer-administered and self-administered questionnaires consistently find higher rates of admissions of socially undesirable behaviors in self-administered conditions (e.g., Aquilino, 1994; Butler et al., 2009; Jobe et al., 1997; Kreuter et al., 2008; Lee et al., 2019; Robertson et al., 2018; Schober et al., 1992; Turner et al., 1992). Tourangeau and Yan (in press) reviewed seven methodological experiments (54 effect sizes) on the effect of modes of administration on self-reports of illicit drug use and estimated that self-administration caused an increase of about 30% in drug use admissions. The findings of mode effects in reporting sensitive information are consistent with the motivated misreporting hypothesis. Face-to-face interviews require participants to verbally disclose socially undesirable information to a third person. Under self-administered conditions, respondents provide their answers directly on a piece of paper or on the computer, removing the interviewer from the data collection process and mitigating the concerns with self-image. In turn, self-administration of surveys provides an increased perception of confidentiality and anonymity which results in an increased willingness to provide socially undesirable information (Schwarz et al., 1991; Sudman & Bradburn, 1974; Tourangeau & Yan, 2007).

Furthermore, the benefits of self-administration tend to be higher for more sensitive topics (Tourangeau et al., 2000; Tourangeau & McNeeley, 2003; Tourangeau & Yan, in press). Methodological experiments testing the effects of self-administration on reports of illicit drug use typically find that the mode effect is larger for reports of cocaine than for marijuana use (e.g., Aquilino, 1994; Schober et al., 1992; Turner et al., 1992). Similarly, Richman et al. (1999) carried out a meta-analysis with 61 methodological experiments (673 effect sizes) and found evidence that self-administration causes an increase in the likelihood of participants reporting sensitive behaviors (e.g., illegal drug use, risky sexual behavior, etc.), while for low sensitivity topics such as job satisfaction and personality scales reports remained similar through the different modes of administration.

In line with these findings, authors such as Bradburn et al. (2004) have suggested that the disclosure of socially undesirable information regarding current behavior is more threatening than disclosing behavioral information that may have occurred in a distant past. In fact, there is evidence that the benefits of self-administration tend to be higher when asking questions about recent behavior compared with questions about behavior that may have occurred in the distant past (Tourangeau et al., 2000; Tourangeau & McNeeley, 2003; Tourangeau & Yan, in press). In their experiments, Turner et al. (1992), as well as Schober et al. (1992), included questions about illicit drug use over the lifetime, the previous year, and the previous month. In these experiments, the benefits of self-administration over

interviewer-administered modes were almost nonexistent for questions with lifetime recall periods, higher for reports during the past year, and highest over the previous month. In sum, these findings show that mode effects tend to increase with topic sensitivity, and survey researchers dealing with sensitive topics, such as offending behavior, must take these aspects into account.

On the other hand, with the technological developments seen in the last decades, self-reports are gradually transitioning from paper-and-pencil to computer-assisted modes. Therefore, survey researchers are interested in understanding whether or not computerization impacts participants' reports. Further, existing evidence that suggests different levels of perceived anonymity and confidentiality between paper-and-pencil and computer-assisted surveys (e.g., Denniston et al., 2010; Trau et al., 2013), causes concern that the computerization of surveys will impact reports of sensitive behavior. However, research on this specific topic has provided inconsistent results. While some survey methodologists have found evidence of an increased rate of disclosure of sensitive behaviors in paper-and-pencil questionnaires (e.g., Beebe et al., 1998), others have found higher reports in computer-assisted modes (e.g., Brener et al., 2006), while others have found no evidence of computers having an impact on participant reports (e.g., Bates & Cox, 2008; Beebe et al., 2006; Knapp & Kirk, 2003). Dodou and de Winter (2014) carried out a meta-analysis (51 studies and 62 effect sizes) to compare social desirability levels in paper-and-pencil and computer-assisted surveys, finding no evidence of differences by mode. However, Richman et al. (1999) found meta-analytical evidence that self-administered surveys using computers resulted in higher rates of reporting sensitive information than using paper-and-pencil questionnaires. Similarly, Gnambas and Kaspar (2015) reviewed experimental comparisons of self-administered modes using computers and paper-and-pencil surveys (39 studies and 460 effect sizes) and found that computer-assisted conditions caused an increased likelihood of reporting sensitive behavior.

Measurement bias in self-reports of offending

The literature on sensitive questions reviewed above provides a perspective on the main issues of collecting sensitive information using the self-report methodology. These findings converge in an accumulated body of knowledge about the processes involved in disclosing information and the measurement techniques that can be implemented in order to improve the quality of the reported data (Schwarz, 1999). The accumulated knowledge on how to ask sensitive questions can and should be considered by researchers in the study of offending behavior. However, the knowledge about asking sensitive questions must be considered with caution by offending researchers because it is mostly based on reports of behavioral information such as sexual behavior and substance use (Tourangeau & Yan,

2007). There is a possibility that these findings may not be directly transferable to reports of all types of offenses (Kleck & Roberts, 2012), making it very important for researchers to carry methodological experiments using questionnaires on offending behavior.

The need to understand the specific measurement biases affecting participants' reports of offending behavior led us to the development of a systematic review of the methodological experiments studying potential sources of bias in collecting data on SRO, presented in Chapter III (Gomes et al., 2019). In this systematic review, the comparison between self-administered surveys using paper-and-pencil and computer-assisted modes was the most replicated manipulation ($k = 10$). Despite the somewhat inconsistent findings, the overall effect showed that data collection using computer-assisted modes resulted in an increased rate of disclosing offending behavior (Gomes et al., 2019). The present results for SRO are consistent with the previous findings for sensitive topics in general (Gnambs & Kaspar, 2015; Richman et al., 1999).

On the other hand, in our systematic review (Gomes et al., 2019), only three studies compared rates of SRO obtained in face-to-face interviews and self-administered surveys (i.e., Hindelang et al., 1981; Krohn et al., 1974; Potdar & Koenig, 2005). Contrary to the solid evidence on sensitive questions (e.g., Tourangeau & Yan, 2007), these studies found no evidence that self-administration impacts participants' reports of offending behavior. Further, Hindelang et al. (1981) concluded that question sensitivity is not an important factor and that "we have no evidence that respondents find efforts to measure their delinquent behavior particularly threatening" (Hindelang et al., 1981, p. 124). As a result, Hindelang et al.'s (1981) highly influential book seemed to have established the validity of SRO once and for all, and methodological research on the best practices of asking questions about offending behavior decreased considerably (Jolliffe & Farrington, 2014).

Hindelang et al.'s (1981) conclusion about the apparent low threat of SRO seems to oppose the literature of sensitive questions reviewed above. SRO perfectly falls under the definition of sensitive topics; questions about offending behavior can be seen as too personal, respondents might fear potential embarrassing or incriminating consequences if their reports become known outside of the research study, and offending is generally regarded as socially undesirable. In order to clarify this issue, we have carried out the study presented in Chapter IV that assesses the sensitivity of questions about offending behavior (Gomes et al., 2021b). In this study, we have developed a measure of item sensitivity based on the three-dimensional definition of sensitive questions of Tourangeau et al. (2000). Evidence found in our study showed that the majority of questions about offending behavior were rated as more sensitive than a question about sexual behavior. Notably, participants rated questions about serious and violent offenses

as highly sensitive. Therefore, similarly to self-reports of other types of sensitive questions, SRO are expected to be affected by reporting bias, such as self-administration. Taking into consideration that failing to identify the biasing effects of modes of administration in SRO can result in flawed estimates and misleading conclusions about the variables affecting offending behavior, more research using questionnaires of offending is clearly needed. In Chapter V we present the last study of this dissertation (Gomes et al., 2021c), where we have tested the impact of mode effects on participants' willingness to disclose offending behavior.

The present dissertation

The main purpose of this dissertation was to provide evidence on the best measurement techniques for offending behavior, as well as to improve our knowledge about the best practices in collecting information about offending. We started by reviewing the main methodologies implemented in crime measurement. Considering the existing evidence on offending measures, we aimed to provide a systematic review of the field observation methods that can be used in real-world settings to assess deviant and dishonest behavior. Further, we aimed to systematically review the main measurement biases in SRO in order to improve self-reported data. In view of the lack of evidence on mode effects on SRO, we aimed to pre-test the sensitivity of offending questions and explored the impact of modes of administration on SRO. In order to reach these objectives, we have carried out a set of studies presented in the five chapters composing this dissertation.

In Chapter I, we started by asking a general but fundamental research question. "What are the main measures of offending behavior?" In order to address this research question, we have carried out a review of the literature about the main methodologies implemented in crime measurement (Gomes et al., 2018). Our review of the literature resulted in three main methodological techniques of crime assessment: official records, observation methods, and SRO. In this chapter, we describe the main characteristics, advantages, limitations, and implications of these measurement techniques. In sum, this review found that official records of crime are deeply biased measures of offending. Observation techniques provide the most valid data, especially when applied in naturalistic field experiments, but their implementation is complex and very rarely used. Finally, despite their limitations, SRO are easily used and provide generally valid estimations of offending behavior, making self-reports the most widely used measurement method in the study of offending.

Taking into account the findings reviewed in the first chapter regarding the benefits of observation methods within field experiments, in Chapter II we addressed a second research question. "What are the

main observation methodologies employed in field experiments to study offending behavior?” Chapter II is a systematic review of field experiments studying stealing or monetary dishonesty reported since 1979 (Gomes et al., 2021a). In this chapter, we reviewed the field observation methods and main results of a total of 60 field experiments, grouped into four categories: Fraudulent/ dishonest behavior, Stealing, Keeping money, and Shoplifting. This study provides an extensive review of the field methods used to assess participants’ deviant and dishonest behavior in the real world. Further, this review highlights the potential of field experiments to study the causes of offending.

Chapter III takes into consideration the main issues of collecting offending information using the self-report methodology reviewed in the first chapter, attempting to provide an answer to our third research question. “What are the measurement biases in SRO?” Chapter III is a systematic review of the methodological experiments studying potential sources of bias in collecting offending data using SRO (Gomes et al., 2019). In order to provide easily accessible information regarding the multiple measurement manipulations, we have used meta-analytical techniques to estimate and synthesize information. This review resulted in a total of 21 methodological experiments, testing 18 different manipulations (33 independent effect sizes), which were grouped into Modes of administration, Procedures of data collection, and Questionnaire design.

In Chapter IV, we considered two research questions derived from the previous chapter. “How sensitive are questions about offending behavior?”; and “Does recall periods impact respondents’ perceptions of question sensitivity?” In this study (Gomes et al., 2021b), we developed a three-dimensional assessment of sensitivity, which allows the assessment and ranking of the sensitivity of each offending question. In doing so, we were able to experimentally explore the impact of different time frames (i.e., lifetime, past-year, and past-month) on participants’ perceptions of question sensitivity. The sample was composed of 249 university students and the offending questions were drawn from the International Self-Report Delinquency 3 questionnaire (ISRD3; Enzmann et al., 2018; Portuguese version by Martins et al., 2015). This study allowed an evaluation and ranking of the sensitivity of offending questions. Additionally, to the extent of our knowledge, this is the only experiment testing the impact of recall periods on question sensitivity. The findings in Chapter IV allow future methodological research to control for the effect of question sensitivity in questionnaires of offending behavior.

Finally, in Chapter V we addressed a research question emerging from the contrasting results of the impact of modes of administration on sensitive questions and SRO (Gomes et al., 2021c). “Does modes of administration impact reports of offending behavior?” In this study, we have developed two methodological experiments; the first was carried out in Portugal with 181 University students, and the

second was a replication study carried out in Florida with 154 University students. The experiments presented a 2 (modes of administration: interviewer-administered *vs.* self-administered) \times 2 (modes of data collection: paper-and-pencil *vs.* computer interviews) factorial design. Based on our literature review, we set out to test two main hypotheses: participants in the self-administered modes would report higher rates of offending behavior than participants in face-to-face interviews (Hypothesis 1); participants in computer-assisted modes would report higher rates of offending compared to participants assigned to the paper-and-pencil modes (Hypothesis 2). In this chapter, we discuss our findings regarding the impact of self-administration and computerization of surveys on SRO.

In the General discussion chapter, we discuss the main findings and conclusions resulting from this dissertation, its theoretical and practical implications on behavioral assessment, the limitations of our studies, and suggestions for future research.

References

- Ahonen, L., FitzGerald, D., Klingensmith, K., & Farrington, D.P. (2020). Criminal career duration: Predictability from self-reports and official records. *Criminal Behaviour and Mental Health, 30*(4), 172-182. <https://doi.org/10.1002/cbm.2152>
- Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use: A field experiment. *Public Opinion Quarterly, 58*(2), 210-240. <https://doi.org/10.1086/269419>
- Auty, K. M., Farrington, D. P., & Coid, J. W. (2015). The validity of self-reported convictions in a community sample: Findings from the Cambridge Study in Delinquent Development. *European Journal of Criminology, 12*(5), 562-580. <https://doi.org/10.1177/1477370815578198>
- Balafoutas, L., Beck, A., Kerschbamer, R., & Sutter, M. (2013). What drives taxi drivers? A field experiment on fraud in a market for credence goods. *The Review of Economic Studies, 80*(3), 876-891. <https://doi.org/10.1093/restud/rds049>
- Bates, S. C., & Cox, J. M. (2008). The impact of computer versus paper-pencil survey, and individual versus group administration, on self-reports of sensitive behaviors. *Computers in Human Behavior, 24*(3), 903-916. <https://doi.org/10.1016/j.chb.2007.02.021>
- Beebe, T. J., Harrison, P. A., Mcrae, J. A., Anderson, R. E., & Fulkerson, J. A. (1998). An evaluation of computer-assisted self-interviews in a school setting. *Public Opinion Quarterly, 62*(4), 623-632. <https://doi.org/10.1086/297863>
- Beebe, T. J., Harrison, P. A., Park, E., McRae, J. A., Jr., & Evans, J. (2006). The effects of data collection mode and disclosure on adolescent reporting of health behavior. *Social Science Computer Review, 24*(4), 476-488. <https://doi.org/10.1177/0894439306288690>
- Bickman, L. (1971). The effect of social status on the honesty of others. *The Journal of Social Psychology, 85*(1), 87-92. <https://doi.org/10.1080/00224545.1971.9918547>
- Bradburn, N. M., Sudman, S., Blair, E., Locander, W., Miles, C., Singer, E., & Stocking, C. (1979). *Improving interview method and questionnaire design: Response effects to threatening questions in survey research*. Jossey-Bass.
- Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design-for market research, political polls, and social and health questionnaires*. Jossey-Bass.
- Brener, N. D., Eaton, D. K., Kann, L., Grunbaum, J. A., Gross, L. A., Kyle, T. M., & Ross, J. G. (2006). The association of survey setting and mode with self-reported health risk behaviors among high

- school students. *Public Opinion Quarterly*, 70(3), 354–374.
<https://doi.org/10.1093/poq/nfl003>
- Buckle, A., & Farrington, D. P. (1984). An observational study of shoplifting. *British Journal of Criminology*, 24(1), 63-73. <https://doi.org/10.1093/oxfordjournals.bjc.a047425>
- Buckle, A., & Farrington, D. P. (1994). Measuring shoplifting by systematic observation: A replication study. *Psychology, Crime and Law*, 1(2), 133-141.
<https://doi.org/10.1080/10683169408411946>
- Butler, S. F., Villapiano, A., & Malinow, A. (2009). The effect of computer-mediated administration on self-disclosure of problems on the Addiction Severity Index. *Journal of Addiction Medicine*, 3(4), 194-203. <https://doi.org/10.1097/ADM.0b013e3181902844>
- Christensen, L. B. (1985). *Experimental methodology* (3rd ed.). Allyn & Bacon.
- Clark, J. P., & Tiffit, L. L. (1966). Polygraph and interview validation of self-reported deviant behavior. *American Sociological Review*, 31(4), 516-523. <https://doi.org/10.2307/2090775>
- Cohn, A., Maréchal, M. A., Tannenbaum, D., & Zünd, C. L. (2019). Civic honesty around the globe. *Science*, 365(6448), 70-73. <https://doi.org/10.1126/science.aau8712>
- Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. *Personnel Psychology*, 47(4), 817-829.
<https://doi.org/10.1111/j.1744-6570.1994.tb01578.x>
- Denniston, M. M., Brener, N. D., Kann, L., Eaton, D. K., McManus, T., Kyle, T. M., Roberts, A. M., Flint, K. H., & Ross, J. G. (2010). Comparison of paper-and-pencil versus Web administration of the Youth Risk Behavior Survey (YRBS): Participation, data quality, and perceived privacy and anonymity. *Computers in Human Behavior*, 26(5), 1054-1060.
<https://doi.org/10.1016/j.chb.2010.03.006>
- Diener, E., Fraser, S. C., Beaman, A. L., & Kelem, R. T. (1976). Effects of deindividuation variables on stealing among Halloween trick-or-treaters. *Journal of Personality and Social Psychology*, 33(2), 178-183. <https://doi.org/10.1037/0022-3514.33.2.178>
- Dodou, D., & de Winter, J. C. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, 36, 487-495.
<https://doi.org/10.1016/j.chb.2014.04.005>
- Druckman, J. N., Gilli, M., Klar, S., & Robison, J. (2015). Measuring drug and alcohol use among college student-athletes. *Social Science Quarterly*, 96(2), 369-380.
<https://doi.org/10.1111/ssqu.12135>

- Enzmann, D., Kivivuori, J., Marshall, I. H., Steketee, M., Hough, M., & Killias, M. (2018). *A global perspective on young people as offenders and victims: First results from the ISRD3 study*. Springer. <https://doi.org/10.1007/978-3-319-63233-9>
- Fagan, A. A., Hawkins, J. D., Catalano, R. F., & Farrington, D. P. (2019). *Communities that Care: Building Community engagement and capacity to prevent youth behavior problems*. Oxford University Press. <https://doi.org/10.1093/oso/9780190299217.001.0001>
- Farrington, D. P. (1973). Self-reports of deviant behavior: Predictive and stable? *Journal of Criminal Law and Criminology*, *64*(1), 99-110. <https://doi.org/10.2307/1142661>
- Farrington, D. P. (1979). Experiments on deviance with special reference to dishonesty. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 12, pp. 207-252). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60263-4](https://doi.org/10.1016/S0065-2601(08)60263-4)
- Farrington, D. P. (1980). External validity: A problem for social psychology. In R. F. Kidd & M. J. Saks (Eds.), *Advances in applied social psychology* (Vol. 1, pp. 184-186). Lawrence Erlbaum. <https://doi.org/10.4324/9781315803005>
- Farrington, D. P. (2005, Ed.). *Integrated developmental and life-course theories of offending: Advances in criminological theory* (Vol. 14). Routledge. <https://doi.org/10.4324/9780203788431>
- Farrington, D. P. (2021). The developmental evidence base: Prevention. In D. A. Crighton & G. J. Towl (Eds.), *Forensic Psychology* (3rd ed., pp. 263-293). Wiley.
- Farrington, D. P., & Coid, J. W. (Eds.). (2003). *Early prevention of adult antisocial behaviour*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511489259>
- Farrington, D. P., & Jolliffe, D. (2004). England and Wales. In D. P. Farrington, P. A. Langan, & M. Tonry (Eds.), *Cross-national studies in crime and justice* (pp. 1-38). Bureau of Justice Statistics. <http://www.ojp.usdoj.gov/bjs>
- Farrington, D. P., Kazemian, L., & Piquero, A. R. (Eds.). (2019). *The Oxford handbook of developmental and life-course criminology*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190201371.001.0001>
- Farrington, D. P., & Kidd, R. F. (1977). Is financial dishonesty a rational decision? *British Journal of Social and Clinical Psychology*, *16*(2), 139-146. <https://doi.org/10.1111/j.2044-8260.1977.tb00209.x>
- Farrington, D. P., & Knight, B. J. (1979). Two non-reactive field experiments on stealing from a 'lost' letter. *British Journal of Social and Clinical Psychology*, *18*(3), 277-284. <https://doi.org/10.1111/j.2044-8260.1979.tb00337.x>

- Farrington, D. P., & Knight, B. J. (1980). Stealing from a “lost” letter: Effects of victim characteristics. *Criminal Justice and Behavior*, 7(4), 423-436. <https://doi.org/10.1177/009385488000700406>
- Farrington, D. P., Lösel, F., Braga, A. A., Mazerolle, L., Raine, A., Sherman, L. W., & Weisburd, D. (2020). Experimental criminology: Looking back and forward on the 20th anniversary of the Academy of Experimental Criminology. *Journal of Experimental Criminology*, 16, 649–673. <https://doi.org/10.1007/s11292-019-09384-z>
- Farrington, D. P., Ttofi, M. M., Crago, R. V., & Coid, J. W. (2014). Prevalence, frequency, onset, desistance and criminal career duration in self-reports compared with official records. *Criminal Behaviour and Mental Health*, 24(4), 241-253. <https://doi.org/10.1002/cbm.1930>
- Feldman, R. E. (1968). Response to compatriot and foreigner who seek assistance. *Journal of Personality and Social Psychology*, 10(3), 202-214. <https://doi.org/10.1037/h0026567>
- Franklin, B. J. (1973). The effects of status on the honesty and verbal responses of others. *The Journal of Social Psychology*, 91(2), 347-348. <https://doi.org/10.1080/00224545.1973.9923060>
- Franzen, A., & Pointner, S. (2013). The external validity of giving in the dictator game. *Experimental Economics*, 16(2), 155-169. <https://doi.org/10.1007/s10683-012-9337-5>
- Gerdtz, M., Yap, C. Y., Daniel, C., Knott, J. C., Kelly, P., & Braitberg, G. (2020). Prevalence of illicit substance use among patients presenting to the emergency department with acute behavioural disturbance: Rapid point-of-care saliva screening. *Emergency Medicine Australasia*, 32(3), 473-480. <https://doi.org/10.1111/1742-6723.13441>
- Gibbons, D. C. (1979). *The criminological enterprise: Theories and perspectives*. Prentice-Hall.
- Gibson, C. L., & Krohn, M. D. (Eds.). (2012). *Handbook of life-course criminology: Emerging trends and directions for future research*. Springer. <https://doi.org/10.1007/978-1-4614-5113-6>
- Giguère, K., Béhanzin, L., Guédou, F. A., Leblond, F. A., Goma-Matsétsé, E., Zannou, D. M., Affolabi, D., Kékè, R. K., Gangbo, F., Bachabi, M., & Alary, M. (2019). Biological validation of self-reported unprotected sex and comparison of underreporting over two different recall periods among female sex workers in Benin. *Open Forum Infectious Diseases*, 6(2), 1-6. <https://doi.org/10.1093/ofid/ofz010>
- Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, 47(4), 1237-1259. <https://doi.org/10.3758/s13428-014-0533-4>

- Gomes, H. S., Farrington, D. P., Defoe, I. N., & Maia, Â. (2021a). Field experiments on dishonesty and stealing: What have we learned in the last 40 years?. *Journal of Experimental Criminology*. Advance online publication. <https://doi.org/10.1007/s11292-021-09459-w>
- Gomes, H. S., Farrington, D. P., Krohn, M. D., Cunha, A., Jurdi, J., Sousa, B., Morgado, D., Hoft, J., Hartsell, E., Kassem, L., & Maia, Â. (2021c). *The impact of modes of administration on self-reports of offending: A two methodological experiment replication* [Manuscript submitted for publication]. School of Psychology, University of Minho.
- Gomes, H. S., Farrington, D. P., Krohn, M. D., & Maia, Â. (2021b). *How sensitive are self-reports of offending?: The impact of recall periods on question sensitivity* [Manuscript submitted for publication]. School of Psychology, University of Minho.
- Gomes, H. S., Farrington, D. P., Maia, Â., & Krohn, M. D. (2019). Measurement bias in self-reports of offending: A systematic review of experiments. *Journal of Experimental Criminology*, *15*(3), 313-339. <https://doi.org/10.1007/s11292-019-09379-w>
- Gomes, H. S., Maia, Â., & Farrington, D. P. (2018). Measuring offending: Self-reports, official records, systematic observation and experimentation. *Crime Psychology Review*, *4*(1), 26-44. <https://doi.org/10.1080/23744006.2018.1475455>
- Gold, M. (1966). Undetected delinquent behavior. *Journal of Research in Crime and Delinquency*, *3*(1), 27-46. <https://doi.org/10.1177/002242786600300103>
- Hardt, R. H., & Peterson-Hardt, S. (1977). On determining the quality of the delinquency self-report method. *Journal of Research in Crime and Delinquency*, *14*(2), 247-259. <https://doi.org/10.1177/002242787701400210>
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*(4), 1009-1055. <https://doi.org/10.1257/0022051043004577>
- Hindelang, M. J., Hirschi, T., & Weis, J. G. (1981). *Measuring delinquency*. Sage.
- Hirschi, T. (1969). *Causes of delinquency*. Transaction.
- Hornstein, H. A., Fisch, E., & Holmes, M. (1968). Influence of a model's feeling about his behavior and his relevance as a comparison other on observers' helping behavior. *Journal of Personality and Social Psychology*, *10*(3), 222-226. <https://doi.org/10.1037/h0026568>
- Hser, Y. I. (1997). Self-reported drug use: Results of selected empirical investigations of validity. In L. Harrison & A. Hughes (Eds.), *The validity of self-reported drug use: Improving the accuracy of survey estimates* (NIDA Research Monograph No. 167, pp. 320-343). U.S. Department of Health and Human Services. <https://www.ojp.gov/pdffiles1/Digitization/167339-167359NCJRS.pdf>

- Huizinga, D., & Elliott, D. S. (1986). Reassessing the reliability and validity of self-report delinquency measures. *Journal of Quantitative Criminology*, 2(4), 293-327. <https://doi.org/10.1007/BF01064258>
- Jobe, J. B., Pratt, W. F., Tourangeau, R., Baldwin, A. K., & Rasinski, K. A. (1997). Effects of interview mode on sensitive questions in a fertility survey. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwartz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 311-329). John Wiley & Sons. <https://doi.org/10.1002/9781118490013.ch13>
- Jolliffe, D., & Farrington, D. P. (2014). Self-reported offending: Reliability and validity. In G. Bruinsma, & D. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice* (pp. 4716-4723). Springer. https://doi.org/10.1007/978-1-4614-5690-2_648
- Jolliffe, D., Farrington, D. P., Hawkins, J. D., Catalano, R. F., Hill, K. G., & Kosterman, R. (2003). Predictive, concurrent, prospective and retrospective validity of self-reported delinquency. *Criminal Behaviour and Mental Health*, 13(3), 179-197. <https://doi.org/10.1002/cbm.541>
- Kabashi, S., Vindenes, V., Bryun, E. A., Koshkina, E. A., Nadezhdin, A. V., Tetenova, E. J., Kolgashkin, A. J., Petukhov, A. E., Perekhodov, S. N., Davydova, E. N., Gamboa, D., Hilberg, T., Lerdal, A., Nordby, G., Zhang, C., & Bogstrand, S. T. (2019). Harmful alcohol use among acutely ill hospitalized medical patients in Oslo and Moscow: A cross-sectional study. *Drug and Alcohol Dependence*, 204, 107588. <https://doi.org/10.1016/j.drugalcdep.2019.107588>
- Kazemian, L., & Farrington, D. P. (2005). Comparing the validity of prospective, retrospective, and official onset for different offending categories. *Journal of Quantitative Criminology*, 21(2), 127-147. <https://doi.org/10.1007/s10940-005-2489-0>
- Kerschbamer, R., Neururer, D., & Sutter, M. (2016). Insurance coverage of customers induces dishonesty of sellers in markets for credence goods. *Proceedings of the National Academy of Sciences*, 113(27), 7454-7458. <https://doi.org/10.1073/pnas.1518015113>
- Kirtadze, I., Otiashvili, D., Tabatadze, M., Vardanashvili, I., Sturua, L., Zabransky, T., & Anthony, J. C. (2018). Republic of Georgia estimates for prevalence of drug use: Randomized response techniques suggest under-estimation. *Drug and Alcohol Dependence*, 187, 300-304. <https://doi.org/10.1016/j.drugalcdep.2018.03.019>
- Kleck, G., & Roberts, K. (2012). What survey modes are most effective in eliciting self-reports of criminal or delinquent behavior? In L. Gideon (Ed.), *Handbook of survey methodology for the social sciences* (pp. 417-439). Springer. https://doi.org/10.1007/978-1-4614-3876-2_24

- Knapp, H., & Kirk, S. A. (2003). Using pencil and paper, Internet and touch-tone phones for self-administered surveys: Does methodology matter? *Computers in Human Behavior, 19*(1), 117-134. [https://doi.org/10.1016/S0747-5632\(02\)00008-0](https://doi.org/10.1016/S0747-5632(02)00008-0)
- Korte, C., & Kerr, N. (1975). Response to altruistic opportunities in urban and nonurban settings. *The Journal of Social Psychology, 95*(2), 183-184. <https://doi.org/10.1080/00224545.1975.9918701>
- Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and Web Surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly, 72*(5), 847-865. <https://doi.org/10.1093/poq/nfn063>
- Krohn, M., Thornberry, T., Bell, K., Lizotte, A., & Phillips, M. (2012). Self-report surveys within longitudinal panel designs. In D. Gadd, S. Karstedt, & S. Messner (Eds.), *The Sage handbook of criminological research* (pp. 23-35). Sage. <https://dx.doi.org/10.4135/9781446268285.n2>
- Krohn, M. D., Waldo, G. P., & Chiricos, T. G. (1974). Self-reported delinquency: A comparison of structured interviews and self-administered checklists. *Journal of Criminal Law and Criminology, 65*(4), 545-553. <https://doi.org/10.2307/1142528>
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity, 47*(4), 2025-2047. <https://doi.org/10.1007/s11135-011-9640-9>
- Kulik, J. A., Stein, K. B., & Sarbin, T. R. (1968). Disclosure of delinquent behavior under conditions of anonymity and nonanonymity. *Journal of Consulting and Clinical Psychology, 32*(5, Pt1), 506-509. <https://doi.org/10.1037/h0026260>
- Lee, H., Kim, S., Couper, M. P., & Woo, Y. (2019). Experimental comparison of PC web, smartphone web, and telephone surveys in the new technology era. *Social Science Computer Review, 37*(2), 234-247. <https://doi.org/10.1177/0894439318756867>
- Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world?. *Journal of Economic Perspectives, 21*(2), 153-174. <https://doi.org/10.1257/jep.21.2.153>
- Liber, A. C., & Warner, K. E. (2018). Has underreporting of cigarette consumption changed over time? Estimates derived from US National Health Surveillance Systems between 1965 and 2015. *American Journal of Epidemiology, 187*(1), 113-119. <https://doi.org/10.1093/aje/kwx196>
- Littlefield, A. K., Brown, J. L., DiClemente, R. J., Safonova, P., Sales, J. M., Rose, E. S., Belyakov, N., & Rassokhin, V. V. (2017). Phosphatidylethanol (PEth) as a biomarker of alcohol consumption in

- HIV-infected young Russian women: Comparison to self-report assessments of alcohol use. *AIDS and Behavior*, 21(7), 1938-1949. <https://doi.org/10.1007/s10461-017-1769-7>
- Livingston, M. (2013). *Assessment of adolescent alcohol use: Estimating and adjusting for measurement bias* (Publication No. 3729218) [Doctoral dissertation, University of Florida]. ProQuest Dissertations Publishing.
- Martins, P., Mendes, S., & Fernandez-Pacheco, G. (2015, September 2-5). *Cross-cultural adaptation and online administration of the Portuguese Version of ISRD3* [Paper presentation]. 15th Annual Conference of the European Society of Criminology, Porto, Portugal.
- Maxfield, M. G., & Babbie, E. R. (2009). *Basics of research methods for criminal justice and criminology* (2nd ed.). Cengage Learning.
- Merritt, C. B., & Fowler, R. G. (1948). The pecuniary honesty of the public at large. *The Journal of Abnormal and Social Psychology*, 43(1), 90–93. <https://doi.org/10.1037/h0061846>
- Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological review*, 100(4), 674-701. <https://doi.org/10.1037/0033-295x.100.4.674>
- Murphy, F. J., Shirley, M. M., & Witmer, H. L. (1946). The incidence of hidden delinquency. *American Journal of Orthopsychiatry*, 16(4), 686–696. <https://doi.org/10.1111/j.1939-0025.1946.tb05431.x>
- Nye, F. I., Short, J. F., & Olson, V. J. (1958). Socioeconomic status and delinquent behavior. *American Journal of Sociology*, 63(4), 381-389. <https://doi.org/10.1086/222261>
- Osgood, D. W., McMorris, B. J., & Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitative Criminology*, 18(3), 267-296. <https://doi.org/10.1023/A:1016008004010>
- Palamar, J. J., Salomone, A., & Keyes, K. M. (2021). Underreporting of drug use among electronic dance music party attendees. *Clinical Toxicology*, 59(3), 185-192. <https://doi.org/10.1080/15563650.2020.1785488>
- Pepper, J. V., & Petrie, C. V. (Eds.). (2003). *Measurement problems in criminal justice research: Workshop summary*. National Academy Press. <https://doi.org/10.17226/10581>
- Piquero, A. R., Schubert, C. A., & Brame, R. (2014). Comparing official and self-report records of offending across gender and race/ethnicity in a longitudinal study of serious youthful offenders. *Journal of Research in Crime and Delinquency*, 51(4), 526–556. <https://doi.org/10.1177/0022427813520445>

- Porterfield, A. L. (1943). Delinquency and its outcome in court and college. *American Journal of Sociology*, 49(3), 199–208. <https://doi.org/10.1086/219369>
- Potdar, R., & Koenig, M. A. (2005). Does audio-CASI improve reports of risky behavior? Evidence from a randomized field trial among young urban men in India. *Studies in Family Planning*, 36(2), 107-116. <https://doi.org/10.1111/j.1728-4465.2005.00048.x>
- Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, 84(5), 754-775. <https://doi.org/10.1037/0021-9010.84.5.754>
- Rijo, D., Miguel, R. R., Paulo, M., & Brazão, N. (2020). The effects of the growing pro-social program on early maladaptive schemas and schema-related emotions in male young offenders: A nonrandomized trial. *International Journal of Offender Therapy and Comparative Criminology*, 64(13-14), 1422-1442. <https://doi.org/10.1177/0306624X20912988>
- Robertson, R. E., Tran, F. W., Lewark, L. N., & Epstein, R. (2018). Estimates of non-heterosexual prevalence: The roles of anonymity and privacy in survey methodology. *Archives of Sexual Behavior*, 47(4), 1069-1084. <https://doi.org/10.1007/s10508-017-1044-z>
- Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, 45, 181-196. <https://doi.org/10.1016/j.joep.2014.10.002>
- Schober, S. E., Caces, M. F., Pergamit, M. R., & Branden, L. (1992). Effect of mode of administration on reporting of drug use in the National Longitudinal Survey. In C. F. Turner, J. T. Lessler, & J. C. Gfroerer (Eds.), *Survey measurement of drug use: Methodological studies* (pp. 267–276). National Institute on Drug Abuse.
- Schore, J., Maynard, R., & Piliavin, I. (1979). *The accuracy of self-reported arrest data*. Mathematica Policy Research.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54(2), 93-105. <https://doi.org/10.1037/0003-066X.54.2.93>
- Schwarz, N., Strack, F., Hippler, H. J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, 5(3), 193-212. <https://doi.org/10.1002/acp.2350050304>
- Sellin, T. (1931). The basis of a crime index. *Journal of Criminal Law and Criminology*, 22(3), 335-356. <https://doi.org/10.2307/1135784>

- Short, J. F., & Nye, F. I. (1958). Extent of unrecorded juvenile delinquency tentative conclusions. *The Journal of Criminal Law, Criminology, and Police Science*, 49(4), 296-302. <https://doi.org/10.2307/1141583>
- Steinberg, J., McDonald, P., & O'Neal, E. (1977). Petty theft in a naturalistic setting: The effects of bystander presence. *The Journal of Social Psychology*, 101(2), 219-221. <https://doi.org/10.1080/00224545.1977.9924010>
- Strang, E., & Peterson, Z. D. (2020). Use of a bogus pipeline to detect men's underreporting of sexually aggressive behavior. *Journal of Interpersonal Violence*, 35(1-2), 208-232. <https://doi.org/10.1177/0886260516681157>
- Sudman, S., & Bradburn, N. M. (1974). *Response effects in surveys: A review and synthesis*. Aldine Publishing Company.
- Sullivan, C. J., & McGloin, J. M. (2014). Looking back to move forward: Some thoughts on measuring crime and delinquency over the past 50 years. *Journal of Research in Crime and Delinquency*, 51(4), 445-466. <https://doi.org/10.1177/0022427813520446>
- Thornberry, T. P., & Krohn, M. D. (2000). The self-report method for measuring delinquency and crime. In D. Duffee (Ed.), *Measurement and analysis of crime and justice* (pp. 33-84). U.S. National Institute of Justice.
- Tourangeau, R., & McNeeley, M. E. (2003). Measuring crime and crime victimization: Methodological issues. In J. V. Pepper, & C. V. Petrie (Eds.), *Measurement problems in criminal justice research: Workshop summary* (pp. 10-42). National Academy Press.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. <https://doi.org/10.1017/CB09780511819322>
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859-883. <https://doi.org/10.1037/0033-2909.133.5.859>
- Tourangeau, R., & Yan, T. (in press). Reporting issues in surveys of drug use. *Substance Use and Misuse*.
- Trau, R. N., Härtel, C. E., & Härtel, G. F. (2013). Reaching and hearing the invisible: Organizational research on invisible stigmatized groups via web surveys. *British Journal of Management*, 24(4), 532-541. <https://doi.org/10.1111/j.1467-8551.2012.00826.x>
- Turner, C. F., Lessler, J. T., & Devore, J. W. (1992). Effects of mode of administration and wording on reporting of drug use. In C. F. Turner, J. T. Lessler, & J. C. Gfroerer (Eds.), *Survey measurement of drug use: Methodological studies* (pp. 177-219). National Institute on Drug Abuse.

- Vinikoor, M. J., Zyambo, Z., Muyoyeta, M., Chander, G., Saag, M. S., & Cropsey, K. (2018). Point-of-care urine ethyl glucuronide testing to detect alcohol use among HIV-hepatitis B virus coinfecting adults in Zambia. *AIDS and Behavior, 22*(7), 2334-2339. <https://doi.org/10.1007/s10461-018-2030-8>
- Wehling, H., & Lusher, J. (2019). People with a body mass index ≥ 30 under-report their dietary intake: a systematic review. *Journal of health psychology, 24*(14), 2042-2059. <https://doi.org/10.1177/1359105317714318>
- Wolter, F., & Laier, B. (2014). The effectiveness of the item count technique in eliciting valid answers to sensitive questions. An evaluation in the context of self-reported delinquency. *Survey Research Methods, 8*(3), 153-168. <https://doi.org/10.18148/srm/2014.v8i3.5819>
- Yan, T., & Cantor, D. (2019). Asking survey questions about criminal justice involvement. *Public Health Reports, 134*(1_suppl), 46S-56S. <https://doi.org/10.1177/0033354919826566>
- Zara, G., & Farrington, D. P. (2016). *Criminal recidivism: Explanation, prediction and prevention*. Routledge.
- Zimny, G. H. (1961). *Method in experimental psychology*. Ronald Press Company. <https://doi.org/10.1037/14006-000>

CHAPTER I
MEASURING OFFENDING:
SELF-REPORTS, OFFICIAL RECORDS, SYSTEMATIC OBSERVATION AND EXPERIMENTATION

Manuscript Published in:

Gomes, H. S., Maia, Â., & Farrington, D. P. (2018). Measuring offending: Self-reports, official records, systematic observation and experimentation. *Crime Psychology Review*, 4(1), 26-44.
<https://doi.org/10.1080/23744006.2018.1475455>

Reprinted in:

Gomes, H.S., Maia, Â., & Farrington, D.P. (2021). Measuring offending: Self-reports, official records, systematic observation and experimentation. In D. Canter & D. Youngs (Eds.), *Reviewing Crime Psychology* (pp. 334-352). Routledge. (Reprinted from "Measuring offending: Self-reports, official records, systematic observation and experimentation," 2018, *Crime Psychology Review*, 4[1], 26-44, <https://doi.org/10.1080/23744006.2018.1475455>)

MEASURING OFFENDING: SELF-REPORTS, OFFICIAL RECORDS, SYSTEMATIC OBSERVATION AND EXPERIMENTATION

Abstract

Criminological knowledge can only be as accurate as the measure of crime itself. Concern with crime measurement starts with the definition of crime, which has consequences for the measurement techniques preferred in different domains. The two main methodologies used to measure criminal behaviour are official records and self-reports of offending (SRO). Although some researchers are concerned about official records being filtered and deeply flawed estimates of criminal activity, others doubt that people can or will provide reliable information about their own criminal behaviour by completing a survey. In this article, we present a historical overview of the development of these techniques and discuss some of the main results of comparing official records and SRO. Throughout this discussion, we explore to what extent criminological conclusions differ depending on the measurement method and the potential implications of these differences. Finally, we present some alternative ways to measure offending, such as systematic observation, which could prove to be very important in improving criminological knowledge. In a period when criminologists seem to be increasingly concerned with the validity of measures of crime, this article reviews the major issues in crime measurement, as well as the advantages and limitations of the primary methodologies.

Keywords: Measurement; Crime; Official records; Self-report; Observation

Introduction

In studying crime, researchers struggle with a variable that is inherently difficult to measure (Osgood et al., 2002). Nevertheless, researchers have developed multiple ways of measuring delinquency and criminal conduct, giving rise to the criminological knowledge that is essential in all developed societies (e.g., crime statistics, criminal careers, risk factors, intervention effectiveness, etc.). Unfortunately, current measures of crime are recognized as being deeply flawed, and it has become a common practice in criminological studies to attach warning labels about potential validity problems and to point out that different methods may result in different estimates of crime (Enzmann, 2013).

In this article, we present the major issues in measuring crime, a brief historical overview of the development of measurement techniques and a review of the primary methods of measuring criminal conduct. In describing these different methodologies, we will consider the advantages and limitations of each method, in order to achieve a broad and integrated understanding about crime and delinquency methods of measurement.

Definition and units of measurement

The generally accepted definition of crime seems to be a legal definition: “any act committed in violation of a law that prohibits it and authorizes punishment for its commission” (Wilson & Herrnstein, 1985, p. 22). However, this definition presents several problems; for example, laws have to be understood in time and space. Because laws are subject to change, many behaviours that were classified as crimes 20 years ago may not be considered as crimes today, and vice versa. Moreover, since laws are the subject of political decisions, what is considered crime in one country may not be viewed as crime in another, which is a potential limitation for comparative research between different countries. For example, in Portugal, the acquisition, possession, and use of small quantities of drugs were decriminalized in 2001 (Greenwald, 2009). Although drug usage is still prohibited, no punishment is applied, so drug usage is not a crime in Portugal. As a result, this might constitute a source of bias when comparing criminal records between Portugal and other countries where drug usage is a crime, as well as within Portugal before and after 2001.

The legal definition of crime has long been criticized. Sellin (1938) pointed out that laws embodied the values of dominant groups and that this causes variation in the definition of ‘crime’ and ‘criminal’. In turn, the absence of an established basic unit of crime, accepted and shared by researchers within the field, constitutes a violation of a fundamental tenet of scientific research and threatens the validity of any

results. Building on the idea that criminal laws do not meet the demands of scientific research, several other constructs have been proposed, such as delinquency, deviance and antisocial behaviour, among others, where the criterion extends beyond the legal definition, granting the utmost importance to the violation of social norms. Nevertheless, the definition of crime still stands today as any behaviour specified and punishable by the criminal law (McLaughlin & Muncie, 2001).

A standardized unit of measurement is crucial for every field of research. In the particular case of criminal behaviour, there are four major elements: the offender, victim, offence, and incident (Maxfield & Babbie, 2009). In a single incident, such as a burglary, one or more offenders may commit several offences (e.g., theft, property damage, etc.) and harm several victims. Therefore, whether or not we focus on one unit of measurement rather than another, the final result may differ significantly. Moreover, different units of measurement are more relevant to different research questions, depending on the objective of a particular study. For example, if a researcher wants to study victimless crimes, such as drug dealing, then he/she would not consider the number of victims as the unit of measurement. However, it is very important to consider which unit of measurement is being used, especially in comparative studies where the results may vary as a function of the unit of measurement.

In attempting to link up results from different data sources, it is important to compare like with like. For example, consider the problem of estimating the probability of a burglary leading to a conviction for burglary. It is possible to compare the number of residential burglaries reported by victims (e.g., in the Crime Survey for England and Wales) – let us call this V – with the number of persons convicted for burglary in England and Wales – C . However, the probability of a burglary is not C/V . This is because a burglary is an offence, but a person convicted for burglary is an offender– offence combination. If three persons are convicted for the same burglary, this produces three convictions for burglary. However, if a burglary is committed by three people, this is still only one offence. To compare like with like, the number of burglaries must be multiplied by the average number of offenders committing each burglary. Then, the number of offender–offence combinations for burglary can be validly compared with the number of offender–offence combinations which are convictions for burglary (see e.g., Farrington et al., 2004a).

Measure of crime

Basically, there are two major methods used to measure criminal behaviour: 1. Official records; and 2. self-reports of offending (SRO) (Piquero et al., 2014). Each of these methods is divided into different forms of criminal behaviour. Official records might derive from crimes known to the police, police arrests, court appearances, convictions, or prison data, whereas SRO might focus on the offender's point of view

or be derived from the victims of crimes. This multitude of techniques brings an even greater complexity to the already complex question of how to measure crime, and is the motivation for the present article.

Brief historical overview

Initially, up to the 1950's, criminal knowledge relied mainly on official records, collecting data from arrests and court appearances in order to study criminal behaviour (Farrington et al., 2007). For example, Durea (1935) used juvenile court appearances to test the relationship between intelligence and delinquent behaviour. Likewise, Bogen (1944) relied on court records to study the effect of economic trends on criminality.

However, official records and crimes known to the police are only the tip of the iceberg of illegal activity. Obviously, official records are a filtered measure of criminal activity, and they encompass the limitations derived from multiple sources, namely, law enforcement, policing, characteristics of specific crimes reported by victims and the definition of crime. In an attempt to estimate this unknown amount of illegal activity, entitled 'hidden delinquency', Murphy et al. (1946) concluded that no violations of city ordinances (0%), about 0.6% of minor offences and only 11% of serious offences were prosecuted as a matter of official complaints. This resulted in a total of less than 1.5% of infractions that originated in official complaints. Later researchers found similar results, of the order of about 1 in 30 offences leading to court referrals and, in the case of marijuana use, around 1 in 1,000 (Farrington et al., 2003).

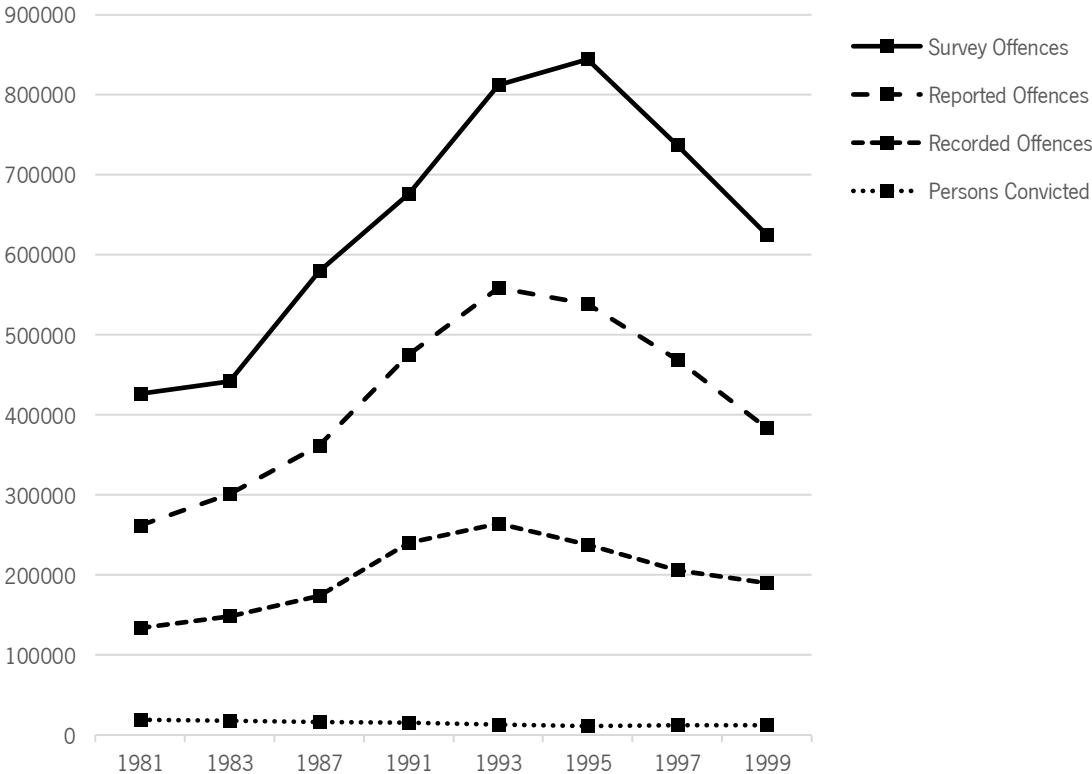
Beyond these limitations, not all crimes that reach the attention of the authorities are officially recorded (Thornberry & Krohn, 2000). Sellin (1931) found that different sources of official data may result in different estimates of criminal behaviour. In this classic study, Sellin found a clear reduction from the estimates of crimes known to the police, to the prevalence of persons tried in court and the numbers convicted and sentenced to imprisonment. He concluded that "the value of a crime rate for index purposes decreases as the distance from the crime itself in terms of procedure increases" (Sellin, 1931, p. 346).

In line with Sellin's (1931) findings, Farrington and Jolliffe (2004, p. 1) stated that the "Criminal Justice System involves a successive funnelling process". Of all crimes committed, only some are reported to the police, then only some are recorded by the police, only some offenders are convicted and then only some are sentenced to custody. Figure 1 shows that, in general, estimates of crime increased from 1981 to 1995; survey offences increased by 98%, reported offences increased by 106% and recorded offences increased by 78%, but persons convicted showed a reduction of 41% over the same period. Moreover, from 1995 to 1999, although survey offences, reported offences and recorded offences

showed reductions of 26%, 29%, and 20%, respectively, the number of persons convicted increased by 8% over the same period. This figure clearly illustrates that different sources provide completely different estimations of offending behaviour, which is very concerning for the validity of the data and for the conclusions derived from it. We must add that, despite the fact that these results refer only to England and Wales, the other countries (i.e., United States, Australia, Canada, the Netherlands, Scotland, Sweden, and Switzerland) considered in the report by Farrington et al. (2004b) follow the same trend.

Figure 1

The prevalence of crimes (i.e., burglary, vehicle theft, robbery, assault, rape and homicide) in England and Wales according to different sources



Note. Source: Figure constructed by the authors based on data presented by Farrington and Jolliffe (2004).

In an attempt to overcome these limitations, researchers developed a new method by using the self-report technique to measure criminal behaviour. Porterfield (1943) was the first to publish an article that used self-report questionnaires in order to compare the prevalence of delinquency between juveniles

in court and college students. However, the groundbreaking results of Nye et al. (1958), about the minimal differences in the prevalence of delinquent behaviour between different socioeconomic strata, revealed the true potential of the SRO technique (Krohn et al., 2010). These works drastically changed criminologists' opinions of SRO, and Hirschi (1969) developed the Social Control Theory based on this methodology.

In 1973, Farrington published the first review of the literature on the psychometric qualities of SRO surveys and concluded that this technique had predictive validity. Although it should not replace entirely the officially recorded data, Farrington (1973, p. 109) suggested that "the most accurate measure of deviant behaviour may yet prove to be some combination of official records and a self-report questionnaire". Hindelang et al. (1981) studied this technique and produced a highly influential book called '*Measuring delinquency*' that was a milestone in the use of the self-report methodology in criminological research, demonstrating that SRO were a valid measure of crime and delinquency. Since then, the self-report method became 'one of the most important innovations in criminological research in the 20th century' (Thornberry & Krohn, 2000, p.34) and since then criminological knowledge (e.g., criminal patterns, delinquency theories, etc.) has relied almost exclusively on data obtained by the self-report methodology (Cops et al., 2016).

Comparing official records and self-reports of offending

Since the development of SRO, researchers have been interested in comparisons between official records and SRO data and have used official records as a standard to study the criterion validity of SRO (Hindelang et al., 1979, 1981). The idea was that, if the two methods measure the same construct, they should be positively correlated. As a matter of fact, Hindelang et al. (1981) found considerable concordance between official records and SRO, which led them to the conclusion that people generally admit their criminal practices. Other authors, such as West and Farrington (1977), also found an association between SRO and officially recorded offending. However, to better understand this relation, we should consider the finding by Farrington (1977) that after criminal convictions – or public labelling – there is an increase in self-reported offending, so convictions could make known offenders more likely to admit their delinquent behaviour. Nevertheless, several researchers have found that SRO significantly predict future convictions among unconvicted people (e.g., Farrington, 2003), which indicates the validity of SRO.

It might be expected that SRO would provide higher estimates of offending since this technique was developed with the objective of overcoming the limitations inherent in official records (Farrington et

al., 2007). Despite the associations described above, researchers looked deeper into the differences between the results obtained by these two methods and their implications for criminological knowledge. In this article, we will focus on the primary differences in conclusions derived from SRO and official records in measuring criminal behaviour.

Scaling-up factor

As stated earlier, the primary limitation of official records of crime is that they provide an underestimate of offending. In an attempt to estimate the real number of crimes per conviction, researchers developed the 'scaling-up factor', which is "estimated by comparing convictions and self-reported offences of the same people at the same ages" (Theobald et al., 2014, p. 265).

Considering the males in the Cambridge Study in Delinquent Development (CSDD; $n = 411$), at the ages of 15–18, 27–32, and 42–47, Farrington et al. (2013) estimated a scaling-up factor of 39 self-reported offences per conviction. In the Pittsburgh Girls Study (PGS), based on a sample of 2,450 girls between ages 12 and 17, a scaling-up factor of five self-reported offences was found for every police charge (Ahonen et al., 2017). In a longer follow-up of the PGS, this rose to a factor of 12 between ages 11 and 19 (Jennings et al., 2018). In this latter study, 33% of low-rate official offenders (with one to four police charges) and 27% of high-rate official offenders (with five or more police charges) self-reported no offences. This highlights possible gender differences in the scaling-up factor. In the Pittsburgh Youth Study (PYS, $n = 506$), with boys aged between 13 and 17, Farrington et al. (2007) found a scaling-up factor of 80. Later, in the same PYS, with boys aged between 13 and 24, Theobald et al. (2014) found a scaling-up factor of nine self-reported offences for each conviction.

Moreover, the evidence seems to suggest that the scaling-up factor changes throughout the life course. Indeed, Theobald et al. (2014) found that this factor increased from 8 at ages 13–15 to 14 at ages 22–24. In the CSDD, younger males (aged 15–18) had a ratio of 47 self-reported offences for each conviction, older males (aged 27–32) had a ratio of 33 and the oldest males (aged 42–47) had the lowest ratio of 17 (Farrington et al., 2013). It is possible that the relationship between the scaling-up factor and age is curvilinear. Clearly, more research on this is needed.

The self-reported offences per conviction ratio also seems to vary as a function of types of crime. For example, in the CSDD, burglary and theft of vehicles had the lowest scaling-up ratios, 6 and 9, respectively, compared with theft from work and drug offences that had alarming ratios of 1,463 and 4,160 self-reported offences per conviction, respectively (Farrington et al., 2013). In the PYS, Farrington et al. (2007) reported that property offences had the lowest scaling-up factor (15), followed by violent

offences (154) and, finally, drug offences had the highest ratio (424). Moreover, Theobald et al. (2014) compared serious (5) and moderate (16) thefts, as well as serious (11) and moderate (13) violence offences, and found that in both cases the scaling-up factor was higher for serious offence types (Theobald et al., 2014).

Another interesting result is that the scaling-up factor seems to change as a function of race. This was found by Ahonen et al. (2017) in the PGS, where African American girls (7) had a much higher ratio than Caucasian girls (2). This difference was even higher at younger ages. At age 13, African Americans had a ratio more than five times higher (27 vs. 5), a difference that gradually decreased with age (at age 17, 5 vs. 2). The results with boys in the PYS followed a different trend, where Caucasian boys had a ratio of 10, slightly higher than African American boys (8). Theobald et al. (2014, p. 274) interpreted this result by explaining that African Americans are more exposed to risk factors, and that this result “does not necessarily mean that the police or the courts are biased against African American boys”, although more research on this topic is clearly needed.

Implications

The discussion on the scaling-up factor clearly shows the different estimates obtained from the two measures of crime, SRO and official records. Moreover, these different estimates of criminal behaviour varied differently with age, type of offences, race, etc. An obvious consequence is the likelihood of drawing different conclusions from the different research methods.

Criminal career research

One other topic where different estimates of criminal behaviour from official records and SRO might have major implications is in the research on criminal careers. Authors such as Blumstein et al. (1986, 1988) demonstrated the importance of criminal career research. Understanding the sequence of offences over time of particular offenders allows us to understand the beginning of offending (i.e., the age of onset), the maintenance of criminal behaviour (i.e., persistence), the moment when they stop offending (i.e., desistance) – and, thus, criminal career duration – as well as knowledge about changes in criminal behaviour, such as specialization or diversification of criminal acts, escalation or de-escalation of the seriousness of crimes, etc.

Because criminal career research requires exact information about the dates of offences, the majority of studies have based the measurement of criminal behaviour on official records, rather than on SRO, a fact that some authors have considered potentially misleading (Farrington et al., 2003). Therefore,

Farrington et al. (2003) suggested that SRO of offending might add value to criminal career research, with a more accurate estimate of the total number of crimes. This, and subsequent studies that based criminal career research on both methods, faced the problems of different estimates when based on official data and when based on SRO.

Age of onset

Considering data from the Seattle Social Development Project (SSDP; $n = 808$), Farrington et al. (2003) found that the first offence reported in the surveys preceded, on average, by 2.4 years the first crime in the official data (i.e., court referral). More exactly, in this study, while the average age of onset based on official records was age 15.1, the average age of onset based on SRO was at age 12.7. Moreover, this study estimated that, on average, an offender commits 26 offences before the first crime is officially recorded.

Concordant results were found by Loeber et al. (2003) in the OJJDP Study Group on Very Young Offenders, where the age of onset based on self-reported serious delinquency was at the age of 11.9, whereas the average age of onset based on official records (i.e., court contact) happened 2.6 years later, at the age of 14.5 years. Kazemian and Farrington (2005) analysed data from the CSDD and found similar results. They found that the age of onset based on SRO was, on average, at 11.9, whereas the age of onset was, on average, 16.9 based on official records (5 years later).

Moreover, Kazemian and Farrington (2005) noticed a relationship between the seriousness of offences and the agreement between the two estimates of the age of onset. The difference between the estimates of the age of onset based on SRO and official records became less pronounced as the seriousness of crimes increased. For example, these authors found a difference of 1.6 years for theft of vehicles (age of onset: SRO = 15.2; official records = 16.8) compared to a difference of 12 years for vandalism (age of onset: SRO = 10.7; official records = 22.7). The fact that serious offences are more likely to result in court convictions, compared with minor offences, may explain these results.

Criminal career duration

To date, we have discussed how SRO provide a much higher estimate of the number of crimes and indicate that criminal activity starts much earlier than according to official records of crime. An obvious consequence seems to be that criminal career duration should be longer if studied with SRO compared with official criminal records. In fact, some authors have found such a result. For example, Le

Blanc and Fréchette (1989; $n = 470$) found a duration of 5.23 years for the criminal career if based on convictions, but a career more than twice as long of 10.76 years based on SRO.

Farrington et al. (2014) published a very informative paper that addressed these questions about career research. Using the data from the CSDD (between ages 8 and 48), these authors showed that, while the average age of onset in SRO was at 10, the first conviction did not happen on average until 19. Similarly, the age of desistance in SRO was at 35, whereas in official records it happened much sooner in life, at the age of 25. There was an average criminal career duration of 25 years according to SRO, compared with an average duration of 6 years based on convictions, a 19-year difference (Farrington et al., 2014).

Implications

Criminal career research provides a great example of the different, and at times contrasting, conclusions that could be derived from different methods for measuring criminal behaviour.

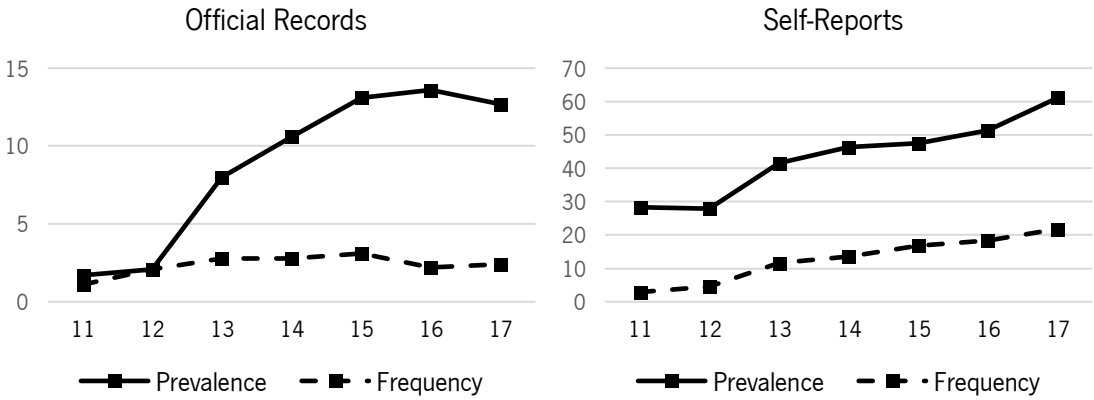
The reader should keep in mind that this is only a part of the problem. There are also differences in criminal patterns. For example, Kazemian and Farrington (2005) found that whereas SRO data indicated a pattern where individuals start with a minor offence and gradually commit more serious offences, the results based on official records of crime showed the opposite serious-to-minor pattern. On the other hand, although some authors argued that criminal features, such as prevalence and frequency, vary similarly with age (e.g., Hirschi & Gottfredson, 1983), others opposed this idea and argued that the age–crime curve was driven by prevalence, while frequency was pretty constant with age (Blumstein et al., 1988). When testing this hypothesis, Farrington et al. (2003) in the SSDP discovered that both prevalence and frequency increased with age in SRO, but only prevalence increased in official records, whereas offending frequency stayed constant with age (Figure 2).

Moreover, in the PYS, Farrington et al. (2007) found that, if based on SRO, the frequency of offences per offender increased with age during adolescence, but, based on official records, the frequency of offences per offender seemed to remain constant across adolescence. This led the authors to the conclusion that “in attempting to explain offending, researchers should always measure both self-reports and official records” (Farrington et al., 2007, p. 246).

It thus seems clear that different conclusions may be obtained from these two methods of collecting criminal data. This is a question that should be considered seriously because it may result in different theoretical and policy implications (Kazemian & Farrington, 2005).

Figure 2

Prevalence and frequency of offending according to different sources



Note. Source: Farrington et al. (2003).

Self-reports of offending

Despite the conclusion of several authors that the best estimate of crime may be achieved by a combination of both SRO and official records (e.g., Farrington, 1973; Farrington et al., 2003), the results presented in this article led some researchers to the conclusion that “official records are biased and yield distorted information about the true characteristics of offenders” (Farrington et al., 2007, p. 229). Others concluded that the “prevalence and mean frequency of self-reported offending is a better indicator of actual delinquent behaviour than is being charged by the police or the frequency of police charges” (Loeber et al., 2015, p. 163).

All these reasons favouring SRO over official records certainly do not mean that SRO are without limitations. Self-reports of human behaviour can be affected by multiple factors. The format of the questionnaire, the wording of questions, the response format, modes of administration, etc. have been previously shown to impact self-reports of behaviour and attitudes (Schwarz, 1999). Furthermore, due to the sensitive and potentially incriminating nature of criminal behaviour, we have reason to believe that SRO would be even more sensitive to these biasing factors (Thornberry & Krohn, 2000). Despite the concern about these potential biases (as shown throughout the present article), current knowledge does not allow us to know which factors impact self-reports in general, to what extent these factors impact SRO in particular and how to control or minimize their effects.

According to our literature review, we can divide the major concerns about self-reports into three main categories: 1) questionnaire design; 2) modes of administration; and 3) testing effects.

Questionnaire design

The way the questionnaire is designed presents several features that constitute potential sources of bias. In 1973, Farrington revealed a concern about the phrasing of the questions. He noted that the vast majority of survey questions on offending were presented to the participants in the same direction, and proposed phrasing questions both positively and negatively, as a way to minimize the potential for acquiescence response bias. However, positive phrasing still stands today as the norm in measuring delinquent behaviour, e.g. 'Have you ever in your life broken into a building to steal something?' (ISRD3 Working Group, 2013) and 'Have you ever stolen something worth more than 50 euros' (Sanches et al., 2016).

Enzmann (2013) focused on the response formats and tested the effects of reversing the response categories 'yes' and 'no', along with a short version of the questionnaire, and omitting follow-up questions. The findings from this experiment showed that the short version, where 'yes' appears first, and with omitted follow-up questions, generated higher estimates of offending, primarily concerning minor crimes. Response order effects have been previously described as resulting from primacy effects or social desirability; for example, participants may see the first response option as the most natural answer (Enzmann, 2013; Schwarz et al., 1991). The impact of follow-up questions is particularly important, because researchers are interested in much more information, rather than only knowing whether a person did or did not commit a certain crime (e.g., how many times; with whom; where it happened, etc.). Since follow-up questions are contingent on affirmative answers, participants might learn to answer negatively in order to avoid further questions and minimize the length of the interview (Thornberry, 1989). It may be that participants obey the law of least effort.

Modes of administration

Administration methods of SRO have long been a concern of researchers (e.g., Gold, 1966). Hindelang et al. (1981) developed a study where participants were randomly assigned to four conditions (i.e., non-anonymous questionnaire, anonymous questionnaire, non-anonymous interview and anonymous interview). The results showed small to no significant differences between the four methods, leading the researchers to the conclusion that SRO are largely independent of the modes of administration. This finding resulted in a substantial decrease in methodological studies of delinquency

surveys. It seemed that the validity of the self-report technique had been established and that further methodological research was unnecessary (Jolliffe & Farrington, 2014).

Fifteen years later, Tourangeau and Smith (1996) summarized multiple studies of method effects, concluding that participants are generally more willing to report illegal activities using self-administration methods, rather than admitting them to interviewers. This experimental study compared three conditions: computer-assisted personal interviewing (CAPI), computer-assisted self-administered interviewing (CASI), and audio computer-assisted self-administered interviewing (ACASI). The findings provided evidence for the presence of method effects, showing that self-administration resulted in higher reports of sexual behaviour and drug use.

More recently, other researchers were also able to demonstrate the impact of method effects on reports of offending. Denniston et al. (2010) conducted an experiment comparing paper-and-pencil versus web administration, and found that participants in the paper-and-pencil condition reported higher perceived privacy and anonymity. Wright et al. (1998) compared self-reports of smoking, alcohol and drug use in computer-assisted versus paper-and-pencil conditions. The results showed higher reports in the computer-assisted condition. Moreover, these authors found an interaction between method effects and the age of the participants, since adolescent participants were more sensitive to method effects. Lucia et al. (2007) showed that paper-and-pencil administration yielded significantly higher reports of delinquency than the internet condition in three out of 22 comparisons (i.e., selling soft drugs, vandalism and theft from the person).

Thornberry and Krohn (2000) published a review of SRO, concluding that computer-assisted self-interviews with audio elicited higher rates of delinquency. Whether or not different modes of administration impact participants' reports of delinquent behaviour is still a debatable subject. Moreover, we seem to know very little about the underlying processes that might explain these effects, the factors that interact with method effects (e.g., age, sensitivity of questions) or even the direction of impact.

Testing and panel effects

Testing and panel effects are a serious threat to longitudinal studies of delinquent behaviour. Considering the prominent place of longitudinal designs in criminological research (Krohn et al., 2012), ensuring the quality of its data should be a priority in this field of science. As described by Thornberry (1989, p. 351), testing effects are "any alterations of a subject's response to a particular item or scale caused by the prior administration of the same item or scale", whereas panel effects refer to "a more

general reaction to being re-interviewed”, rather than a specific reaction to questionnaire characteristics (Thornberry, 1989, p. 361).

In his study, Thornberry (1989) was able to find a decrease in the prevalence of delinquent behaviour as a function of the number of prior interviews. Other researchers also demonstrated reductions in participants’ reports of delinquency in prospective longitudinal studies that were inconsistent with the age–crime curve (e.g., Bosick, 2009; Lauritsen, 1998). However, whether these results are due to testing/panel effects is still questionable, and other features could explain such declines, for example, scale construction, item-specific age–crime curves or selective attrition (Bosick, 2009). More recently, Krohn et al. (2012) reviewed the role of surveys within longitudinal studies and appealed for the importance of further investigating the potential for testing and panel effects.

Future directions for self-report methods

This list of biasing effects, by no means exhaustive, may constitute a real threat to the quality of SRO and, by extension, to the validity of criminological knowledge. Nonetheless, these potential biases of the self-report technique should not be seen as reasons not to use questionnaires to measure delinquency. On the contrary, it is urgent to develop experimental studies to test to what extent these effects might contaminate the quality of results and to try to understand the ways in which they interact with each other. By doing this, we can strive to obtain ever better results, closer and closer to the reality of offending. Furthermore, developing knowledge about the best way to survey participants could also facilitate standardized self-report measures of offending.

Alternative methods for measuring crime

Considering the literature reviewed in this paper, even if we accept that the self-report technique provides a better indicator of offending behaviour than official records, we are led to the conclusion that survey methods are ‘also rather biased and indirect measures of offending’ (Buckle & Farrington, 1984). If anything, the previous discussion on the biases of SRO shows that there is still much work to do in order to fully understand how different methodological features interfere with the reporting of criminal behaviour.

In response, some researchers have turned to observation methods to measure criminal behaviour. Obviously, most criminal acts are unpredictable, some are very difficult to observe directly (e.g., white collar crime), and offenders try to conceal their illicit activity. However, direct observation can be a very useful technique in specific domains. For example, Konecni et al. (1976) carried out systematic

observation of drivers' behaviour and found that younger males were more likely to violate a red light in an intersection.

Buckle and Farrington (1984) systematically observed shoppers and were able to observe nine out of 503 (1.8%) customers shoplifting. In a later study, Buckle and Farrington (1994) used the same technique and observed this illegal behaviour by 15 out of 988 (1.5%) customers. The findings from these studies included information about the personal characteristics of shoplifters (e.g., mostly males and more likely to be over 55 years old), about the offence itself (mostly small low-cost items were stolen) and about offenders' behaviour during the offence (most checked if they were being observed by anyone before shoplifting) and after the offence (most purchased other goods to allay suspicion). Despite some concerns about this technique, namely regarding the generalization of results, these authors concluded that this method of measuring shoplifting was valid, and that it should be used more frequently.

Also on the theme of shoplifting, Buckle et al. (1992) tested the technique of systematic counting. In this study, researchers repeatedly and systematically counted the number of specific items in each shop daily in order to detect item loss. By comparing items missing with items purchased in a total pool of 29 stores, the researchers found that 10.9% of items leaving the store were stolen. In the worst store, more than one-third of minor items were stolen. The authors analysed the qualities of this technique and concluded that systematic counting to measure shoplifting produced valid results. Shortly after, Farrington et al. (1993) carried out an experiment to evaluate the effectiveness of three situational interventions in preventing shoplifting, using systematic counting as the behavioural measure. This study provided evidence that one of these situational programs (i.e., electronic tagging) caused significant decreases in shoplifting over time.

In the late 1970s and early 1980s, researchers were interested in field experiments in criminology, which resulted in some interesting methods of measuring deviant behaviour. Farrington and Kidd (1977) conducted a field experiment with the purpose of studying the decision process in committing financially dishonest behaviour. Basically, these authors provided random citizens the opportunity to dishonestly accept a lost coin, asking them if the supposed lost coin was actually theirs. Out of 84 participants, 31 claimed the coin dishonestly. Using this 'dishonesty' measure, Farrington and Kidd (1977) were able to conclude that people would act more dishonestly when the coin was less valuable (10p versus 50p) and when the experimenter was female rather than male. Interestingly, the cost had no effect on dishonesty if the experimenter was female. In another field experiment, Farrington et al. (1980) interviewed youth and asked them to participate in a coin-sorting test with the implicit purpose of providing them with an opportunity to steal. The final results showed that 10 out of a total 25 participants stole

during the coin-sorting test. Interestingly, the boys who actually stole were not significantly more likely than the remainder to say that they would steal in a hypothetical situation.

In 1979, Farrington and Knight used the lost letter technique to study stealing. In this study, London pedestrians found an apparently lost unsealed stamped addressed letter containing information about the addressed person and a sum of money. The findings showed that people would more likely steal a letter that contained cash (compared with control letters containing no cash) if the addressed victim was a higher class male (compared with an old lady), and younger people were more likely to steal. In a follow-up study, Farrington and Knight (1980) used the lost letter technique and found that people were more likely to steal the letter if the addressed victim was male. Younger participants were more likely to steal the letter, but the participant's sex did not influence stealing behaviour. Whether the victim was young or old, rich or poor, or an individual or an association had no significant influence on stealing behaviour. Overall, when the lost letter contained cash, 39% ($n = 112$) of letters were not returned. Once again, the authors expressed their concerns about the external validity of the lost letter methodology to generalize for other types of stealing, a topic that they thought should be empirically tested.

Conclusions

On the one hand, criminologists agree that crime is inherently difficult to measure and that it is virtually impossible to achieve a perfect measurement of criminal behaviour (Krohn et al., 2012; Osgood et al., 2002). On the other hand, the challenge of measuring crime is not very different from the challenges of measuring many other aspects of human behaviour. The literature reviewed in the present article shows that measures of offending are limited and in some cases deeply flawed. However, criminologists have come a long way, and the acquired knowledge about criminal behaviour has made criminology an essential science in all societies.

In this article, we presented evidence that some measures provide more valid results than others. However, it is not argued that criminological research should be solely based on one measurement method. Clearly, different measurement techniques may be more suitable to different research questions. As described by Maxfield and Babbie (2009), official records of crime are a better method for measuring serious crimes, such as murder, as well as crimes in which the victim is a business or a commercial establishment. On the other hand, SRO measure more accurately crimes that do not have readily identifiable victims, such as drug offences. Official records and victim surveys may be more suitable for measuring crime rates of areas, whereas SRO may be more suitable for measuring offending behaviour by individuals. Official records offer more precise dates of offending and a broader range of illicit

behaviours, whereas SRO provide a better estimation of the real number of offences and other crime variables, such as age of onset, criminal desistance, etc. Researchers should be aware of the advantages and disadvantages of the crime measurement techniques and select the research method that best fits their specific research questions.

All measures of crime have considerable merit, and using a variety of methods may very well prove to be the best way to fully understand the phenomenon. Nevertheless, one clear conclusion of this review article is the great need for more methodological research, which is crucial to the development of criminology as a science. Many of the limitations and methodological concerns presented in this article have been well known to researchers for a long time. However, we still do not fully understand the underlying mechanisms that explain these limitations, how they interact with other variables, to what extent they cast doubt on or invalidate criminological knowledge or, in some cases, even if they actually exist.

Recognizing the limitations of our measurement techniques, which was one of the main objectives of this article, should not inhibit researchers from developing scientific studies to test hypotheses, theories, and programmes. On the contrary, since measurement is the basis of all sciences, knowledge about methodological limitations and strengths will greatly improve the validity of the conclusions that can be derived from our studies. Criminology must be viewed as a science, with knowledge based on solid evidence. Therefore, as is true for all sciences, methodological research in criminology is one of the most important aspects in achieving valid cumulative knowledge about offending.

For that reason, researchers should focus their efforts on improving SRO, on understanding the role of testing and panel effects in longitudinal studies, as well as on using experimental designs to test theoretical hypotheses. To improve the self-reported methodology, it is crucial to determine the impact of modes of administration and questionnaire design on SRO. The best way would be to apply the experimental method and randomly allocate participants to different conditions (e.g., interview versus paper-and-pencil questionnaire; questionnaires with follow-up questions versus without follow-up questions). To control which condition might be subject to underreporting or overreporting, researchers should collect physiological data, such as blood, saliva or hair samples. This physiological data would provide an external criterion with which it would be possible to compare the rates of agreement with self-reports of substance use in the different experimental conditions and determine which condition offers the most valid results.

In their review of SRO within longitudinal studies, Krohn et al. (2012) proposed an interesting study design to determine the extent of testing and panel effects. Basically, this is “a longitudinal study

in which the sample is randomly divided into groups, with some groups receiving all assessments starting at time T and others entering the panel at later assessments, $T + 1$, $T + 2$, and so on. If there are systematic differences in responses at $T + 1$, or at subsequent assessments, it would provide direct evidence of testing effects" (Krohn et al., 2012, p. 32). Furthermore, we think that the use of an experimental design, randomly allocating participants to different modes of administration and/or questionnaire designs, would make it possible to conclude whether any differences are a result of testing effects or panel effects.

Finally, we suggest that researchers in criminology should rely more on field experiments to test their hypotheses. In the last 35 years, very few experiments of this kind have been carried out in criminology or psychology, but they have been carried out in behavioural economics (see e.g., Rosenbaum et al., 2014). With regard to measuring offending, the ability to observe participants' offending behaviour as it happens (as in the experiments reviewed above) provides a level of validity that other methods cannot achieve. With regard to the scientific method, adopting the experimental design would make it possible to carry out sets of solid and replicable field experiments to test the foundations of criminological theories. Perhaps the way we think of criminological experiments today makes it hard to include the most serious and violent offending behaviours, but dishonesty and offenses such as stealing and vandalism can be studied in real-life experiments (Farrington, 2008). The basis of all sciences is systematic observation and experimentation, and this should also be at the heart of criminology.

References

- Ahonen, L., Loeber, R., Farrington, D. P., Hipwell, A. E., & Stepp, S. D. (2017). What is the hidden figure of delinquency in girls? Scaling up from police charges to self-reports. *Victims and Offenders, 12*(5), 761-776. <https://doi.org/10.1080/15564886.2016.1185486>
- Blumstein, A., Cohen, J., & Farrington, D. P. (1988). Criminal career research: Its value for criminology. *Criminology, 26*(1), 1-35. <https://doi.org/10.1111/j.1745-9125.1988.tb00829.x>
- Blumstein, A., Cohen, J., Roth, J. A., & Visher, C. A. (Eds.). (1986). *Criminal careers and "career criminals"* (Vol. 1). National Academy Press. <https://doi.org/10.17226/922>
- Bogen, D. (1944). Juvenile delinquency and economic trend. *American Sociological Review, 9*(2), 178-184. <https://doi.org/10.2307/2086310>
- Bosick, S. J. (2009). Operationalizing crime over the life course. *Crime and Delinquency, 55*(3), 472-496. <https://doi.org/10.1177/0011128707307223>
- Buckle, A., & Farrington, D. P. (1984). An observational study of shoplifting. *British Journal of Criminology, 24*(1), 63-73. <https://doi.org/10.1093/oxfordjournals.bjc.a047425>
- Buckle, A., & Farrington, D. P. (1994). Measuring shoplifting by systematic observation: A replication study. *Psychology, Crime and Law, 1*(2), 133-141. <https://doi.org/10.1080/10683169408411946>
- Buckle, A., Farrington, D. P., Burrows, J., Speed, M., & Burns-Howell, T. (1992). Measuring shoplifting by repeated systematic counting. *Security Journal, 3*(3), 137-146. <https://www.criminologycyprus.org/wp-content/uploads/2015/12/buckle92.pdf>
- Cops, D., Boeck, A. D., & Pleysier, S. (2016). School vs. mail surveys: Disentangling selection and measurement effects in self-reported juvenile delinquency. *European Journal of Criminology, 13*(1), 92-110. <https://doi.org/10.1177/1477370815608883>
- Denniston, M. M., Brener, N. D., Kann, L., Eaton, D. K., McManus, T., Kyle, T. M., Roberts, A. M., Flint, K. H., & Ross, J. G. (2010). Comparison of paper-and-pencil versus Web administration of the Youth Risk Behavior Survey (YRBS): Participation, data quality, and perceived privacy and anonymity. *Computers in Human Behavior, 26*(5), 1054-1060. <https://doi.org/10.1016/j.chb.2010.03.006>
- Durea, M. A. (1935). Mental and social maturity in relation to certain indicators of the degree of juvenile delinquency. *Child Development, 6*(2), 154-160. <https://doi.org/10.2307/1125498>
- Enzmann, D. (2013). The impact of questionnaire design on prevalence and incidence rates of self-reported delinquency: Results of an experiment modifying the ISRD-2 questionnaire. *Journal of*

Contemporary Criminal Justice, 29(1), 147-177.
<https://doi.org/10.1177/1043986212470890>

- Farrington, D. P. (1973). Self-reports of deviant behavior: Predictive and stable? *Journal of Criminal Law and Criminology*, 64(1), 99-110. <https://doi.org/10.2307/1142661>
- Farrington, D. P. (1977). The effects of public labelling. *British Journal of Criminology*, 17(2), 112-125. <https://doi.org/10.1093/oxfordjournals.bjc.a046802>
- Farrington, D. P. (2003). Key results from the first 40 years of the Cambridge Study in Delinquent Development. In T. P. Thornberry, & M. D. Krohn (Eds.), *Taking stock of delinquency: An overview of findings from contemporary longitudinal studies* (pp. 137-183). Kluwer/Plenum. <https://doi.org/10.1007/b105384>
- Farrington, D. P. (2008). Criminology as an experimental science. In C. Horne & M. J. Lovaglia (Eds.), *Experiments in criminology and law: A research revolution* (pp. 175-179). Rowman and Littlefield.
- Farrington, D. P., Bowen, S., Buckle, A., Burns-Howell, T., Burrows, J., & Speed, M. (1993). An experiment on the prevention of shoplifting. In R. V. Clarke (Ed.), *Crime prevention studies* (Vol. 1, pp. 93-119). Criminal Justice Press.
- Farrington, D. P., & Jolliffe, D. (2004). England and Wales. In D. P. Farrington, P. A. Langan, & M. Tonry (Eds.), *Cross-national studies in crime and justice* (pp. 1-38). Bureau of Justice Statistics. <http://www.ojp.usdoj.gov/bjs>
- Farrington, D. P., Jolliffe, D., Hawkins, J. D., Catalano, R. F., Hill, K. G., & Kosterman, R. (2003). Comparing delinquency careers in court records and self-reports. *Criminology*, 41(3), 933-958. <https://doi.org/10.1111/j.1745-9125.2003.tb01009.x>
- Farrington, D. P., Jolliffe, D., Loeber, R., & Homish, D. L. (2007). How many offenses are really committed per juvenile court offender?. *Victims and Offenders*, 2(3), 227-249. <https://doi.org/10.1080/15564880701403934>
- Farrington, D. P., & Kidd, R. F. (1977). Is financial dishonesty a rational decision? *British Journal of Social and Clinical Psychology*, 16(2), 139-146. <https://doi.org/10.1111/j.2044-8260.1977.tb00209.x>
- Farrington, D. P., Knapp, W. S., Erickson, B. E., & Knight, B. J. (1980). Words and deeds in the study of stealing. *Journal of Adolescence*, 3(1), 35-49. [https://doi.org/10.1016/S0140-1971\(80\)80011-X](https://doi.org/10.1016/S0140-1971(80)80011-X)

- Farrington, D. P., & Knight, B. J. (1979). Two non-reactive field experiments on stealing from a 'lost' letter. *British Journal of Social and Clinical Psychology*, *18*(3), 277-284. <https://doi.org/10.1111/j.2044-8260.1979.tb00337.x>
- Farrington, D. P., & Knight, B. J. (1980). Stealing from a "lost" letter: Effects of victim characteristics. *Criminal Justice and Behavior*, *7*(4), 423-436. <https://doi.org/10.1177/009385488000700406>
- Farrington, D. P., Langan, P. A., & Tonry, M. (Eds.). (2004b). *Cross-national studies in crime and justice*. U.S. Bureau of Justice Statistics.
- Farrington, D. P., Langan, P. A., Tonry, M., & Jolliffe, D. (2004a). Introduction. In D. P. Farrington, P. A. Langan, & M. Tonry (Eds.), *Cross-national studies in crime and justice* (pp. iii-xiv). U.S. Bureau of Justice Statistics.
- Farrington, D. P., Piquero, A. R., & Jennings, W. G. (2013). *Offending from childhood to late middle age: Recent results from the Cambridge Study in Delinquent Development*. Springer. <https://doi.org/10.1007/978-1-4614-6105-0>
- Farrington, D. P., Ttofi, M. M., Crago, R. V., & Coid, J. W. (2014). Prevalence, frequency, onset, desistance and criminal career duration in self-reports compared with official records. *Criminal Behaviour and Mental Health*, *24*(4), 241-253. <https://doi.org/10.1002/cbm.1930>
- Gold, M. (1966). Undetected delinquent behavior. *Journal of Research in Crime and Delinquency*, *3*(1), 27-46. <https://doi.org/10.1177/002242786600300103>
- Greenwald, G. (2009). *Drug decriminalization in Portugal: Lessons for creating fair and successful drug policies*. Cato Institute. <http://dx.doi.org/10.2139/ssrn.1464837>
- Hindelang, M. J., Hirschi, T., & Weis, J. G. (1979). Correlates of delinquency: The illusion of discrepancy between self-report and official measures. *American Sociological Review*, *44*(6), 995-1014. <https://doi.org/10.2307/2094722>
- Hindelang, M. J., Hirschi, T., & Weis, J. G. (1981). *Measuring delinquency*. Sage.
- Hirschi, T. (1969). *Causes of delinquency*. Transaction.
- Hirschi, T., & Gottfredson, M. (1983). Age and the explanation of crime. *American Journal of Sociology*, *89*(3), 552-584. <https://doi.org/10.1086/227905>
- ISRD3 Working Group. (2013). *Questionnaire ISRD3: Standard student questionnaire* (ISRD3 Technical Report Series #2). https://web.northeastern.edu/isrd/wp-content/uploads/2016/01/ISRD3_TechRep_2.pdf

- Jennings, W. G., Loeber, R., Ahonen, L., Piquero, A. R., & Farrington, D. P. (2018). An examination of developmental patterns of chronic offending from self-report records and official data: Evidence from the Pittsburgh Girls Study (PGS). *Journal of Criminal Justice, 55*, 71-79. <https://doi.org/10.1016/j.jcrimjus.2017.12.002>
- Jolliffe, D., & Farrington, D. P. (2014). Self-reported offending: Reliability and validity. In G. Bruinsma, & D. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice* (pp. 4716-4723). Springer. https://doi.org/10.1007/978-1-4614-5690-2_648
- Kazemian, L., & Farrington, D. P. (2005). Comparing the validity of prospective, retrospective, and official onset for different offending categories. *Journal of Quantitative Criminology, 21*(2), 127-147. <https://doi.org/10.1007/s10940-005-2489-0>
- Konecni, V. J., Ebbesen, E. B., & Konecni, D. K. (1976). Decision processes and risk taking in traffic: Driver response to the onset of yellow light. *Journal of Applied Psychology, 61*(3), 359-367. <https://doi.org/10.1037/0021-9010.61.3.359>
- Krohn, M., Thornberry, T., Bell, K., Lizotte, A., & Phillips, M. (2012). Self-report surveys within longitudinal panel designs. In D. Gadd, S. Karstedt, & S. Messner (Eds.), *The Sage handbook of criminological research* (pp. 23-35). Sage. <https://dx.doi.org/10.4135/9781446268285.n2>
- 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*(4), 509-525. <https://doi.org/10.1007/s10940-010-9119-1>
- Lauritsen, J. L. (1998). The age-crime debate: Assessing the limits of longitudinal self-report data. *Social Forces, 77*(1), 127-154. <https://doi.org/10.1093/sf/77.1.127>
- Le Blanc, M., & Fréchette, M. (1989). *Male criminal activity from childhood through youth: Multilevel and developmental perspectives*. Springer-Verlag. <https://doi.org/10.1007/978-1-4612-3570-5>
- Loeber, R., Farrington, D. P., Hipwell, A. E., Stepp, S. D., Pardini, D., & Ahonen, L. (2015). Constancy and change in the prevalence and frequency of offending when based on longitudinal self-reports or official records: Comparisons by gender, race, and crime type. *Journal of Developmental and Life-Course Criminology, 1*(2), 150-168. <https://doi.org/10.1007/s40865-015-0010-5>
- Loeber, R., Farrington, D. P., & Petechuk, D. (2003). *Child delinquency: Early intervention and prevention* (NCJ 186182). U.S. Office of Juvenile Justice and Delinquency Prevention. <https://www.ojp.gov/pdffiles1/ojdp/186162.pdf>
- Lucia, S., Herrmann, L., & Killias, M. (2007). How important are interview methods and questionnaire designs in 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(1), 39-64. <https://doi.org/10.1007/s11292-007-9025-1>
- Maxfield, M. G., & Babbie, E. R. (2009). *Basics of research methods for criminal justice and criminology* (2nd ed.). Cengage Learning.
- McLaughlin, E., & Muncie, J. (Eds.). (2001). *The Sage dictionary of criminology*. Sage.
- Murphy, F. J., Shirley, M. M., & Witmer, H. L. (1946). The incidence of hidden delinquency. *American Journal of Orthopsychiatry*, 16(4), 686–696. <https://doi.org/10.1111/j.1939-0025.1946.tb05431.x>
- Nye, F. I., Short, J. F., & Olson, V. J. (1958). Socioeconomic status and delinquent behavior. *American Journal of Sociology*, 63(4), 381-389. <https://doi.org/10.1086/222261>
- Osgood, D. W., McMorris, B. J., & Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitative Criminology*, 18(3), 267-296. <https://doi.org/10.1023/A:1016008004010>
- Piquero, A. R., Schubert, C. A., & Brame, R. (2014). Comparing official and self-report records of offending across gender and race/ethnicity in a longitudinal study of serious youthful offenders. *Journal of Research in Crime and Delinquency*, 51(4), 526-556. <https://doi.org/10.1177/0022427813520445>
- Porterfield, A. L. (1943). Delinquency and its outcome in court and college. *American Journal of Sociology*, 49(3), 199-208. <https://doi.org/10.1086/219369>
- Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology*, 45, 181-196. <https://doi.org/10.1016/j.joep.2014.10.002>
- Sanches, C., Gouveia-Pereira, M., Marôco, J., Gomes, H. S., & Roncon, F. (2016). Deviant behavior variety scale: Development and validation with a sample of Portuguese adolescents. *Psicologia: Reflexão e Crítica*, 29(31), 1-8. <https://doi.org/10.1186/s41155-016-0035-7>
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54(2), 93-105. <https://doi.org/10.1037/0003-066X.54.2.93>
- Schwarz, N., Strack, F., Hippler, H. J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, 5(3), 193-212. <https://doi.org/10.1002/acp.2350050304>
- Sellin, T. (1931). The basis of a crime index. *Journal of Criminal Law and Criminology*, 22(3), 335-356. <https://doi.org/10.2307/1135784>

- Sellin, T. (1938). *Culture conflict and crime: A report of the subcommittee on delinquency of the Committee on Personality and Culture*. Social Science Research Council.
- Theobald, D., Farrington, D. P., Loeber, R., Pardini, D. A., & Piquero, A. R. (2014). Scaling up from convictions to self-reported offending. *Criminal Behaviour and Mental Health, 24*(4), 265-276. <https://doi.org/10.1002/cbm.1928>
- Thornberry, T. P. (1989). Panel effects and the use of self-reported measures of delinquency in longitudinal studies. In M. W. Klein (Ed.), *Cross-national research in self-reported crime and delinquency* (pp. 347-369). Springer. https://doi.org/10.1007/978-94-009-1001-0_16
- Thornberry, T. P., & Krohn, M. D. (2000). The self-report method for measuring delinquency and crime. In D. Duffee (Ed.), *Measurement and analysis of crime and justice* (pp. 33–84). U.S. National Institute of Justice.
- Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60*(2), 275-304. <https://doi.org/10.1086/297751>
- West, D. J., & Farrington, D. P. (1977). *The delinquent way of life*. Heinemann.
- Wilson, J. Q., & Herrnstein, R. J. (1985). *Crime and human nature*. Free Press.
- Wright, D. L., Aquilino, W. S., & Supple, A. J. (1998). A comparison of computer-assisted and paper-and-pencil self-administered questionnaires in a survey on smoking, alcohol, and drug use. *Public Opinion Quarterly, 62*(3), 331-353. <https://doi.org/10.1086/297849>

CHAPTER II

FIELD EXPERIMENTS ON DISHONESTY AND STEALING:
WHAT HAVE WE LEARNED IN THE LAST 40 YEARS?

Manuscript Published in:

Gomes, H. S., Farrington, D. P., Defoe, I. N., & Maia, Â. (2021). Field experiments on dishonesty and stealing: What have we learned in the last 40 years?. *Journal of Experimental Criminology*. Advance online publication. <https://doi.org/10.1007/s11292-021-09459-w>

FIELD EXPERIMENTS ON DISHONESTY AND STEALING: WHAT HAVE WE LEARNED IN THE LAST 40 YEARS?

Abstract

Objectives. Field experiments combine the benefits of the experimental method and the study of human behavior in real-life settings, providing high internal and external validity. This article aims to review the field experimental evidence on the causes of offending.

Methods. We carried out a systematic search for field experiments studying stealing or monetary dishonesty reported since 1979.

Results. The search process resulted in 60 field experiments conducted within multiple fields of study, mainly in economics and management, which were grouped into four categories: Fraudulent/ dishonest behavior, Stealing, Keeping money, and Shoplifting.

Conclusions. The reviewed studies provide a wide variety of methods and techniques that allow the real-world study of influences on offending and dishonest behavior. We hope that this summary will inspire criminologists to design and carry out realistic field experiments to test theories of offending, so that criminology can become an experimental science.

Keywords: Field experiments; Naturalistic experiments; Stealing; Dishonesty; Systematic review

Introduction

The main aim of this article is to encourage criminologists to carry out naturalistic field experiments to investigate the causes of offending. Theories of offending are usually tested in cross-sectional or longitudinal studies. However, in trying to isolate the effect of a particular variable on offending, these methods can only attempt to control for other measured extraneous influences. Because of numerous unknown and unmeasured variables that might influence offending, these methods have low internal validity. In contrast, a randomized field experiment that manipulates influences on offending has higher internal validity, because the randomized design's logic controls for all measured and unmeasured extraneous influences on offending, providing that a large number of units are randomly assigned (Weisburd, 2003). Information that is relevant to criminological theories can and should be drawn from the many experiments on prevention and intervention in criminology (see e.g., Robins, 1992), but conclusions can be drawn more directly by testing theories in naturalistic field experiments.

This article presents a systematic review of field experiments on dishonesty and stealing that have been published in the 40 years since the seminal review by Farrington (1979). Remarkably, most of these experiments have been carried out by economists rather than by criminologists, and most have been designed to test ideas of rational decision-making influenced by subjectively expected benefits, costs, and probabilities. We believe that most criminologists are not familiar with this body of knowledge from the economics literature, and so we present summaries of all the experiments. We hope that these summaries will inspire criminologists to design and carry out realistic field experiments to test theories of offending, so that criminology can become a more experimental science (see e.g., Farrington, 2008).

Experimental approach

Experiments are the most important technique in developing scientific knowledge. The experimental approach implies the manipulation of variables under strictly controlled situations, which allows for the study of cause-and-effect relationships to provide unambiguous conclusions about the variables that influence behavior. The potential of experiments is very important in the study of criminal behavior because they can provide conclusive evidence about factors influencing offending, as well as predicting and preventing future offending behaviors. However, the vast majority of research in the field of criminology and criminal behavior is nonexperimental.

Many researchers have pointed out the limitations of the experimental approach, mainly referring to the artificiality of laboratory settings which may contaminate the experiment, yielding inconclusive

results that are not easily generalizable to the real world. Field experiments, on the other hand, are very useful techniques that overcome these limitations by testing cause-and-effect relationships in real-world settings. In 1979, Farrington carried out a review of field experiments on deviance, urging researchers to carry out more of these realistic experiments. Recently, despite the limited number of field experiments developed in the field of criminology, several realistic field experiments on stealing and dishonesty have been carried out by behavioral economists (Farrington et al., 2020). The goal of the present article is to systematically review the field experiments on deviant behavior that have been carried out in the last 40 years, after the publication of Farrington (1979). In doing so, we explore the experimental designs, measurement techniques, and main findings of relevant studies in the interest of increasing the use of this robust technique in criminology.

The scientific process or method is a systematic approach to acquiring knowledge that, through objective observation and hypothesis testing, enables an ever-growing body of knowledge (Christensen, 1985). Within the multiple types of studies and tools that constitute the scientific method, the experimental approach stands out because it allows for the testing of cause-and-effect relationships. Experiments can be described as “objective observation of phenomena which are made to occur in a strictly controlled situation in which one factor is varied and the others are kept constant” (Zimny, 1961, p. 35). The ability to control extraneous variables and precisely manipulate the independent variable (or variables) are key to arriving at unambiguous causal conclusions and provide pathways to ever more impactful treatments with fewer negative side effects, as well as cost-benefit estimations.

Laboratory versus field experiments

In this regard, laboratory experiments are the main experimental technique, since they maximize control. Therefore, the primary advantage of laboratory experiments is internal validity. In the laboratory, researchers are able to account for and minimize the influence of extraneous stimuli in an attempt to control the effect of environmental factors irrelevant to the study. However, the gains in internal validity conferred by the laboratory control come at the cost of artificial and sterile settings. This, in turn, may influence the results and reduce the study’s external validity, limiting the relevance for predicting behavior in the field as well as generalizability to the real world (Farrington, 1980; Harrison & List, 2004).

On the other hand, field experiments are not subject to this artificiality problem, since they are carried out in real-life settings. Therefore, the main advantage of field experiments is external validity. Compared with cross-sectional and longitudinal studies, field experiments have high internal validity. However, the limited ability in some cases to control for extraneous variables in naturalistic environments

may cause a reduction in the internal validity of field experiments (Christensen, 1985). In some field experiments, this lack of control over the factors influencing behavior opens the possibility for alternative explanations, which may compromise the study of causal relationships (Pierce & Balasubramanian, 2015).

A further potential disadvantage of field experiments is selection bias in the random selection of participants (Christensen, 1985). For example, a field experiment designed to study dishonest behavior of people buying journals may be affected by selection bias, since this sample (i.e., journal customers) may not represent the population of interest, namely, the offender population (e.g., Pruckner & Sausgruber, 2013). However, laboratory research may also be subject to selection bias because experiments, especially in psychological and social science research, are generally carried out with undergraduate students as participants, further limiting the external validity of laboratory experiments (Farrington, 1979).

Finally, researchers must consider that in laboratory experiments, people are aware that their behavior is being scrutinized. This makes laboratory experiments subject to multiple sources of bias, such as social desirability, thus compromising their internal validity (Levitt & List, 2007). This is especially relevant in the study of deviance. The study of deviant behavior brings about additional concerns, because it is a highly sensitive topic that people try to conceal, possibly due to guilt, shame, or fear of repercussions (Gomes et al., 2019). Taking this into account, naturalistic field experiments on deviance, carried out in real-life contexts in which participants are unaware that their behavior is being studied, may offer the greatest internal and external validity of all methods (Farrington, 1979).

Field experiments in the study of deviance

Despite the apparent consensus on the relevance of experiments in the development of criminological knowledge and crime prevention practice, most research on deviance is nonexperimental, and naturalistic field experiments are still scarce in social science (Franzen & Pointner, 2013; Gomes et al., 2018). A quick search for the terms “*crim* OR delinq**” in the Scopus database (i.e., article title, abstract, and keywords) results in a total of 267,523 documents up to 2018. On the other hand, the same search including the term “*experiment*” results in a total of 11,005 documents, which represents 4.11% of the studies. The same search for “*field experiment*,” however, results in only 239 documents, which represents 2.17% of all experiments and less than 0.1% (0.09%) of criminological research. Hence, field experiments on deviance are sorely needed.

In 1979, Farrington carried out a pioneering review of field experiments on deviance, with special reference to dishonesty. In that review, studies were included where members of the public were given the opportunity to dishonestly claim money, referring to such techniques as the lost coin where the experimenters pretend to pick up money (e.g., Farrington & Kidd, 1977; Feldman, 1968; Korte & Kerr, 1975) or leave coins in a telephone booth (e.g., Bickman, 1971; Franklin, 1973); experiments that provided opportunities for members of the public to engaged in offending behavior, for example, theft of candies (e.g., Diener et al., 1976), theft of shampoo out of a purposely forgotten expensive shampoo bottle (Steinberg et al., 1977), taking bags without paying (Lenga & Kleinke, 1974), and stealing money out of lost letters and/or wallets (e.g., Farrington & Knight, 1979, 1980; Hornstein et al., 1968). However, Farrington (1979) noted that, despite the wide variety of deviance that was studied, there were no studies on vandalism or property damage. He mentioned the famous study by Zimbardo (1969) but concluded that it did not meet the criteria for an “experiment” because of its inadequate control of independent and extraneous variables.

Theoretical framework: factors influencing deviance

Farrington (1979) proposed that engaging in the above-described dishonest behaviors can be considered a risky decision-making process. Therefore, a relevant specific theory would include the evaluation of the benefits and costs that follow from the choice to commit dishonest behavior (Farrington, 1979). Hence, in Farrington’s work (1979), the subjective expected utility (SEU) perspective was used as the main theoretical framework (see also Farrington & Knight, 1980). The SEU theory suggests that, in situations of risk (i.e., uncertainty), a decision about the alternative choices is based on (1) utility (i.e., subjective benefit or attractiveness), (2) subjective costs, and (3) their associated probabilities. Thus, each alternative choice has a total SEU, and, in the end, the decision-maker chooses the option with the highest SEU (Farrington & Knight, 1980). At the same time, Farrington (1979) also noted that solely focusing on costs and benefits is too simplistic to predict complex human behavior such as deviance. However, it is useful to start off with a simple and testable theory, and identify which results cannot be explained by it to subsequently determine in which ways it needs modifying or extending, rather than starting off with a complex theory that is less testable.

Accordingly, in the current review, we explore whether the manipulation of benefits and costs predicted dishonest behavior in field experiments. Similar to Farrington (1979), in the current review, financial gains in some form are regarded as “benefits for the perpetrator.” Additionally, factors such as the suffering of other persons (victims) because of the actions of the perpetrator are regarded as “costs

for the other.” Of note, Farrington (1979) described conditions where the victims were less deserving (e.g., stealing from a young rich person) as less unpleasant and thus “low cost,” whereas conditions where victims were more deserving (e.g., stealing from an old poor person) were regarded as more unpleasant and thus “high cost.” We use the same definitions in the current review. Finally, Farrington (1979) also demonstrated that the likelihood of apprehension (i.e., *costs for the perpetrator*) is also a relevant predictor of deviance. In the current review, we divide costs into *costs for the other* (i.e., costs for the victim) and *costs for the self* (i.e., costs for the perpetrator).

The present study

Farrington (1979) highlighted the benefits of naturalistic experimentation and expressed his wish “that psychologists will have the ingenuity, determination, and social responsibility to meet the challenge of experiments on deviance” (Farrington, 1979, p. 242).

In order to provide criminology researchers with an updated review of the field experimental evidence relevant to the study of deviance, the present article aims to systematically review field experiments seeking to study the causes of offending or monetary dishonesty that have been reported since the review of Farrington (1979). We focus on field experiments on deviance that included financial dishonesty, as this overlaps most with an experimental way of studying delinquency (cf., Farrington, 1979). Unlike the review of Farrington (1979), the current review only includes studies with deviance as an outcome measure (whereas Farrington, 1979 also included studies that investigated deviance as an independent variable).

In order to provide relevant information to researchers who might consider developing field experiments to test their hypotheses, the present review of field experiments on deviance will focus on the methods used to assess participants’ deviant or dishonest behavior in the field. Moreover, inspired by Farrington (1979), who concluded that many field experiments were motivated by cost-benefit theories such as SEU, we additionally coded the studies on whether they investigated independent variables that are related to benefits and costs (i.e., *costs for the self* and *costs for the other*). In other words, we explore whether studies that manipulated these benefit and cost variables found that increases in benefits increases deviance, while decreases in costs increases deviance.

Methods

Search strategy

The search for field experiments was carried out in four different steps: (1) a systematic search in general databases; (2) a systematic search in specialized journals; (3) a reference search; and (4) a citation search. All searches were carried out in July 2019. Regarding the first step, we entered the following keywords (“field experiment” or “naturalistic experiment”) and (“steal*” or “dishonest*” or “theft” or “shoplift*”) in several relevant data bases, i.e., Scopus, EBSCO, PubMed, Web of Science, Google Scholar, ProQuest, and Ethos.

Secondly, we searched the same keywords in several specialized journals that publish field experimental findings in the fields of criminology, psychology, and economics, i.e., *Journal of Economic Behavior and Organization*, *Journal of Economic Psychology*, *Journal of Organizational Behaviour Management*, *Journal of Behavioral and Experimental Economics*, *Experimental Economics*, *Journal of Public Economics*, *Journal of Social Psychology*, *Journal of Applied Psychology*, *Journal of Applied Social Psychology*, *Journal of Experimental Social Psychology*, *Personality and Social Psychology Bulletin*, *Psychological Science*, *Journal of Experimental Criminology*, and *Security Journal*.

In the third step of this systematic search process, we reviewed the articles found in the previous steps and carried out a search of their references. Taking into consideration the large amount of referenced material, we searched for any of the previous keywords in the articles’ titles (i.e., “field” OR “naturalistic” OR “experiment” OR “steal*” OR “dishonest*” OR “theft” OR “shoplift*”). The studies included in this third process were also subject to a second sweep of the reference search process, where we repeated the same process of searching the titles of the referenced materials.

Finally, in the citation process, we carried out a citation search in order to identify all studies that have cited the studies included in the previous steps. Similarly to the reference search step, in order to deal with the large amount of cited articles, we conducted a search of the keywords in the articles’ titles (i.e., “field” OR “naturalistic” OR “experiment” OR “steal*” OR “dishonest*” OR “theft” OR “shoplift*”).

Inclusion criteria

In the present review of field experimental evidence relevant to the study of deviance, we have included all published and unpublished field experiments seeking to study the causes of offending or monetary dishonesty reported since 1979 in English, French, Spanish, or Portuguese, that met the two inclusion criteria described below.

1. Field experiment

As described above, we have used List's (2007) definition of a field experiment, i.e., an experimental study carried out in the natural environment. Taking into account this definition of a field experiment, we have included in this review all experimental studies that used members of the public in the real world who were unaware that their deviant behavior was being assessed. Studies conducted in laboratory settings or where participants are aware that their deviant behavior was being measured were not included.

2. Studied the causes of offending or dishonesty

In order to review research designs relevant to the study of criminology, we have included studies with measures of offending or monetary dishonesty. Within this definition, we have considered field experiments that included measures of behaviors that ranged from stealing to acting dishonestly in order to obtain goods, whether it may be money or any other item (e.g., candy, flowers, newspaper). On the other hand, we have excluded experiments studying other types of deviant behaviors such as littering (e.g., Ramos & Torgler, 2012), illegal disposal of household garbage (e.g., Dur & Vollaard, 2019), and jaywalking (e.g., study 2 of Keuschnigg & Wolbring, 2015). Furthermore, since in this article we are mostly interested in the study of the causes of offending, other criminological field experiments or interventions, such as the ones on hot spots (e.g., Weisburd, 2005), were not included.

Additionally, whenever we found an unpublished study, such as a doctoral thesis (e.g., Korbelt, 2013), that was later published (e.g., Chytilová & Korbelt, 2014), we gave preference to the published version of the field study and treated the unpublished version as a repeated study. In the same manner, studies that used the same sample to study similar hypotheses were treated as repeated and only the first publication was entered in this review (e.g., Armantier & Boly, 2011, 2013).

Search for eligible studies

1. Systematic search in general databases

The search equation for studies in the systematic search in general databases resulted in a total of 383 studies. After eliminating all repeated studies ($k = 99$), a total of 284 studies underwent the screening for inclusion criteria. As illustrated in Figure 3, 75 studies did not report field experimental evidence, and 177 studies lacked a measure of deviance and were excluded. Additionally, two studies meeting the inclusion criteria (Farrington & Knight, 1979, 1980) were already analyzed in the original

review of Farrington (1979) and were not included in the present review. In sum, a total of 30 studies found in the first systematic search met the inclusion criteria and were included in the present review.

2. Systematic search in specialized journals

In the second systematic search process we searched our keywords in specialized journals. This search resulted in a total of 164 studies. From these, 31 repeated studies were eliminated and a total of 133 studies forwarded to the eligibility search. Our search revealed 88 studies that failed the field experiment criteria and 39 failed to measure deviance. After excluding the studies that failed to meet the eligibility criteria, six more studies were included in our review.

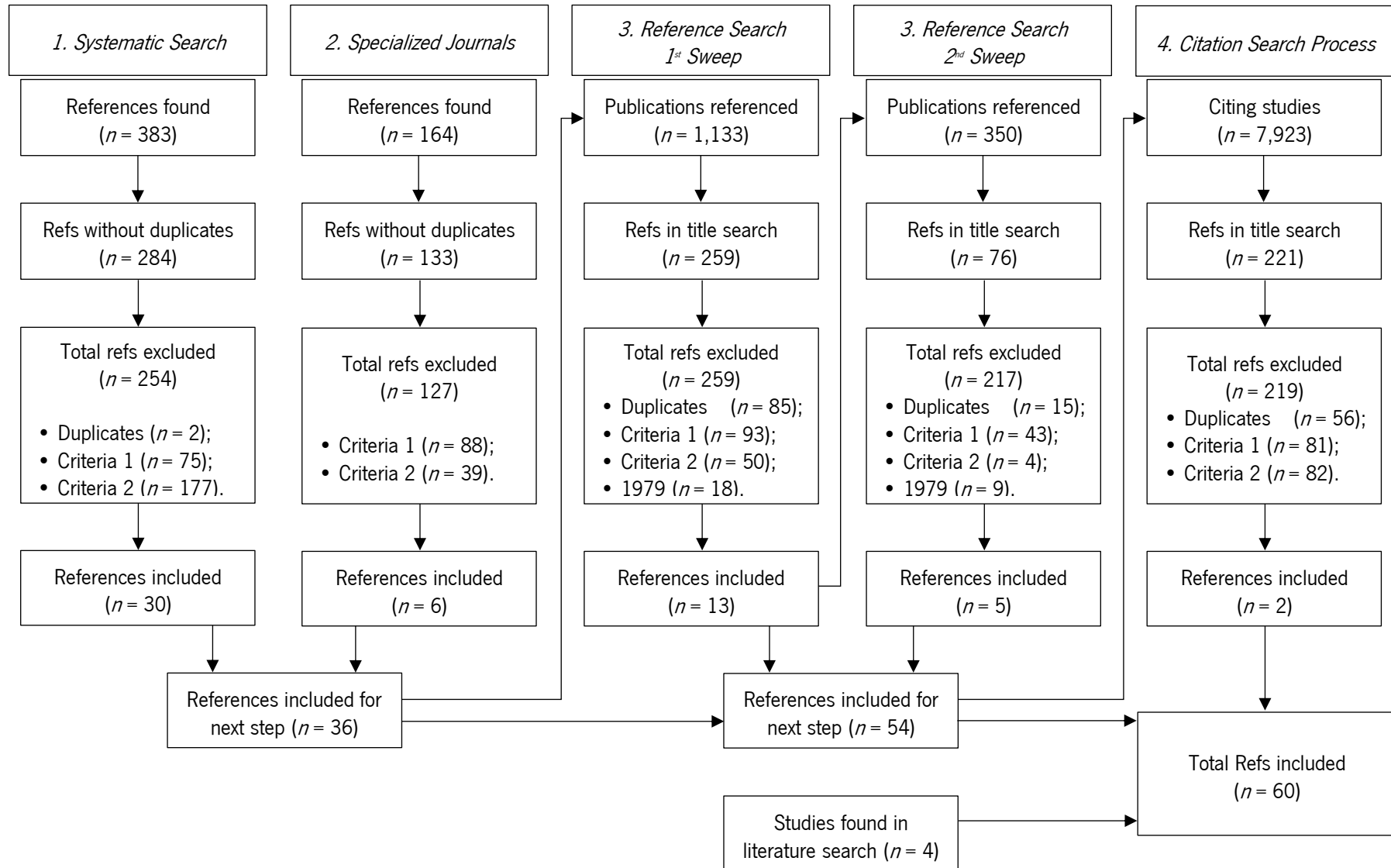
3. Reference search

In the two previous steps, we found a total of 36 field experiments. In the third step, we analyzed the studies referenced in these field experiments. As showed in Figure 3, a total of 1,133 studies were referenced. In order to make the search feasible, we carried out a search for our keywords in articles' titles. As a result, we found 259 referenced studies with at least one of the keywords in its title. Following this procedure, 85 repeated studies and 18 studies dated before 1979 were removed. The analysis for the remaining studies resulted in the removal of 93 studies for failing to meet criteria 1 and 50 for failing criteria 2. This resulted in a total of an additional 13 field experiments included in our study.

In order to maximize the number of field experiments in this review, we carried out a second sweep for referenced studies using the 13 newly found field experiments. This time, 350 studies were referenced, of which 76 included at least one of our keywords in the title. Out of the total 76 references, 15 studies were repeated, 9 were published before 1979, 43 failed criteria 1, and 4 failed criteria 2. Therefore, the second sweep resulted in the inclusion of five new field experiments in our review.

Figure 3

Flowchart of the systematic search processes



4. Citation search

In the final step, we considered all the 54 field experiments found in the previous steps and carried out a search for the studies that cited these experiments using Google Scholar (Figure 3). In all, 7,923 studies cited the field experiments included in the present review. Following the same procedures used in the previous step, we carried out a search for the keywords in the titles of the studies: “field” OR “naturalistic” OR “experiment” OR “steal*” OR “dishonest*” OR “theft” OR “shoplift*” and found 221 studies that were selected for criteria analysis. In this analysis, 56 studies were repeated, 81 failed the field experiment criteria, and 82 failed the measure of deviance criteria and were deleted. This resulted in the inclusion of two more field experiments in our systematic review.

Finally, four more studies (Cohn et al., 2019; Hayes & Downs, 2011; Hayes et al., 2011; Johns et al., 2017) were found during the literature review of field experiments. These studies failed to enter in any of the search processes considered in this article, but proved to be relevant field experiments for the present systematic review and were included. In conclusion, a total of 60 field experiments studying the causes of monetary dishonesty were included in our systematic review.

Results

Table 1 summarizes some key descriptive features of the studies included in this review. Field experiments reported results from multiple countries, including two studies that published multinational reports: List and Momeni (2017) included samples in the USA and India; and Cohn et al. (2019) included samples from 40 different countries. This resulted in 106 samples from a total of 44 different countries, where the European countries (including Russia $n = 1$ and Turkey $n = 1$) were the most frequently sampled (38.7%, $n = 41$), mostly represented by Germany ($n = 7$) and the UK ($n = 5$). The second most sampled continent was North America (33.0%, $n = 35$), mostly USA ($n = 28$), followed by Canada ($n = 6$) and one Mexican sample. Studies considering Asian samples (14.2%, $n = 15$) included mostly Israel ($n = 4$) and India ($n = 4$). A total of seven samples were considered from African countries, namely Burkina Faso, Ghana, Kenya, Morocco, Nigeria, South Africa, and Tanzania. South American (4.7%, $n = 5$) countries included Peru ($n = 2$), followed by Argentina, Brazil, and Chile. Finally, three samples from the Australian continent (2.8%) were considered, two from Australia and one from New Zealand.

Most field experiments included in the present review were published in English (98.3%, $k = 59$), with one exception published in French (i.e., Tremblay et al., 2000). Regarding the studies' date, as illustrated in Table 1, about half of the studies ($k = 28$) were reported from 1979 to 2009, whereas 32 (53.3%) field experiments were carried out in the last 10 years (i.e., 2010–2019). Most of these studies

were published in scientific journals (90.0%, $k = 54$), while the remaining six studies were in reports and doctoral dissertations. As for the discipline of these studies, half of the studies were carried out under or published in economics or management journals (50.0%, $k = 30$), while the remaining studies were in the psychology and social sciences ($k = 14$) or criminology ($k = 16$). Since we believe that few criminologists read economics or management journals, we think that it is important to communicate these studies to criminologists and encourage them to carry out field experiments on stealing and dishonesty. These experiments have been reviewed in an economics journal (Rosenbaum et al., 2014) but not recently, and not in a criminology journal.

Table 1

Descriptive information on 60 studies in the systematic review

Descriptive Variables	Categories	Frequency
Sample origin (44 countries, 106 samples)	Europe	41 (38.7%)
	North America	35 (33.0%)
	Asia	15 (14.2%)
	Africa	7 (6.6%)
	South America	5 (4.7%)
	Australia	3 (2.8%)
Study language	English	59 (98.3%)
	French	1 (1.7%)
Report Date	1979/80s	13 (21.7%)
	1990s	6 (10.0%)
	2000s	9 (15.0%)
	2010s	32 (53.3%)
Publications	Journal articles	54 (90.0%)
	Reports	3 (5.0%)
	Dissertations	3 (5.0%)
Field of study	Criminology	16 (26.7%)
	Economics / Management	30 (50.0%)
	Psychology / Social Sciences	14 (23.3%)

Furthermore, these field experiments presented multiple and creative methodologies in attempts to answer to different research questions that we were able to group into four different main topics, namely, fraudulent/dishonest behavior ($k = 21$), stealing ($k = 16$), keeping money ($k = 9$), and shoplifting ($k = 14$). Detailed information about all of these studies is presented in the results chapter.

Fraudulent/dishonest behavior

Within the fraud category, we have included field experiments that used a dependent variable related to illegal or dishonest behavior resulting in monetary or personal gain. This resulted in multiple types of measures of deviance, from low seriousness dishonesty such as sellers' overcharging or methods usually applied in laboratory experiments such as the coin toss or the dice roll tasks, to more serious offensive practices such as insurance fraud (see Table 2).

Five studies reported field experimental evidence related to overcharging. Balafoutas et al. (2013) carried out a naturalistic field experiment designed to study fraudulent behavior of taxi drivers. In this study, confederates posed as passengers and the taxi driver's perception about the passenger was manipulated by the way passengers spoke and dressed, showing different degrees of familiarity with the city. By using GPS data, researchers were able to precisely record the chosen route and compare it to an estimated correct fare for the given distance, with the difference measuring the amount of overcharging. They found that taxi drivers more frequently overcharged passengers unfamiliar with the city, taking them on an average detour that more than doubled the length of the journey of familiar passengers.

Conrads et al. (2015), as well as Dugar and Bhattacharya (2017), developed field experiments in order to study dishonesty in real-life pay-per-weight pricing markets. In these two studies, the purchased goods were weighted by the researchers after the transaction and the actual weight compared to the weight reported by the sellers. Conrads et al. (2015) employed this methodology to study overcharging occurring in candy stores. In this experiment, the authors found that overcharging occurred in 38% of purchases, though the apparent status of the buyer (high vs. low) and the quantity of candy bought (high vs. low) did not impact sellers' dishonesty. In the case of Dugar and Bhattacharya (2017), overcharging was studied in fish markets. Results showed that most sellers overcharged (89%). Moreover, these results also showed how overcharging varied as a function of the potential economic benefit (i.e., the type and size of fish).

The remaining two field experiments on overcharging (Jesilow & O'Brien, 1980; Schneider, 2012) had confederates visiting auto repair garages and submitting a test vehicle for repair with a prearranged set of defects. Findings from the study of Schneider (2012) showed that mechanics recommended

unnecessary repairs in 33% of visits. Moreover, when the researcher presented himself as one-time business, the total amount of repair cost increased significantly, compared with possible repeated business. Jesilow and O'Brien (1980) resorted to a similar methodology to study the effectiveness of deterrence interventions. In this experiment, the authors matched two areas by the degree of auto fraud and then subjected the experimental area to a deterrence intervention that included broadcasts of the existence of a state agency to which the public could report questionable repair dealers and a letter sent to the repair garages reminding them of the law and the consequences of violation. The opportunity for fraudulent behavior was created by having female confederates enter the repair facilities requesting the shops to test their car batteries (i.e., the "battery test"). Findings showed that the percentage of shops wrongly recommending a new battery in the post intervention phase was much higher in the control group compared to the intervention group.

Similar methodologies were employed to study insurance fraud. In the field experiment developed by Tracy and Fox (1989), confederates visited random auto body repair shops and obtained estimates of repair costs. In this case, experimenters manipulated whether the car was or was not being covered by insurance, as well as the sex of the driver. The results showed much higher repair estimates for insured vehicles, showing that the auto shops would inflate the prices in the insured condition. Furthermore, these results also showed, not only a sex-of-the-driver effect, where the estimated repair costs were much higher to female drivers, but also a sex-coverage interaction in which the male-female differences were even greater in the non-covered condition, suggesting that male drivers were better able to "get a break." More recently, Kerschbamer et al. (2016) also studied insurance fraud in computer repair shops. Confederates entered the repair shop and submitted manipulated test computers for repair. Results clearly showed a much higher average repair price when confederates were covered by insurance, compared to when they were not covered.

Taking into consideration that insured clients who are victims of theft have the opportunity to boost their losses in order to achieve monetary gains, three studies focused on insurance fraud to study the impact of deterrent letters on insurance customers (Blais & Bacher, 2007; Shu et al., 2012; Tremblay et al., 2000). Tremblay et al. (2000) manipulated whether the claimants received a deterrent or permissive letter, as well as whether the claim regulation was carried out on the telephone or by having regulators visiting the insurer's home. Findings showed main effects of claim regulation, where settling the claim over the telephone led to higher losses per claim. Interaction effects also showed that the permissive letter increased the average claim amounts only when the claims were settled remotely by

telephone, while the deterrent letter decreased the average amount claimed only in the face-to-face condition.

Blais and Bacher (2007) also applied measures of insurance fraud, in this particular case to study the effects of the threat of legal sanctions on offending behavior of insurance customers. In this study, insurance companies randomly assigned reports of property theft to the control (business as usual) or experimental group. Claimants in the experimental group received a deterrent letter reminding them of the sanctions associated with claim padding. Findings showed that the deterrent letter decreased the likelihood of claim padding. In Shu et al.'s (2012) field experiment, the authors manipulated the policy review form by making insurance customers report the current odometer mileage of their insured cars and sign it either at the beginning or at the end of the form. Seeing that a lower odometer mileage indicated a lower risk of accidents and, thus, lower insurance premiums, participants were expected to underestimate their car's mileage. Results of this field experiment showed that customers who signed at the beginning provided about 10% higher mileage estimates than those who signed at the end.

Nagin et al. (2002) designed a field experiment to study the effects of monitoring of employees' fraudulent behavior. Participants in this experiment were telephone solicitors at a call center, and their salary increased with the number of successful solicitations (i.e., contributions from potential donors). Given this incentive to claim higher solicited donations, the company monitored for falsely reported donations (i.e., "bad calls"). In this field experiment, the audit rate for bad calls that were reported back to employees was manipulated. Results showed that a perceived reduction in monitoring was quickly followed by more fraudulent behavior by employees in the number of bad calls.

List and Momeni (2017, 2020) carried out two field experiments to study workers' fraudulent behavior. In these field experiments, the authors employed online workers through MTurk (i.e., an online labor market platform) to perform a transcription task for payment. Workers had to transcribe 10 scanned images of short German texts. If the images were unreadable, workers could report and skip that image, moving on to the next image. This provided an opportunity for workers to misreport readable images as unreadable, allowing them to get the payment with less effort. A different way to behave fraudulently in this experiment was to take the upfront payment without completing the job. In the first experiment, List and Momeni (2017) paid 10% of the total payment upfront, and manipulated the total wage (i.e., \$0.90; \$1.20; \$1.26) and Corporate Social Responsibility (CSR) by making a charity donation on behalf of the firm or on behalf of the workers. Results showed that the decrease in the wage and the increase of the expenditure on CSR, especially when framed as a pro-social act on behalf of the workers, caused an increase in the number of employees acting dishonestly. In List and Momeni's (2020) second field

experiment, the authors followed the same design and manipulated the amount of upfront payment (i.e., 0%, 10%, 50%, and 90% of the total pay). Findings showed that, compared to the baseline condition, all conditions with upfront payment decreased dishonesty. On the other hand, within the upfront conditions, larger upfront payments were related to increases in the workers' dishonest behavior.

Olken (2007) developed a field experiment in order to study the impact of monitoring in the fraudulent behavior of villagers in Indonesia. In this experiment, funds were awarded to villages for the construction of roads. The information provided about government auditing (i.e., "external audits") and direct participation in the monitoring process by villagers was manipulated. In order to assess fraudulent behavior in the construction of roads, core samples of the roads after the projects were completed were dug up, and the quantity of materials used was estimated. The difference between the amount the village claimed to have spent on the project and the engineers' estimated price was the measure of fraudulent behavior in this study. Findings showed that, contrary to direct participation which did not affect village fraud, increasing the probability of external audits caused a substantial reduction in missing funds in the project.

Bertrand et al. (2007) carried out a field experiment in order to study whether the allocation of driver's licenses in India was influenced by a candidate's willingness to pay. In this experiment, driver's license candidates were randomly assigned to the control group, given free driving lessons, or given a large financial reward (i.e., the bonus group) if they obtained the driver's license in 32 days, two days longer than the minimum legal time of 30 days. Furthermore, upon obtaining the driver's license, participants were invited to a final session and enrolled in a surprise practical driving test in order to assess their driving skills. Results showed that participants in the bonus group were more likely to make extralegal payments and to obtain licenses without really knowing how to drive.

Green (1985) studied fraudulent behavior by auditing homes which had a "basic" cable service but which stole premium cable television signals with an unauthorized descrambler. This field experiment aimed to study general deterrence hypotheses by sending people known to be stealing signals a written legal threat, providing an amnesty period to rectify the situation without being prosecuted. The cable terminals were re-audited immediately after and 6 months after the intervention. Results of this experiment showed that about two-thirds of the original violators stopped stealing cable signals and this effect was maintained during the follow-up period.

The remaining five field experiments included in this category resorted to techniques frequently used in laboratory settings to study dishonest behavior, namely the dice roll task (Chytilová & Korbil, 2014; Okeke & Godlonton, 2014; Siniver & Yaniv, 2018) and the coin toss task (Buccioli & Piovesan,

2011; Houser et al., 2016). Experiments using these methodologies ask participants to roll a dice or toss a coin and report the outcome, knowing that different outcomes result in different rewards. These tasks are usually performed in unmonitored conditions, in order to assure participants that only they know the true outcome, creating an opportunity for them to act dishonestly for financial gain. These methods are unable to assess individual dishonest behavior, but the comparison of reported outcomes and the baseline distribution makes it possible to measure cheating at the aggregate level (see Rosenbaum et al., 2014).

Chytilová and Korbelt (2014) used the dice roll task with school students in order to study whether group settings influence dishonest behavior. The reward for completing a questionnaire was equal to the dice outcome, with the exception of the number “6” which would result in no payoff. Students rolled the dice either individually or in groups of three. Groups could also be determined randomly (exogenous groups) or students formed the groups themselves (endogenous groups). The main findings of this study showed that students in group settings (independently of the exogenous or endogenous formation) were more likely to act dishonestly.

Okeke and Godlonton (2014) applied the dice roll technique to study whether pro-social preferences lead to dishonest behavior. These authors recruited female interviewers to carry out interviews in the community. Interviewers visited households and distributed discounted price vouchers. Interviewers were supposed to ask the interviewees to roll the dice, and the amount of the voucher depended on the score they rolled. The misallocation of price vouchers was the measure of interviewer dishonesty. Results of this experiment showed that interviewers were more likely to allocate higher value vouchers to the poorest interviewees.

In the Siniver and Yaniv (2018) field experiment, participants were recruited after purchasing and scratching scratch cards at selling kiosks in order to study the effect of winning and losing in the lottery on dishonest behavior. Participants were asked to carry out the dice roll task under a cup and the monetary reward was determined by the participants’ report of the outcome. Results showed that lottery losers acted more dishonestly than lottery winners. Furthermore, the higher the lottery losses, the higher the dishonest behavior.

Regarding experiments using the coin toss task, Bucciol and Piovesan (2011) studied children’s dishonest behavior by asking summer campers to toss a fair coin in private. Depending on the reported outcome, they earned a prize, thus providing an incentive for them to act dishonestly. In this experiment, researchers manipulated whether or not they mentioned the possibility of cheating to the participants, requesting the experimental group not to cheat. Results showed that participants cheated somewhat in both groups and throughout the different ages (from 5 to 15 years), although boys cheated more than

girls. Nevertheless, the honesty request made to the experimental group reduced dishonest behavior by 16%.

Houser et al. (2016) used the coin toss technique to study the dishonest behavior of parents when the payoff was a toy for the child or cash for the parent. The presence of the parent's child in the room during the coin toss was also manipulated in order to study whether the presence of the child would increase scrutiny and thus lessen dishonest behavior. Accordingly to the authors' predictions, parents were more likely to act dishonestly to benefit their child than to benefit themselves. Also, dishonest behavior was expected to be higher when the child was not present. However, this effect was only found when the daughter was present, and parents' dishonest behavior did not change in the presence of their sons.

Stealing

A total of 16 field experiments were included in the stealing category (Table 3). In this category, field studies used two main methodologies. The first was the "lost" letter technique (and some adapted versions of this technique), which consists of leaving stamped, addressed, and apparently lost letters in determined places, typically containing a sum of money. The failure to return a "lost" letter containing money was defined as stealing. The second group of methods in this category used multiple techniques that provided the opportunity for participants to steal things such as pens, newspapers, and money.

Within the studies using the "lost" letter technique, the research conducted by Gabor and Barker (1989) used "lost" letters in order to study the prevalence of dishonesty in Canada. In their field experiment, researchers planted letters under the windshield wipers of cars of selected participants, with a note stating "found near your car." These envelopes contained a coin and a letter either appearing to be a personal and trivial letter or an official letter stating that the value of the coin was \$150. Overall, about one-quarter of sample failed to return the "lost" letter. However, the stated value of the coin failed to significantly impact the stealing of the letter. In agreement with previous experiments, participants' sex had little effect on stealing, contrary to their age, where younger participants were less likely to return the apparently lost letter.

The study conducted by Cohn et al. (2019) reported three large-scale field experiments conducted in 40 countries, using an adaptation of the "lost" letter technique to study civic honesty, by providing participants with the opportunity to return or steal a "lost" wallet. In these field experiments, confederates approached an employee at the counter (e.g., banks, hotels) and said that he/she found a wallet on the street and asked the employee to take care of it. Wallets included the "owners' personal information"

which allowed the employee to voluntarily return the “lost” wallet. In the first field experiment, the authors manipulated the money in the wallet, either no money or \$13.45 USD. Overall, results showed that citizens were much more likely to return the “lost” wallets with money than without. Moreover, despite dishonesty rates varying from 86% to 24% of cases, analyses showed that in none of the 40 countries the money condition increased significantly the likelihood to steal the “lost” wallet. In their second field experiment, Cohn et al. (2019) tested the same hypothesis with a larger amount of money contained in the “lost” wallet (i.e., \$94.15 USD). Dishonesty rates decreased even further with the big money condition, showing that the honest return rates for the “lost” wallet were higher when the larger amount of money was added. In the third field experiment, the authors manipulated whether the wallets with money included or did not include a key, in order to study the effect of an item valuable to the owner. Results of this last study showed that adding the key increased the return rates of the “lost” wallet, suggesting people’s concern for harm to the owner.

A further adaptation of the “lost” letter technique to study stealing was used in three studies (Keizer et al., 2008; Keuschnigg & Wolbring, 2015; Lanfear, 2018). This methodology consisted of leaving an envelope, visibly containing money, either hanging out of or nearby to a mailbox, and observing passerby behavior. Keizer et al. (2008) carried out six field experiments in order to study whether setting cues of violation of a contextual norm (e.g., graffiti in an anti-graffiti area) impacted deviant behavior. For the purposes of the present review, we are only going to focus on the last two field experiments referring to stealing, since the previous field experiments focused on littering and trespassing. In the fifth and sixth field experiments, “lost” letters visibly containing cash were left hanging out of a mailbox. The authors manipulated whether the setting was or not covered with graffiti (i.e., experiment 5) and whether or not there was litter on the floor around the mailbox (i.e., experiment 6). Results of both field experiments showed an increased odds of stealing the “lost” letter in the disorder conditions.

Keuschnigg and Wolbring (2015) carried out three field experiments that sought to replicate Keizer et al.’s (2008) field experiments on littering and stealing, and added an adaptation of these experiments to jaywalking. Similar to the previous study, only the field experiment on stealing falls within the scope of the present review. In the stealing experiment, apparently lost letters were left in front of a mailbox with visible cash in them. The authors manipulated the amount of money in the envelope (€5, €10, or €100). Also, the area surrounding the mailbox was either kept clean or there were two heavily wrecked bicycles next to the mailbox. Results clearly replicated the previous experiment, showing an increased odds of stealing the “lost” letter in the physical disorder condition. Furthermore, this spillover effect of the norm violation on stealing behavior was influenced by the amount of cash contained in the

envelopes, where the effect was the strongest when the envelopes contained the €5 note (i.e., people steal more in the disorder condition), weaker for the €10 note, and disappeared completely with the €100 note, showing that “once stakes are high, the relevance of environmental cues diminishes” (Keuschnigg & Wolbring, 2015, p. 120).

Lanfear (2018) carried a similar field experiment to study some key features of the broken windows theory. In this experiment, local physical disorder was manipulated by the addition or not of both litter and graffiti, and using the adapted “lost” letter technique with envelopes containing a \$5 bill left near the mailbox. This experiment failed to replicate the results of Keizer et al. (2008) as well as Keuschnigg and Wolbring (2015). Local disorder failed to impact passerby behavior on stealing the “lost” letter. Nevertheless, evidence indicated that in the disorder condition, participants were less likely to act pro-socially by mailing the “lost” letter.

The remaining 11 field experiments included in the stealing category used multiple methodologies that created an opportunity for participants to steal. Castillo et al. (2014), for example, sent out envelopes to Lima, Peru, from two cities in the USA via normal mail services. Researchers manipulated whether or not the envelopes contained cash, as well as the sender’s name, i.e., a foreign name (i.e., J. Tucker, M. Scott) or a local name (i.e., M. Sosa, L. Cordova). This methodology was developed to study whether the very nature of the mail influences stealing behavior of the people who handle the mail. Results showed that the envelopes containing money were much more likely to be lost. Furthermore, the mail was much more likely to be lost if the sender’s last name matched the recipient’s last name (i.e., a local name).

Belot and Schröder (2015), as well as Greenberg (2002), created the opportunity for participants to steal cash. In Belot and Schröder’s (2015) field experiment, the authors recruited students for a paid job of identifying the provenance of euro coins collected in different countries. Contrary to what participants were led to believe, a fixed number of coins was given to each participant, allowing the researchers to count the cash and assess the number of stolen coins. The authors manipulated whether or not participants were monitored, as well as incentives associated with monitoring, where participants’ mistakes in the coin sorting task were penalized either mildly or harshly. Results showed that about 10% of participants stole coins, though monitoring or incentives had no impact on participants’ stealing.

Greenberg (2002) used a sample of employees of a financial services company and asked them to complete a survey regarding working conditions in exchange for a payment. After completing the task, participants walked into an unsupervised room where they found a bowl of pennies, from which they should count the \$2 USD that was due to them. The researchers knew the total number of pennies that were in the bowl, allowing them to figure out whether or not the participant stole coins. These employees

belonged to two different locations, in one of which an ethics program was in place. Furthermore, the authors also manipulated whether the payment was coming from either personal funds or the company. The results showed that participants attending the corporate ethics program had a lower likelihood of stealing coins, and that participants stole more often when the money was said to come from a company.

Greenberg (1990) carried a field experiment that also focused on employee theft, in this particular case, concerning the inventory of a firm. Employees of several manufacturing plants were either or not subjected to a 15% pay reduction during a period of time. The groups receiving the wage cuts were divided into two groups. One group received an adequate explanation for the wage cut by the company president, while the other group was in the inadequate-explanation condition. Employee theft was assessed by the percentage of unaccounted inventory lost. Results revealed that employee theft increased during the pay reduction period. Furthermore, the theft rate in the inadequate-explanation condition was much higher than in the adequate-explanation condition.

Cohn et al. (2014) also studied the impact of wage cuts on employee theft. In this specific case, hired workers were asked to sell promotional cards. While selling these promotional cards, workers were supposed to collect information from the customer. This created the opportunity for willing workers to steal the cash sales and fake customer information. Incorrect customer information could be checked by the research team. Sales were carried out in groups of two, and, in the first phase, all workers were given the same hourly wage. In the second phase, the wage either stayed the same, both group members received a 25% wage cut, or only one group member suffered the 25% cut. Results showed that the wage cut created an increased likelihood of employee theft, but this only happened for the employees who were directly affected by the wage cuts in the unilateral condition.

Widner (1998) developed a field experiment in order to assess the effectiveness of a series of intervention techniques aimed to reduce the theft of petrified wood in a national park. The three interventions tested in this study included a uniformed volunteer, deterrent signs, and a signed pledge, and each was randomly in place for 10 days. The theft of petrified wood was assessed by direct field observation carried out by the research team. Using this methodology, researchers were able to observe a theft rate of 2.1% in the control condition, and this reduced to about 1.4% in the intervention conditions. These results revealed that the three interventions were effective in the reduction of theft, when compared to the control condition. Furthermore, these interventions showed no differential effectiveness.

Schlüter and Vollan (2015) developed a field experiment where they studied the theft of flowers in a farmer's field using an honor system. This was an unattended system that allowed the customer to enter the flower field, cut the intended flowers themselves, and pay the respective sum in a cashbox,

relying entirely on the honesty of customers. Researchers left a message near the cashbox which varied between legal threats, moral persuasion, and referencing a family business or a consulting firm. Theft of flowers was assessed through direct observation carried out by the researchers through a semi-transparent window by counting the flowers and the respective payment into the cashbox. Findings suggested no main effect of the legal or moral messages. However, flower theft increased when the flower field was framed as a company business, compared to the family business condition.

Two other field experiments used the honor payment system in order to study stealing, in their case of newspapers (Geller et al., 1983; Pruckner & Sausgruber, 2013). In order to ensure unmonitored transactions, experimenters placed just one paper in the sales booth and checked for payments at specific intervals. If the newspaper had been taken, the cashbox would be emptied recording the amount paid (Pruckner & Sausgruber, 2013). In the field experiment conducted by Geller et al. (1983), two anti-theft sign messages were implemented; one appealed to moral, internal control and the second showed a legal threat. Results supported the effectiveness of both messages in reducing newspaper theft. Similarly, in the second field experiment on theft of newspapers using the honor system (Pruckner & Sausgruber, 2013), the authors also tested the impacts of a moral, a legal, and a neutral control message. Findings revealed that about two-thirds of customers stole the newspaper, and those who paid did so by depositing much less than the indicated price (i.e., €0.60). The treatments showed no effects on newspaper theft. However, the appeal for honesty in the moral condition caused an average increase on the amount paid, compared to both control and legal treatments.

The final two field experiments included in the stealing category focused on university students. In the experiment conducted by Cagala et al. (2014), students were randomly allocated to two groups with different levels of monitoring during a university exam. Students in both groups were provided with a high-quality pen that they were supposed to deliver in the post-exam phase, where the levels of monitoring were the same throughout the experimental groups. Results showed that the monitoring in the exam phase caused an intertemporal spillover effect, where participants in the low monitoring group were much more likely to steal the pen. Finally, Wortley and McFarlane (2011) carried out a field experiment in a university library and created the opportunity for students to steal a photocopying card. Researchers left a photocopying card unattended on a library table and observed passerby behavior from a distance. Researchers manipulated ownership of the card by using either a signed or an unsigned card, and manipulated guardianship, by placing the card either next to library books (giving the impression that the owner was nearby) or on its own. Both variables of symbolic territoriality (i.e., signed cards/next to

books) decreased the likelihood of photocopying card theft, providing evidence of effective crime prevention.

Keeping money

The keeping money category was composed of a total of nine studies that used field experimental designs where the researchers created the opportunity for participants to dishonestly keep money that did not belong to them. These methods included situations where the participants could keep wrongly received money (e.g., extra change or money received on their phones), keep lost money, or even accept a bribe (Table 4).

Four experiments created the opportunity for dishonest behavior by handing people extra change. In these experiments, behavior was considered dishonest whenever participants noticed the extra money and still kept it to themselves. In the study of Azar et al. (2013), restaurant customers who paid with cash received excessive change. The amount of the extra change was manipulated, either a smaller amount (the equivalent to about \$3 USD) or a larger one (about \$12 USD). Analyses showed that participants in the condition where they received less extra money were more likely to keep it. Further results showed that only about a third of customers returned the excessive change, though repeated customers as well as female customers returned the excessive change more often.

Yuchtman-Yaar and Rahav (1986) developed a field experiment with bus passengers, where bus drivers gave passengers extra change. The temptation to keep the extra money was manipulated by giving an extra 7% or 25% of the total ticket cost. Out of the total passengers noticing the extra change, the level of temptation showed no main effect on the levels of dishonesty. However, these authors found an interaction between level of temptation and passengers' sex, where male passengers were more likely to keep the extra change in the low temptation condition, while female passengers were more likely to keep the extra change when the monetary temptation was higher.

Gabor et al. (1986), as well as Rabinowitz et al. (1993), carried out field experiments where confederates gave extra money to store cashiers. In the Gabor et al. (1986) experiment, a confederate walked into a store, picked up a local newspaper costing 30 cents, paid for it with a single Canadian dollar bill, and proceeded towards the door without awaiting the change. One of three confederates (i.e., a Caucasian male, a Caucasian female, and a male of East Indian descent) visited either chain-type or family-type stores. The type of store, along with confederates' ethnic origin, did not affect cashiers' dishonest behavior. On the other hand, cashiers were significantly more likely to act dishonestly to male customers than towards the female confederate. Regarding the Rabinowitz et al. (1993) field experiment,

American confederates visited Austrian shops with female employed cashiers and purchased two postcards costing 4 shillings each (equivalent to \$9 USD). While making the payment, confederates either overpaid or underpaid the cashier by 1 shilling and walked away slowly. Overall, cashiers dishonestly kept the overpaid money in 26% of cases (after taking account of carelessness). Furthermore, results showed that cashiers were more likely to act dishonestly to female confederates than to males.

Similar to the extra change paradigm, Yap et al. (2013) gave extra money to participants and watched their behavior in order to test whether expansive postures lead to dishonest behavior. In this field experiment, community members were invited to participate in a study about the relationship between stretching and impression formation, in exchange for a \$4 payment. Participants were randomly assigned to hold either an expansive or a contractive pose for 1 minute. After completing the study, participants were handed \$8, comprising three \$1 bills and one \$5 bill, giving the impression to the participants that this was an accidental overpayment. Participants who checked the money and kept the extra payment were considered to be acting dishonestly (i.e., “stealing by omission”). According to the authors’ predictions, the one-minute pose had a significant effect on participants’ dishonest behavior, where participants in the expansive pose were much more likely to keep the extra payment.

Alem et al. (2018) transferred money (the equivalent of \$12 USD) to mobile phones and immediately afterwards sent a text message asking participants to return the supposedly misdirected payments. The authors manipulated these text messages in three experimental conditions, either framed neutrally, offering part of the money as a gift, or trying to induce a feeling of guilt in recipients. Results of this experiment showed that both kindness and guilt messages resulted in higher return rates (i.e., reduced dishonest behavior) compared with the neutral message.

Using a different paradigm, Newman (1979) carried out a field experiment in a social situation, in which a confederate dropped a coin while approaching an unsuspecting participant, and an observer recorded the participant’s behavior. The experimental situation occurred either in a university campus or in a central shopping area. Furthermore, the value of the dropped coin was also manipulated (2p or 10p GBP). Dishonest behavior was much more common in the city than on the campus setting, where dishonest behavior was virtually nonexistent. Considering the coin value, within the city condition, dishonesty increased with the higher value coin.

Gire and Williams (2007) left a “lost” dollar bill in specific places and noted whether passerby college members collected the money. In this field experiment, researchers manipulated whether the note was “lost” either in public (sidewalk) or in private (bathroom) settings at campuses of two colleges, a military and a nonmilitary college. Contrary to the nonmilitary college, the military college had a very

stringent honor code that was rigorously enforced, which made picking the “lost” note an honor violation that could lead to dismissal from the college. This allowed for a real-world test of the effect of severe threats on people’s behavior. In agreement with the authors’ predictions, military college members were less likely to take the dropped money than nonmilitary college members. Moreover, while nonmilitary college members took the note at the same rate regardless of setting, military college members were much more likely to take the money in the private setting.

The final field experiment in the keeping money category was conducted by Armantier and Boly (2011), who recruited participants for a paid part-time job to spell-check a set of 20 exam papers. The 11th paper came with a bribe and a message asking the person to find few mistakes in the exam. The amount of the bribe, the wage paid to graders, and the level of monitoring (with punishment) were manipulated. In the control condition, nearly half (49%) of participants accepted the bribe. Regarding the experimental manipulations, doubling the bribe, a lower wage, and when the job was not monitored and punished increased the likelihood of bribe acceptance.

Shoplifting

In our final category, we have included 14 field experiments that explored the effectiveness of multiple anti-theft programs in market stores, either by customers or by store employees (Table 5). Studies included in this category resorted to methods such as the randomized controlled trial, pre-posttest design, and multiple baseline design. Regarding the measurement methods, these field experiments used very similar methodologies, where the inventory of stocks was compared to the number of sold items, and the difference between the two figures corresponded to shoplifted products.

The two field experiments carried out by McNees et al. (1980a) and Carter et al. (1988) used the multiple baseline design to study employee theft. Through this methodology, the authors are able to provide feedback to the participants regarding a specific type of product and see its effect on stealing. Then, after a period of time, researchers changed the type of product address in the feedback information and observed its effect on the theft of that specific product category. In the field experiment of McNees et al. (1980a), signs were posted clearly stating the number of items stolen by employees, and requesting them not to steal. After a period of time, the sign switched to a different item category (i.e., either potato chips, milk, ice cream bars, or cold sandwiches). Findings showed that providing product-specific information about employee theft effectively reduced theft rates of that type of items. Regarding Carter et al. (1988), the study was conducted in a grocery store in order to estimate employee theft. Store employees were given information about the inventory loss of specific items through graphs posted in the

lunchroom, while customers were not made aware of this information. The specific items changed on a biweekly basis. Results showed that the amount of stock lost from the specific items' categories was reduced immediately following the introduction of the intervention.

Carter et al. (1980) also carried out a field experiment using the multiple baseline design. In this case, researchers varied the information put up in three signs, informing customers that items marked with red dots were frequently stolen. The items marked with the red dots changed on a weekly basis (i.e., lip gloss, Elvis Presley records, leather coats, and small wrenches). Results showed a significant decrease in item losses following their public identification as frequently taken.

Five other shoplifting studies used randomized controlled trial designs in order to explore the efficacy and cost-benefit of preventive mechanisms of theft (Hayes & Downs 2011; Hayes et al., 2011, 2012, 2019; Johns et al., 2017). In the field experiment developed by Hayes and Downs (2011), researchers aimed at testing the efficacy of three situational crime prevention treatments, namely, in-aisle closed-circuit television (CCTV) with public view monitors (PVMs), in-aisle CCTV domes, and protective keeper or safer boxes. The presence of PVMs and domes are expected to increase the concern for detection and deter theft. Similarly, the keeper or safer boxes, because they are difficult to open, increase both the offender's theft effort as well as the perceived risk of detection. These three interventions, along with a no-intervention control, were randomly assigned to the stores entering this study. Findings showed all three treatments to be effective anti-shoplifting interventions, causing significant reductions in the stock losses of 57% in in-aisle CCTV with PVMs, of 27% in in-aisle CCTV domes, and of 61% in protective keeper or safer boxes, compared to the control stores.

Regarding the trial carried out by Hayes et al. (2011), researchers tested the efficacy of keeper or safer boxes and found this intervention to significantly reduce theft by 52% compared to control conditions. Hayes et al. (2012) aimed at testing the efficacy of two situational crime prevention treatments, namely a protective product handling, which involved increased attention paid to the high-loss test product and a reduced general access to the product, and a protective product display, consisting of an audio alert tone. Findings showed both treatments to be effective, causing a significant reduction in the stock losses of over 50% compared to the control condition.

Johns et al. (2017) tested the efficacy of protective display fixtures, a mechanism that forces customers to press a button to obtain the product, and enhanced PVMs (ePVM), an intervention in which the attention to the presence of the in-aisle monitors is highlighted by flashing lights. Findings showed that only the protective display fixture caused a statistically significant reduction of stock loss, resulting in a reduction of theft by 41% in these stores. Finally, Hayes et al. (2019) randomly assigned anti-theft wire

wraps to highly shoplifted categories of products (i.e., cordless electric drills, weight loss supplements, and skincare products). Results of this field experiment showed that the effectiveness of wire wraps varied by product category. This anti-theft intervention reduced theft rates of cordless electric drills compared to the control stores, but failed to show the same result for the other product categories.

The remaining six field experiments included in the shoplifting chapter used pretest/posttest designs. McNees et al. (1980b) tested the efficacy of an anti-shoplifting program directed at young students, where young customers gained a token in every purchase which they could trade for prizes when holding five tokens. This field experiment was conducted in a convenience food market located near an elementary school. Researchers estimated the baseline regular amount of merchandise stolen during the pretest, posttest, and during a follow-up period. Findings showed that, during the implementation of this program, the rate of stolen items decreased by over 50%, compared to the baseline. However, the rate of shoplifted items increased after the program was terminated.

Thurber and Snow (1980) tested the effectiveness of anti-theft posted signs on the shoplifting of cigarettes. Contrary to the authors' expectations, the posting of anti-shoplifting signs was associated with increases in cigarette theft rates when compared to the pretest period. Carter and Holmberg (1993) evaluated the anti-theft effectiveness of a public identification intervention. In a grocery store, researchers publicly identified high-risk items using red dots. The implementation of this intervention was associated with reductions in the rate of theft, an effect that lasted up to a 15-week period. Farrington et al. (1993) carried out a field experiment in order to test the anti-shoplifting effectiveness of three methodologies. These methods included electronic tagging (where an alarm would sound if a tagged item was taken through the door), store redesign (which lessened the opportunities for shoplifting), and a uniformed guard. Regarding the results of this field experiment, contrary to the uniformed guard that failed to show any effect on shoplifting, the two interventions of electronic tagging and store redesign showed a significant reduction in the shoplifted items. However, from these two effective methodologies, only the electronic tagging caused a decrease in the shoplifting rate that was maintained over time.

DiLonardo and Clarke (1996) carried out a field experiment in order to compare the effectiveness of two anti-shoplifting techniques. The authors resorted to four stores that, despite the use of electronic article surveillance (EAS), presented high shoplifting rates and replaced the anti-theft method by ink tags (i.e., tags attached to the products that would break and stain the garment if tampered; a warning to this effect was printed on the tag). Results showed that ink tags were associated with a reduction of 42% in the amount of stock loss compared to EAS. Finally, Hayes and Blackwood 2006 carried out a field experiment to evaluate the anti-shoplifting effects of EAS in retail stores. In this study, the hidden EAS

was implemented either at the 50% level (every other item was tagged) or the 100% level. Contrary to the authors' predictions, the implementation of EAS had no effect on the stock loss.

Discussion

The present study had two aims. First, we aimed to systematically review the field experiments on stealing and dishonesty that had been published after the review of Farrington (1979). In doing so, we especially coded detailed information about the experimental design, in order to provide relevant information to researchers who are interested in designing and conducting field experiments. Secondly, inspired by SEU, we additionally coded the studies in whether the manipulated variables were related to benefits and costs (i.e., *costs for the self* versus *costs for the other*), in order to establish whether variations in costs and benefits predict levels of dishonesty (cf., Farrington 1979). In line with Farrington (1979), *benefits for the perpetrator* included financial gains. "Costs for the other" included factors such as the suffering of the victim because of the actions of the perpetrator. Finally, "costs for the self" included factors related to the likelihood of apprehension. In the current review, we used these definitions as guidelines for coding studies on *costs for the other*, *costs for the self*, and benefits.

In our literature search, we identified four categories of field experiments on deviance: fraudulent/dishonest behavior, stealing, keeping money, and shoplifting. Below, we summarize and describe the implications of the findings for each category. This is followed by a critical overview of how far the study of field experiments has come, and what is still needed to make greater advances in the future. Finally, we end with a conclusion.

Fraudulent/dishonest behavior

Of the studies that manipulated *costs for the other*, two out of four found significant results. These two studies (Kerschbamer et al., 2016; Tracy & Fox, 1989) manipulated costs for insurance companies (low costs for the victim) versus costs for the clients (high costs for the victim), and showed that, especially when the costs were low, the perpetrators overcharged their clients more. However, the two remaining studies (Balafoutas et al., 2013; Conrads et al., 2015) that found non-significant results manipulated the perceived income of the victim, and showed that, when the victim was perceived to have high SES (i.e., low costs for the victim), this did not predict higher overcharging by the perpetrators. Only two studies manipulated benefits and showed that higher financial benefits predicted more fraudulent behavior (List

& Momeni, 2020; Tremblay et al., 2000). As for *costs for the self*, all of the seven studies that manipulated this factor showed that a lower likelihood of apprehension predicted higher levels of fraud.

Stealing

The three studies in the stealing category that manipulated “costs for the other” found significant effects. One of these studies showed that theft increased when payment came from a company (lower costs) compared with personal funds (higher costs) (Greenberg, 2002). A second study showed that, when the owner’s name was not signed on a photocopying card in a library (i.e., low costs), it was stolen more often (Wortley & McFarlane, 2011). The other study showed that, when “lost” wallets contained a personal item valuable to the owner (i.e., high cost for the other), the return rates of the “lost” wallet increased (Cohn et al., 2019). Next, “costs for the self” was investigated in three studies. Two of these studies found significant deterrent effects of monitoring; namely, Cagala et al. (2014) found that high monitoring during the exam phase decreased pen theft in the post-exam phase, whereas Widner (1998) found that having anti-theft interventions decreased petrified wood theft. The other study did not find that monitoring decreased the theft of coins (Belot & Schröder, 2015).

Finally, four field experiments manipulated the amount of benefits to the self. Castillo et al. (2014) found that letters containing money increased mail theft. Keuschnigg and Wolbring (2015) found a significant effect in interaction with another variable (i.e., disorder environmental cues). One other study found that “lost” letter theft was not affected by the apparent value of the contained coins (Gabor & Barker, 1989). Moreover, two field experiments carried out by Cohn et al. (2019) found an opposite effect compared to our hypothesis, where the higher the amount of money in a “lost” wallet, the less stealing was committed by employees (i.e., in such cases, the employees at the counters more often mailed the wallets back to the hotel guests).

Keeping money

Concerning the *costs for the other* manipulation, we only located one study (Gabor et al., 1986) that fitted this description. Gabor et al. (1986) investigated cashiers’ dishonesty in keeping the change of customers in chain stores (low costs for the other) versus family stores (high costs for the other). However, unexpectedly, the chain stores condition did not lead to more cashiers’ dishonesty regarding keeping the change of customers. As for *costs for the self*, the only such study (Armantier & Boly, 2011) in the keeping

money category showed that low (versus high) monitoring, coupled with punishment if caught, led to increases in accepting a bribe.

Finally, we located five studies in the keeping money category that manipulated benefits. Three of those studies (Armantier & Boly, 2011; Newman, 1979; Rabinowitz et al., 1993) consistently showed that higher benefits predicted more instances of participants keeping or accepting money that was not theirs (i.e., picking up a dropped coin; acceptance of a bribe; keeping due change). However, although the remaining two studies also found significant effects, the effects were in the opposite direction compared to our hypothesis. In Azar et al. (2013), customers of a restaurant received extra change after paying, and the amount of extra change was manipulated. Higher amounts of extra change actually decreased the instances in which customers kept the “extra” change. Similarly, in Yuchtman-Yaar and Rahav (1986), bus drivers gave passengers extra change and the amount of extra change was manipulated. For females, higher amounts of extra change increased dishonesty, but for males, higher amounts of extra change actually decreased keeping the extra change. It is of note is that both studies that found the opposite effect for benefits originated from Israel.

Shoplifting

For the shoplifting category, regarding components of the SEU theory, we only found studies that manipulated *costs for the self*. The intervention study of DiLonardo and Clarke (1996) investigated security measures to prevent shoplifting and showed that ink tags (versus electronic article surveillance; EAS) reduced shoplifting. Of note is that both ink tags and EAS increased the chances of apprehension (i.e., *costs for the self*) compared with a condition without security measures. Thus, the intervention in DiLonardo and Clarke (1996) would have been a more stringent test of the *costs for the self* hypothesis, if its security conditions had been compared to a condition in which no security measures were used.

On the other hand, Hayes and Downs (2011), Hayes et al. (2011, 2012, 2019), and Johns et al. (2017) compared control conditions to anti-shoplifting interventions (i.e., CCTV, keeper or safer boxes, protective product display, or anti-theft wire wraps) and showed that these interventions reduced the stores' theft rates. McNees et al. (1980b) showed that an anti-shoplifting intervention directed to elementary school students reduced the rates of theft, though these findings were not maintained over time. Finally, Farrington et al. (1993) conducted a series of experiments and showed that electronic tagging reduced shoplifting, and this effect was maintained over time; however, a uniformed guard did not affect shoplifting.

Benefits versus costs for the other versus costs for the self

In sum, the above-described results show that when the chance of apprehension (*costs for the self*) is low, more dishonest behavior takes place. This pattern of findings was found in all the seven studies on fraud, in two out of the three studies on the stealing, in the study on the keeping money, and in all eight studies on the shoplifting category. Concerning *costs for the other*, there were too few studies that manipulated this factor in order to draw strong conclusions for each category. In the shoplifting category, there were no studies that manipulated *costs for the other*. Across the categories, four out of eight studies that manipulated *costs for the other* found that when costs are low for the victims (e.g., an insurance company versus an individual), then dishonest behavior increases.

Finally, when it comes to benefits, the studies across the different categories consistently showed that high benefits predicted dishonest behavior, as seven of the 11 studies found such significant effects. However, of note is that, in the fraudulent behavior category, only two studies manipulated benefits (and both studies found significant effects), and in the shoplifting category, no study manipulated benefits. Seven of the 10 studies that found significant results (i.e., three studies for the stealing category, five studies for the keeping money category, and two studies for the fraudulent category) found an effect in the hypothesized direction showing that more benefits led to more stealing and keeping money. However, in the remaining three cases, the opposite pattern of effect was reported: fewer benefits predicted more dishonest behaviors of perpetrators when the studies manipulated the amount of extra change given to customers or when the experiment manipulated how much money was in a lost wallet. Perhaps the relation between the benefits and the probability of dishonest behavior follows an inverted-U shape.

Past, present, and future

The current review shows that researchers in many different parts of the world have carried out field experiments to study financial dishonesty. Such cultural diversity is very welcome, in order to determine to what extent theories are generalizable. Of course, legal definitions of deviance (e.g., theft) might vary substantially across cultures. Such discrepancies should be kept in mind when interpreting the findings of the studies highlighted in this review. However, a further advantage of the field experimental methodology to study offending and dishonest behavior is the fact that the reviewed experiments focused on naturally recurring behaviors, and are generally independent of the legalistic definitions of offending.

Within the present study, in order to review the field experimental evidence relevant to the study of deviance, including the field experiments on stealing and dishonesty developed by behavioral

economists, we have systematically reviewed the field experimental studies on stealing and monetary dishonesty. However, in doing so, we have not included the field experiments on other types of deviant behavior that might be of interest to the study of criminal behavior, such as littering, jaywalking, or vandalism (e.g., property damage). Hence, readers should bear in mind that the findings in the present review might not be generalizable to other types of deviant behavior, and we encourage researchers to investigate these topics in the future. Especially in the study of vandalism, this type of deviance should be relatively easy to investigate in field experiments, considering that (1) it often happens in public view and (2) it is less ethically sensitive compared to other types of deviancy (e.g., theft, sexual assault, physical assault) (Farrington, 1979). For example, vandalism experiments could be conducted in areas where vandalism already takes place in public view (Zimbardo, 1969). Therefore, researchers would need to worry less about ethical considerations associated with providing individuals with the opportunity to act in a deviant manner, which is typically the case in field experiments on deviance.

Costs and benefits were the focus of this review, in part because these are immediate situational factors that are suitable for manipulation in short-term experiments (Farrington & Knight, 1980). However, it should be noted that dishonesty is a complex behavior, which cannot solely be explained by such immediate situational factors. Future studies should also attempt to vary other non-situational variables (e.g., impulsivity), as well as social environmental factors (e.g., the presence of peers) (Defoe et al., 2019; Defoe, in press). Studies that manipulate the social context remain rare in field experiments in the criminology literature. However, the few available studies suggest that the immediate social context also plays a role (Farrington, 1979).

Conclusion

Our review shows that it is worthwhile for criminologists to study influences on offending using field experiments within a SEU framework. This review clearly demonstrates that variations in the benefits and costs (particularly the likelihood of apprehension) associated with a dishonest act are important predictors of offending. Specifically, higher levels of financial benefits and lower probabilities of apprehension predict higher levels of dishonesty. Interestingly, some studies found that fewer benefits led to more stealing. More research is needed on why this effect is sometimes in the opposite direction, and why higher benefits sometimes lead to less dishonesty. Perhaps in such cases, there could be an interaction with costs and benefits that are driving the effects. Therefore, future studies are also encouraged to investigate potential interactions between costs and benefits.

The present review shows how immediate situational influences on dishonesty (e.g., costs and benefits) can be manipulated in field experiments to better understand the causes of stealing and dishonesty. Although many economists have undertaken this challenge, such experiments in criminology remain rare (for an overview see Clarke, 1995; Clarke & Cornish, 1985). However, field experiments on financial dishonesty overlap considerably with everyday delinquency, and hence, such experiments could be a powerful tool for criminologists (Farrington, 1979; Farrington et al., 2020). In fact, targeting immediate situational factors that predict crime could be just as successful as prevention programs that solely target individual characteristics (e.g., impulsivity). We conclude that criminologists should seek to carry out naturalistic field experiments on offending to investigate theories and explanations of offending.

Table 2*Summary of field experiments in the Fraudulent/ dishonest behavior category*

Study	Participants	Design	Measure	Main findings	SEU
Balafoutas et al. (2013)	Taxi drivers (348 taxi rides)	Task: taxi ride. Manipulation - taxi driver's perception of customers: <u>Information about the city</u> : Local vs. Non-local natives; <u>Information about the tariff system</u> : Native vs. Foreigner; <u>Income</u> : Low vs. High income.	Overcharging	Non-local natives increased overcharging. Foreigner customers increased overcharging. Customer's perceived income did not impact overcharging.	Costs for the other
Bertrand et al. (2007)	822 driver's license candidates	Task: Obtaining a driver's license. Manipulation: <u>Prize</u> : Bonus (large financial reward if obtained in 32 days) vs. Free driving lessons vs. Control.	Extra-legal payments	Bonus group members are more likely to make extra-legal payments to obtain licenses.	*Benefits
Blais and Bacher (2007)	Insurance customers (765 claims)	Task: Insurance companies randomly assigned claims of property theft to study groups. Manipulation: <u>Deterrence</u> : Conventional vs. Deterrent letter.	Insurance fraud (i.e. claim padding)	The deterrent letter decreased fraudulent behavior.	*Costs for the self

Buccioli and Piovesan (2011)	160 children attending a summer camp Italy	Task: Summer campers were asked to carry out the coin toss task as a typical camp activity. Manipulation: <u>Honesty request</u> : Control vs. Explicit request to refrain from cheating.	Coin toss task.	Honesty request reduced cheating.	NA
Chytilova and Korbela (2014)	226 school students Czech Republic	Task: Students were recruited for a task. Reward would be determined by the dice roll task. Manipulation: <u>Setting</u> : Individual vs. Groups of three; <u>Group formation</u> : Exogenous (randomly formed groups) vs. Endogenous (groups formed by themselves).	Dice roll task.	Group settings increased dishonesty. Group formation did not impact students' dishonesty.	NA
Conrads et al. (2015)	Candy sellers in 50 markets (200 observations) Germany	Task: Confederates entered the market and bought a bag of candy. Manipulation: <u>Status of the buyer</u> : Wealthy vs. Poor; <u>Quantity of candy bought</u> : High (150g) vs. Low (50g).	Overcharging	Overcharging in 38% of purchases.	Costs for the other
Dugar and Bhattacharya (2017)	Fish sellers in 10 markets (160 observations)	Task:	Overcharging.	Overcharging in 89% of purchases.	NA

	India		Confederates entered markets and purchase a pre-determined quantity of fish. Manipulation: <u>Size of fish</u> : Small (less expensive) vs. Large (more expensive); <u>Type of fish</u> : Rohu (less expensive) vs. Catla (more expensive).		Within less expensive type of fish, large fish increased overcharging. Within more expensive type of fish, small fish increased the probability of overcharging.
Green (1985)	67 subjects found to be stealing cable television signals	USA	Pretest / Posttest design. (1) Researchers identified houses that illegally tampered with terminals; (2) a deterrent letter threatening criminal prosecution was sent; (3) a re-audit was developed after the letter was sent; and (4) follow-up audit six months after.	Stealing cable television signal.	The deterrent letter decrease cable crime. Deterrent effect lasted at least six months. *Costs for the self
Houser et al. (2016)	249 parent-child pairs	USA	Task: Parents of 3-6 year-old children were recruited for a task. Reward would be determined by the coin toss task. Manipulation <u>Scrutiny</u> : Parent is alone vs. Child in the room during the coin toss; <u>Moral cost</u> : Low (prize pack for the child) vs. High (\$10 for the parent).	Coin toss task.	Low moral cost NA increased cheating. No scrutiny increased cheating.

Jesilow and O'Brien (1980)	145 auto shops	Pretest / Posttest design	Fraud in the vehicle repair price.	Deterrent intervention decreased fraudulent behavior.	*Costs for the self
USA		Task: Female confederates entered repair facilities and requested to test the car batteries because they their cars would not start.			
		Manipulation: <u>Intervention</u> : Control vs. Deterrent intervention (deterrent announcements and letters).			
Kerschbamer et al. (2016)	61 computer repair shops	Task: Confederate entered computer repair shops and asked for a repair. Computers were manipulated with a destroyed RAM module.	Fraud in the computer repair price.	Average repair price in Control and Insurance groups was 70.17€ and 128.68€, respectively.	*Costs for the other
Austria		Manipulation: <u>Insurance</u> : Control vs. Insurance		Insurance increased fraudulent behavior.	
List and Momeni (2017)	3,022 hired workers through MTurk	Task: Participants were contracted online to transcribe 10 images. Before starting, workers reported whether the image was readable. If not readable, the transcription was not necessary.	Dishonesty	Decrease in wage increased dishonest behavior.	NA
USA and India				Implementation of CSR, especially on behalf of the workers, increased dishonest behavior.	

Manipulation:

Wage: Low (\$0.90) vs. Medium (\$1.20) vs. High (\$1.26);

Corporate Social Responsibility (CSR): Charity donation on behalf of the firm vs. On behalf of the workers.

List and Momeni (2020)	2,000 hired workers through MTurk USA	Task: Participants were contracted online to transcribe 10 images. Before starting, workers reported whether the image was readable. If not readable, the transcription was not necessary. Manipulation: <u>Upfront payment</u> : 0% vs. 10% vs. 50% vs. 90% of the total pay.	Dishonesty	Upfront payment decreased dishonesty, when compared to 0% upfront. Considering upfront conditions, the higher the upfront payment the higher the probability to behave dishonestly.	*Benefits
Nagin et al. (2002)	Employees of a large call center company USA working	Task: Employees called potential donors and request contributions. Payment was a base salary and a bonus for the number of successful solicitations. Manipulation <u>Reported monitoring to employees</u> : audit rates varied.	Fraud (i.e. "Bad calls").	A reduction in monitoring increased fraud.	*Costs for the self

Okeke & Godlonton (2014)	10 women interviewers	Task: Interviewers were hired to conduct household visits and distribute discounted price vouchers. The price vouchers would be determined by the dice roll task.	Dice roll task.	Interviewers were more likely to allocate the higher value vouchers to the poorest beneficiaries.	NA
	Nigeria				
Olken (2007)	608 villages	Task: Funding was awarded to villages to build a road. Public meetings were implemented to encourage public participation in the monitoring process.	Fraud in the road cost (i.e. road samples dug and analyzed).	Information about audits decreased fraudulent behavior.	*Costs for the self
	Indonesia	Manipulation: <u>Monitoring</u> : No audit vs. Audit; <u>Direct participation</u> : No invitation vs. Invitation (villagers received invitations to attend meetings) vs. Invitation and comment form (villagers received invitations and an anonymous form).		Villagers' participation did not affect fraudulent behavior.	
Schneider (2012)	40 auto repair garages	Task: Confederates submitted a test vehicle with a prearranged set of defects to garages for repairs. The mechanic was asked to inspect the vehicle and provide a price estimate.	Overcharging.	Low-reputation increased overcharging.	NA
	Canada	Manipulation:		Average overcharge for low-reputation and high reputation was \$59.75	

			<p><u>Reputation:</u> Low-reputation (Confederate said to be moving out the city) vs. High-reputation (Confederate said to be moving into the city)</p>	and \$37.70, respectively.	
Shu et al. (2012)	Insurance customers (13,488 policy forms) USA	<p>Task: Customers of an insurance company were requested to report the current odometer mileage of their cars</p> <p>Manipulation <u>Signature:</u> At the beginning vs. At the end of the report.</p>	Insurance fraud	<p>Signing at the end of the report increased fraudulent behavior.</p> <p>Signing at beginning of the form led to a 10.25% increase in implied miles driven.</p>	*Costs for the self
Siniver and Yaniv (2018)	300 kiosks customers Israel	<p>Task: People were observed purchasing scratch cards. Upon completing scratching their cards, participants were invited to participate in a simple task with monetary payoffs.</p> <p>Manipulation: <u>Scratch outcome:</u> Winners (Profit > 0) vs. Break eveners (Profit = 0) vs. Losers (Profit < 0).</p>	Dice roll task.	<p>Losing in the lottery increased dishonesty.</p> <p>Winners in the lottery reported, on average, a lower outcome (7.75), followed by break eveners (8.20), and, finally, by losers (9.80).</p>	NA
Tracy and Fox (1989)	96 auto body repair shops USA	<p>Task: Confederates entered auto repair shops and obtained estimates of repair costs for their cars.</p>	Fraud in the vehicle repair price.	Insurance coverage increased fraudulent behavior.	*Costs for the other

		<p>Manipulation: <u>Insurance</u>: No insurance vs. Insurance; <u>Sex of drives</u>: Female vs. Male.</p>		<p>Within not covered by insurance, female drivers increased fraudulent behavior.</p>
Tremblay et al. (2000) Canada (French)	Insurance customers (321 claims)	<p>Task: Customers of an insurance company who reported theft and burglary for compensation for losses received letters from the insurance company.</p> <p>Manipulation <u>Letters</u>: Control (civil warning) vs. Deterrence (criminal warning) vs. Permissive (without any warning) <u>Claim regulation</u>: Internally (by telephone) vs. Externally (face-to-face).</p>	Insurance fraud	<p>Permissive letter increased fraudulent behavior, only when claim was settled by phone.</p> <p>Deterrent letter decreased insurance fraud, only when claim was settled face-to-face.</p>

Note. SEU = Availability of Costs or Benefits Manipulation; NA = not applicable; * = the manipulation of costs or benefits was significant.

Table 3*Summary of field experiments in the Stealing category*

Study	Participants	Design	Measure	Main findings	SEU	
Belot and Schröder (2015)	91 University students	Germany	Task: Participants were recruited to sort boxes of euro coins. Manipulation: <u>Monitoring</u> : No vs. Monitoring. <u>Incentives</u> : Mild vs. Harsh incentives.	Theft of coins.	10% of the participants stole coins. Monitoring did not affect theft. Incentives did not affect theft.	Costs for the self
Cagala et al. (2014)	766 University students	Germany	Task: Students taking an exam were provided with a high-quality pen. Manipulation: <u>Monitoring</u> : Low vs. High monitoring during the exam.	Theft of pen in the post-exam phase.	High monitoring during exam phase decreased pen theft in the post-exam phase.	*Costs for the self
Castillo et al. (2014)	Postal workers (541 observations)	Peru	Task: Envelopes were sent from the USA to Lima, Peru. Manipulation: <u>Content</u> : No money vs. Money (two \$1 bills); <u>Sender's name</u> : Foreign name vs. Same family name as the recipient.	Theft of pieces of mail.	Money content increased theft of pieces of mail. Sender's last name matching the recipient's increased theft of pieces of mail.	*Benefits

Cohn et al. (2014)	96 hired workers	Task: Participants were hired to sell promotional cards that permitted entrance to nightclubs. After a baseline treatment where workers got paid the same hourly wage, the firm introduced cuts of 25%.	Theft of cash sales.	Unilateral wage cut increased theft of cash sales.	NA
	Germany	Manipulation: <u>Wage-cut target:</u> General (both group members) vs. Unilateral (only one group member).		Within unilateral group, the workers that received the cuts were more likely to steal money from the firm.	
Cohn et al. (2019)	Employees	Task: Confederates turned in “lost” transparent wallets to an employee at the counter.	Lost wallet technique.		
	<u>Field Exp. 1</u> 17,303 wallets in 40 countries	“lost” 355 Manipulation: <u>Wallet content:</u> No money vs. Money (\$13.45).		Money conditions decreased wallet theft.	*Benefits (opposite effect)
	<u>Field Exp. 2</u> 2,932 wallets in US, UK, and Poland	“lost” Manipulation: <u>Wallet content:</u> No money vs. Money (\$13.45) vs. Big-Money (\$94.15).		Big-Money conditions decreased wallet theft.	*Benefits (opposite effect)
	<u>Field Exp. 3</u> 2,932 wallets in US, UK, and Poland	“lost” Manipulation: <u>Wallet content:</u> No money vs. Money-NoKey vs. Money-Key.		Key conditions decreased wallet theft.	*Costs for the other
Gabor and Barker (1989)	Members of the public (112 lost letters)	Task: “Lost” letters were planted under the windshield wipers of cars with message “found near your car”.	Adapted “Lost” letter technique.	Apparent value of coin did not affect letter theft.	Benefits
	Canada				

		<p>Manipulation: <u>Coin value</u>: Low (one penny) vs. High (one penny valued at \$150).</p>		<p>Younger subjects were less likely to return the "lost" letters.</p>	
Geller et al. (1983)	<p>Newspaper customers (166 days of observation) USA</p>	<p>Task: Newspaper theft from free-access racks was observed. After a baseline treatment, sign messages were posted.</p> <p>Manipulation: <u>Antitheft sign messages</u>: Internal control (stated politely) vs. External control (stating consequences, e.g., legal threat)</p>	<p>Theft of newspapers.</p>	<p>Internal control messages decreased newspaper theft.</p> <p>External control messages decreased newspaper theft.</p>	<p>NA</p>
Greenberg (1990)	<p>143 employees of manufacturing plants. USA</p>	<p>Task: Plants introduced temporarily cutting wages by 15%. Employee theft was measured before, during, and after the pay cut.</p> <p>Manipulation: <u>Explanation for wage cut</u>: Control (No wage cut) vs. Adequate explanation vs. Inadequate explanation.</p>	<p>Theft of firm inventory.</p>	<p>Introducing wage cuts increased employee theft.</p> <p>Inadequate-explanation increased employee theft.</p>	<p>NA</p>
Greenberg (2002)	<p>270 employees of a financial</p>	<p>Task: Employees completed a survey in exchange for \$2. Participants took</p>	<p>Theft of coins.</p>	<p>Payment coming from the company increased employee theft.</p>	<p>*Costs for the other</p>

USA	services company	<p>payment in private from a bowl of pennies.</p> <p>Manipulations: <u>Victim of theft</u>: Organization vs. Individual (money was being paid from personal funds); <u>Corporate ethics program</u>: Control vs. Office in which an ethics program in place.</p>		Working in an office in which there was no ethics program in place increased employee theft.	
Keizer et al. (2008)	203 members of the public	<p>Task: Participants passed by a mailbox and noticed an envelope visibly containing a 5€ note hanging out of a mailbox.</p> <p>Manipulation: <u>Norm violation</u>: Control (clean) vs. DisorderGraff (mailbox covered with graffiti) vs. DisorderLitter (litter on the ground).</p>	Adapted “lost” letter technique.	<p>Graffiti disorder increased stealing compared to control.</p> <p>Litter disorder increased stealing compared to control.</p>	NA
Keuschnigg and Wolbring (2015)	270 members of the public	<p>Task: Participants passed by a mailbox and noticed an envelope visibly containing money in front of a mailbox.</p> <p>Manipulation:</p>	Adapted “lost” letter technique.	<p>Disorder condition increased stealing.</p> <p>Disorder effect was stronger for 5€ condition and marginally significant for 10€ condition.</p>	*Benefits (but only significant in an interaction with the disorder manipulation)
	Netherlands				
	Germany				

		<p><u>Amount money</u>: 5€ vs. 10€ vs. 100€.</p> <p><u>Norm violation</u>: Control (clean) vs. Disorder (two heavily wrecked bicycles next to the mailboxes).</p>		<p>Within 100€ condition, disorder did not affect stealing.</p>	
Lanfear (2018)	2786 members of the public	<p>Task: Participants passed by a mailbox and noticed an envelope visibly containing a \$5 note near the mailbox.</p> <p>Manipulation: <u>Norm violation</u>: Control (clean) vs. Disorder (graffiti and litter)</p>	Adapted “lost” letter technique.	<p>Norm violation did not affect stealing.</p> <p>Disorder condition decreased pro-social behavior (i.e., mailing the dropped envelope).</p>	NA
USA					
Pruckner and Sausgruber (2013)	Newspaper customers (120 observations)	<p>Task: Newspaper transactions in booths on the streets via an “honor system” where costumers are supposed to make a payment without monitoring.</p> <p>Manipulation: <u>Reminder</u>: Control (“The paper costs €0.60.”) vs. Legal (“... Stealing a paper is illegal”) vs. Moral (“... Thank you for being honest”).</p>	Theft of newspapers.	<p>Legal reminder did not affect newspaper theft.</p> <p>Moral reminder did not affect newspaper theft.</p> <p>Moral reminder increased the average amount paid.</p>	NA
Austria					
Schlüter and Vollan (2015)	Flower customers (336 observations)	<p>Task:</p>	Theft of flowers.	<p>Both reminder messages (i.e., legal</p>	NA

Germany		In an unattended flower field, customers picked and paid for the flowers via an "honor system".		and moral) did not affect theft of flowers.	
		Manipulation: <u>Reminder</u> : Control (No message) vs. Legal (threatening message) vs. Moral (thankful message); <u>Who is asking</u> : Control vs. Family vs. Business (consulting firm).		Family business condition decreased theft of flowers.	
Widner (1998)	National Park visitors (40 days of observation) USA	Task: The behavior of visitors was directly observed.	Theft of petrified wood.	All interventions decreased theft of petrified wood, when compared to control.	*Costs for the self
		Manipulation: <u>Anti-theft interventions</u> : Control vs. Uniformed Volunteer vs. Sign (depicting the progressive loss of petrified wood) vs. Pledge (visitors signed an anti-theft pledge before entering the park).		No differences between intervention effectiveness were found.	
Wortley and McFarlane (2011)	University students (2,098 minutes of observation) Australia	Task: In a University library, students passed by an unattended photocopy card.	Theft of photocopy cards.	Unsigned cards increased card theft.	*Costs for the other
		Manipulation: <u>Territoriality ownership</u> : Signed ("M. Smith") vs. Unsigned card;		Card on its own increased card theft.	

Territoriality guardianship: Card next to two library books vs. Card on its own.

Note. SEU = Availability of Costs or Benefits Manipulation; NA = not applicable; * = the manipulation of costs or benefits was significant.

Table 4*Summary of field experiments in the Keeping money category*

Study	Participants	Design	Measure	Main findings	SEU
Alem et al. (2018)	225 farmers Tanzania	Participants received an amount of money on their phone. Then, they received an SMS asking to return the money. Manipulation: <u>Message frame</u> : Control (neutral message) vs. Kindness (gift of 25%) vs. Guilt.	Keeping money wrongly received.	Kindness framed message reduced unethical behavior compared to control. Guilt inducing message reduced unethical behavior compared to control.	NA
Armantier and Boly (2011)	247 adults with university degrees or enrolled at a university Burkina Faso	Task: Participants were recruited to grade a set of 20 exam papers. The 11th paper came with a bribe and a request to find few mistakes. Manipulations: <u>Amount of bribe</u> – No bribe vs. Low bribe vs. High bribe; <u>Wage</u> – Low vs. High wage; <u>Monitoring</u> – Low vs. High monitoring.	Acceptance of bribe.	High bribe amount increased acceptance of bribe. High wage decreased acceptance of bribe. Monitoring and punishment decreased acceptance of bribe.	*Benefits *Costs for the self

Azar et al. (2013)	192 customers at a restaurant Israel	Task: After paying, customers received extra change. Manipulation: <u>Extra change</u> : Low (\$3) vs. High (\$12) amount of change.	Keeping extra change.	High amount of extra change decreased unethical behavior.	*Benefits (but the effect is in the opposite direction)
Gabor et al. (1986)	Cashiers at 125 convenience stores Canada	Task: A confederate bought a newspaper (\$0.30) with a single dollar bill and left without awaiting the change. Manipulation: <u>Sex of confederate</u> : female vs. male; <u>Type of store</u> : chain type vs. family store.	Keeping due change.	Male confederates increased cashiers' dishonesty. Type of store did not affect dishonesty.	Costs for the other
Gire and Williams (2007)	Colleges' members (80 lost bills) USA	Task: Money (one dollar note) was left at the campuses. Manipulation: <u>Type of college</u> : Military vs. Nonmilitary; <u>Type of setting</u> : Public (sidewalk) vs. Private (bathroom).	"Lost" dollar bill.	Within military colleges, private setting increased dishonesty. Within nonmilitary colleges, type of setting did not affect dishonesty.	NA
Newman (1979)	80 university students and	Task:	Picking a "dropped" coin.	City site increased dishonesty.	*Benefits

UK	adult members of the public		A female confederate dropped a coin while approaching an unsuspecting participant.			Higher value of coin increased dishonesty.
			Manipulation: <u>Amount of money</u> : Low (2p) vs. High (10p); <u>Site</u> : University campus vs. City (shopping area).			
Rabinowitz et al. (1993)	96 female souvenir shop cashiers.	Austria	Task: Confederates purchased two postcards costing 4 shillings (\$.36 USD) and left without awaiting the request or return of the money.	Keeping change.	due	Payment did not affect dishonesty. Female confederates increased dishonesty.
			Manipulation: <u>Sex of confederate</u> : female vs. male; <u>Payment</u> : Overpayment (+1) vs. Underpayment (-1 shilling).			*Benefits
Yap et al. (2013)	88 members of the public	USA	Task: Participants were recruited for a study in exchange for \$4. While making the payment, the experimenter "accidentally" handed \$8.	Keeping extra money.		Holding an expansive pose increased dishonesty. NA
			Manipulation:			

Postural expansiveness: Hold an expansive pose vs. Hold an contractive pose for 1 min.

Yuchtman-Yaar and Rahav (1986)	328 passengers	bus	Task: Bus drivers gave passengers extra change.	Keeping extra change.	Within females, higher incentives increased dishonesty.	*Benefits (but in opposite direction for males)
Israel			Manipulation: <u>Incentive</u> : Low (extra change was 7% of the fare) vs. High (extra change was 25% of the fare).		Within males, higher incentives decreased dishonesty.	

Note. SEU = Availability of Costs or Benefits Manipulation; NA = not applicable; * = the manipulation of costs or benefits was significant.

Table 5*Summary of field experiments in the Shoplifting category*

Study	Participants	Design	Measure	Main findings	SEU
Carter et al. (1980)	Customers of a grocery store Sweden	Multiple baseline design. Intervention: <u>Public identification</u> : signs and red dots alerting customers for frequently shoplifted items.	Shoplifting.	Public identification reduced shoplifting.	NA
Carter and Holmberg (1993)	Customers of a grocery store Sweden	Pretest / Posttest design. Intervention: <u>Public identification</u> : signs and red dots alerting customers for frequently shoplifted items.	Shoplifting.	Public identification reduced shoplifting.	NA
Carter et al. (1988)	Employees of a grocery store. Sweden	Multiple baseline design. Intervention: <u>Product identification</u> : Oral presentation, list of target items, and data on losses graphed biweekly the lunchroom.	Employee theft.	Information on product identification reduced employee shoplifting.	NA
DiLonardo and Clarke (1996)	Customers of 4 stores	Pretest / Posttest design. Intervention: Replacement of EAS (i.e. Electronic Article Surveillance) with ink tags.	Shoplifting	Ink tags reduced shoplifting when compared to EAS condition.	*Costs for the self

Farrington et al. (1993)	Customers of 9 stores UK	Pretest / Posttest design. Intervention: <u>Anti-shoplifting intervention</u> : Control vs. Electronic tagging vs. Store redesign vs. Uniformed guard.	Shoplifting	Electronic tagging reduced shoplifting, maintained over time. Store redesign reduced shoplifting, but was not maintained over time. Uniformed guard did not affect shoplifting.	*Costs for the self
Hayes and Blackwood (2006)	Customers of 21 stores USA	Pretest / Posttest design. Intervention: <u>Source-tagged products</u> : Control vs. 50% (half the products received a hidden EAS) vs. 100% (all products received a hidden EAS).	Shoplifting.	Electronic Article Surveillance did not affect shoplifting when compared to control.	NA
Hayes and Downs (2011)	Customers of 47 stores USA	Randomized Controlled Trial Intervention: <u>Anti-shoplifting intervention</u> : Control vs. In-aisle closed-circuit television (CCTV) public view monitor vs. In-aisle CCTV dome vs. Keeper/safer box.	Shoplifting	All three interventions reduced shoplifting	*Costs for the self
Hayes et al. (2011)	Customers of 10 stores	Randomized Controlled Trial	Shoplifting	Keeper/safer boxes reduced shoplifting	*Costs for the self

	USA		Intervention: <u>Anti-shoplifting intervention</u> : Control vs. Keeper/safer box.					
Hayes et al. (2012)	Customers of 57 drug stores	Randomized Controlled Trial	Shoplifting.	Protective handling shoplifting	product reduced	*Costs for the self		
	USA		Intervention: <u>Anti-shoplifting intervention</u> : Control vs. Protective product display.		Protective display fixtures reduced shoplifting.			
Hayes et al. (2019)	Customers of 60 retail stores	Randomized Controlled Trial	Shoplifting.	Anti-theft reduced shoplifting, but only in specific product categories.	wire-wraps	*Costs for the self		
	USA		Intervention: <u>Anti-theft wraps</u> : Control vs. Wire-wraps.					
Johns et al. (2017)	Customers of 42 stores	Randomized Controlled Trial	Shoplifting.	Protective fixtures shoplifting.	display reduced	*Costs for the self		
	USA		Intervention: <u>Anti-shoplifting intervention</u> : Control vs. Enhanced public view monitor (ePVM).		ePVM did not significantly affect shoplifting.			
McNees et al. (1980a)	Employees of a fast-food snack bar located on a university campus	Multiple baseline design.	Employee theft.	Implementation of product specific signs decreased employee theft.		NA		
	USA		Intervention: <u>Product identification</u> : Sequential signs introduced at different time					

points describing employee theft rates.

McNees et al. (1980b)	Customers of a food market near an elementary school USA	Pretest / Posttest design. Intervention: <u>Anti-shoplifting intervention</u> : Program (poster and buyers received tokens that could be exchange for prizes). After 12 weeks, the program was terminated and its results were followed-up for 10 days.	Shoplifting.	Anti-shoplifting program reduced students' shoplifting. After program termination, the average shoplifting increased.	*Costs for the self
Thurber and Snow (1980)	Customers of a retail supermarket USA	Pretest / Posttest design. Intervention: Anti-shoplifting signs. After 2 weeks, the intervention was terminated and its results were followed-up for 1 week.	Shoplifting.	Anti-shoplifting intervention increased shoplifting of cigarettes. After intervention termination, the average shoplifting decreased.	NA

Note. SEU = Availability of Costs or Benefits Manipulation; NA = not applicable; * = the manipulation of costs or benefits was significant.

References

- Alem, Y., Eggert, H., Kocher, M. G., & Ruhinduka, R. D. (2018). Why (field) experiments on unethical behavior are important: Comparing stated and revealed behavior. *Journal of Economic Behavior & Organization*, *156*, 71–85. <https://doi.org/10.1016/j.jebo.2018.08.026>
- Armantier, O., & Boly, A. (2011). A controlled field experiment on corruption. *European Economic Review*, *55*(8), 1072–1082. <https://doi.org/10.1016/j.euroecorev.2011.04.007>
- Armantier, O., & Boly, A. (2013). Comparing corruption in the laboratory and in the field in Burkina Faso and in Canada. *The Economic Journal*, *123*(573), 1168–1187. <https://doi.org/10.1111/eoj.12019>
- Azar, O. H., Yosef, S., & Bar-Eli, M. (2013). Do customers return excessive change in a restaurant?: A field experiment on dishonesty. *Journal of Economic Behavior & Organization*, *93*, 219–226. <https://doi.org/10.1016/j.jebo.2013.03.031>
- Balafoutas, L., Beck, A., Kerschbamer, R., & Sutter, M. (2013). What drives taxi drivers? A field experiment on fraud in a market for credence goods. *The Review of Economic Studies*, *80*(3), 876–891. <https://doi.org/10.1093/restud/rds049>
- Belot, M., & Schröder, M. (2015). The spillover effects of monitoring: A field experiment. *Management Science*, *62*(1), 37–45. <https://doi.org/10.1287/mnsc.2014.2089>
- Bertrand, M., Djankov, S., Hanna, R., & Mullainathan, S. (2007). Obtaining a driver's license in India: An experimental approach to studying corruption. *The Quarterly Journal of Economics*, *122*(4), 1639–1676. <https://doi.org/10.1162/qjec.2007.122.4.1639>
- Bickman, L. (1971). The effect of social status on the honesty of others. *The Journal of Social Psychology*, *85*(1), 87–92. <https://doi.org/10.1080/00224545.1971.9918547>
- Blais, E., & Bacher, J. L. (2007). Situational deterrence and claim padding: Results from a randomized field experiment. *Journal of Experimental Criminology*, *3*(4), 337–352. <https://doi.org/10.1007/s11292-007-9043-z>
- Buccioli, A., & Piovesan, M. (2011). Luck or cheating? A field experiment on honesty with children. *Journal of Economic Psychology*, *32*(1), 73–78. <https://doi.org/10.1016/j.joep.2010.12.001>
- Cagala, T., Glogowsky, U., & Rincke, J. (2014). A field experiment on intertemporal enforcement spillovers. *Economics Letters*, *125*(2), 171–174. <https://doi.org/10.1016/j.econlet.2014.08.034>

- Carter, N., Hansson, L., Holmberg, B., & Melin, L. (1980). Shoplifting reduction through the use of specific signs. *Journal of Organizational Behavior Management*, 2(2), 73–84. https://doi.org/10.1300/J075v02n02_01
- Carter, N., & Holmberg, B. (1993). Theft reduction in a grocery store through product identification. *Journal of Organizational Behavior Management*, 13(1), 129–135. https://doi.org/10.1300/J075v13n01_08
- Carter, N., Holmström, A., Simpanen, M., & Melin, L. (1988). Theft reduction in a grocery store through product identification and graphing of losses for employees. *Journal of Applied Behavior Analysis*, 21(4), 385–389. <https://doi.org/10.1901/jaba.1988.21-385>
- Castillo, M., Petrie, R., Torero, M., & Viceisza, A. (2014). Lost in the mail: A field experiment on crime. *Economic Inquiry*, 52(1), 285–303. <https://doi.org/10.1111/ecin.12046>
- Christensen, L. B. (1985). *Experimental methodology* (3rd ed.). Allyn & Bacon.
- Chytilová, J., & Korbela, V. (2014). *Individual and group cheating behavior: A field experiment with adolescents* (IES Working Paper No. 06/2014). Institute of Economic Studies. <https://www.econstor.eu/bitstream/10419/102591/1/78366057X.pdf>
- Clarke, R. V. (1995). Situational crime prevention. *Crime and Justice*, 19, 91–150. <https://doi.org/10.1086/449230>
- Clarke, R. V., & Cornish, D. B. (1985). Modeling offenders' decisions: A framework for research and policy. *Crime and Justice*, 6, 147–185. <https://doi.org/10.1086/449106>
- Cohn, A., Fehr, E., Herrmann, B., & Schneider, F. (2014). Social comparison and effort provision: Evidence from a field experiment. *Journal of the European Economic Association*, 12(4), 877–898. <https://doi.org/10.1111/jeea.12079>
- Cohn, A., Maréchal, M. A., Tannenbaum, D., & Zünd, C. L. (2019). Civic honesty around the globe. *Science*, 365(6448), 70-73. <https://doi.org/10.1126/science.aau8712>
- Conrads, J., Ebeling, F., & Lotz, S. (2015). (Dis-) honesty: Measuring overcharging in a real-world market. *Journal of Behavioral and Experimental Economics*, 57, 98–102. <https://doi.org/10.1016/j.socec.2015.05.003>
- Defoe, I. N. (in press). Towards a hybrid criminological and psychological model of risk behavior: The developmental neuro-ecological risk-taking model (DNERM). *Developmental Review*.
- Defoe, I. N., Dubas, J. S., & Romer, D. (2019). Heightened adolescent risk-taking? Insights from lab studies on age differences in decision-making. *Policy Insights from the Behavioral and Brain Sciences*, 6(1), 56-63. <https://doi.org/10.1177/2372732218801037>

- Diener, E., Fraser, S. C., Beaman, A. L., & Kelem, R. T. (1976). Effects of deindividuation variables on stealing among Halloween trick-or-treaters. *Journal of Personality and Social Psychology*, *33*(2), 178-183. <https://doi.org/10.1037/0022-3514.33.2.178>
- DiLonardo, R. L., & Clarke, R. V. (1996). Reducing the rewards of shoplifting: An evaluation of ink tags. *Security Journal*, *7*(1), 11–14. https://popcenter.asu.edu/sites/default/files/60-dilonardo_clarke-reducing_the_rewards_of_shoplifting_a.pdf
- Dugar, S., & Bhattacharya, H. (2017). Fishy behavior: A field experiment on (dis) honesty in the marketplace. *Journal of Behavioral and Experimental Economics*, *67*, 41–55. <https://doi.org/10.1016/j.socec.2017.02.002>
- Dur, R., & Vollaard, B. (2019). Salience of law enforcement: A field experiment. *Journal of Environmental Economics and Management*, *93*, 208-220. <https://doi.org/10.1016/j.jeem.2018.11.011>
- Farrington, D. P. (1979). Experiments on deviance with special reference to dishonesty. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 12, pp. 207-252). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60263-4](https://doi.org/10.1016/S0065-2601(08)60263-4)
- Farrington, D. P. (1980). External validity: A problem for social psychology. In R. F. Kidd & M. J. Saks (Eds.), *Advances in applied social psychology* (Vol. 1, pp. 184-186). Lawrence Erlbaum. <https://doi.org/10.4324/9781315803005>
- Farrington, D. P. (2008). Criminology as an experimental science. In C. Horne & M. J. Lovaglia (Eds.), *Experiments in criminology and law: A research revolution* (pp. 175–179). Rowman and Littlefield.
- Farrington, D. P., Bowen, S., Buckle, A., Burns-Howell, T., Burrows, J., & Speed, M. (1993). An experiment on the prevention of shoplifting. In R. V. Clarke (Ed.), *Crime Prevention Studies* (Vol. 1, pp. 93–119). Criminal Justice Press.
- Farrington, D. P., & Kidd, R. F. (1977). Is financial dishonesty a rational decision? *British Journal of Social and Clinical Psychology*, *16*(2), 139-146. <https://doi.org/10.1111/j.2044-8260.1977.tb00209.x>
- Farrington, D. P., & Knight, B. J. (1979). Two non-reactive field experiments on stealing from a 'lost' letter. *British Journal of Social and Clinical Psychology*, *18*(3), 277-284. <https://doi.org/10.1111/j.2044-8260.1979.tb00337.x>
- Farrington, D. P., & Knight, B. J. (1980). Four studies of stealing as a risky decision. In P. Lipsitt & B. D. Sales (Eds.), *New directions in psycholegal research* (pp. 26–50). Van Nostrand-Reinhold.

- Farrington, D. P., Lösel, F., Braga, A. A., Mazerolle, L., Raine, A., Sherman, L. W., & Weisburd, D. (2020). Experimental criminology: Looking back and forward on the 20th anniversary of the Academy of Experimental Criminology. *Journal of Experimental Criminology*, *16*, 649–673. <https://doi.org/10.1007/s11292-019-09384-z>
- Feldman, R. E. (1968). Response to compatriot and foreigner who seek assistance. *Journal of Personality and Social Psychology*, *10*(3), 202-214. <https://doi.org/10.1037/h0026567>
- Franklin, B. J. (1973). The effects of status on the honesty and verbal responses of others. *The Journal of Social Psychology*, *91*(2), 347-348. <https://doi.org/10.1080/00224545.1973.9923060>
- Franzen, A., & Pointner, S. (2013). The external validity of giving in the dictator game. *Experimental Economics*, *16*(2), 155-169. <https://doi.org/10.1007/s10683-012-9337-5>
- Gabor, T., & Barker, T. (1989). Probing the public's honesty: A field experiment using the "lost letter" technique. *Deviant Behavior*, *10*(4), 387–399. <https://doi.org/10.1080/01639625.1989.9967824>
- Gabor, T., Streat, J., Singh, G., & Varis, D. (1986). Public deviance: An experimental study. *Canadian Journal of Criminology*, *28*, 17–29. <http://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=12068737>
- Geller, E. S., Koltuniak, T. A., & Shilling, J. S. (1983). Response avoidance prompting: A cost-effective strategy for theft deterrence [Abstract]. *Behavioral Counseling & Community Interventions*, *3*(1), 28–42. <https://psycnet.apa.org/record/1985-12586-001>
- Gire, J. T., & Williams, T. D. (2007). Dissonance and the honor system: Extending the severity of threat phenomenon. *The Journal of Social Psychology*, *147*(5), 501–509. <https://doi.org/10.3200/SOCP.147.5.501-510>
- Gomes, H. S., Farrington, D. P., Maia, Â., & Krohn, M. D. (2019). Measurement bias in self-reports of offending: A systematic review of experiments. *Journal of Experimental Criminology*, *15*(3), 313-339. <https://doi.org/10.1007/s11292-019-09379-w>
- Gomes, H. S., Maia, Â., & Farrington, D. P. (2018). Measuring offending: Self-reports, official records, systematic observation and experimentation. *Crime Psychology Review*, *4*(1), 26-44. <https://doi.org/10.1080/23744006.2018.1475455>
- Green, G. S. (1985). General deterrence and television cable crime: A field experiment in social control. *Criminology*, *23*(4), 629–645. <https://doi.org/10.1111/j.1745-9125.1985.tb00367.x>

- Greenberg, J. (1990). Employee theft as a reaction to underpayment inequity: The hidden cost of pay cuts. *Journal of Applied Psychology, 75*(5), 561–568. <https://doi.org/10.1037/0021-9010.75.5.561>
- Greenberg, J. (2002). Who stole the money, and when? Individual and situational determinants of employee theft. *Organizational Behavior and Human Decision Processes, 89*(1), 985–1003. [https://doi.org/10.1016/S0749-5978\(02\)00039-0](https://doi.org/10.1016/S0749-5978(02)00039-0)
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009–1055. <https://doi.org/10.1257/0022051043004577>
- Hayes, R., & Blackwood, R. (2006). Evaluating the effects of EAS on product sales and loss: Results of a large-scale field experiment. *Security Journal, 19*(4), 262–276. <https://doi.org/10.1057/palgrave.sj.8350025>
- Hayes, R., & Downs, D. (2011). Controlling retail theft with CCTV domes, public view monitors and protective containers: A randomized controlled trial. *Security Journal, 24*(3), 237–250. <https://doi.org/10.1057/sj.2011.12>
- Hayes, R., Downs, D. M., & Blackwood, R. (2012). Anti-theft procedures and fixtures: A randomized controlled trial of two situational crime prevention measures. *Journal of Experimental Criminology, 8*(1), 1–15. <https://doi.org/10.1007/s11292-011-9137-5>
- Hayes, R., Johns, T., Scicchitano, M., Downs, D., & Pietrawska, B. (2011). Evaluating the effects of protective Keeper boxes on ‘hot product’ loss and sales: A randomized controlled trial. *Security Journal, 24*(4), 357–369. <https://doi.org/10.1057/sj.2011.2>
- Hayes, R., Strome, S., Johns, T., Scicchitano, M., & Downs, D. (2019). Testing the effectiveness of anti-theft wraps across product types in retail environments: A randomized controlled trial. *Journal of Experimental Criminology, 15*(4), 703–718. <https://doi.org/10.1007/s11292-019-09365-2>
- Hornstein, H. A., Fisch, E., & Holmes, M. (1968). Influence of a model's feeling about his behavior and his relevance as a comparison other on observers' helping behavior. *Journal of Personality and Social Psychology, 10*(3), 222–226. <https://doi.org/10.1037/h0026568>
- Houser, D., List, J. A., Piovesan, M., Samek, A., & Winter, J. (2016). Dishonesty: From parents to children. *European Economic Review, 82*, 242–254. <https://doi.org/10.1016/j.euroecorev.2015.11.003>
- Jesilow, P., & O'Brien, M. J. (1980). *Deterring automobile repair fraud - A field experiment* (NIJ Reference Service No. 89242). National Institute of Justice. <https://www.ncjrs.gov/pdffiles1/Digitization/89242NCJRS.pdf>

- Johns, T., Hayes, R., Scicchitano, M., & Grottini, K. (2017). Testing the effectiveness of two retail theft control approaches: An experimental research design. *Journal of Experimental Criminology*, *13*(2), 267–273. <https://doi.org/10.1007/s11292-017-9284-4>
- Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, *322*(5908), 1681–1685. <https://doi.org/10.1126/science.1161405>
- Kerschbamer, R., Neururer, D., & Sutter, M. (2016). Insurance coverage of customers induces dishonesty of sellers in markets for credence goods. *Proceedings of the National Academy of Sciences*, *113*(27), 7454–7458. <https://doi.org/10.1073/pnas.1518015113>
- Keuschnigg, M., & Wolbring, T. (2015). Disorder, social capital, and norm violation: Three field experiments on the broken windows thesis. *Rationality and Society*, *27*(1), 96–126. <https://doi.org/10.1177/1043463114561749>
- Korbel, V. (2013). *Children and cheating: A field experiment with individuals and teams* [Master's thesis, Charles University in Prague]. Thesis Repository of Charles University in Prague. https://dspace.cuni.cz/bitstream/handle/20.500.11956/58563/DPTX_2011_2_11230_0_355601_0_125761.pdf?sequence=1&isAllowed=y
- Korte, C., & Kerr, N. (1975). Response to altruistic opportunities in urban and nonurban settings. *The Journal of Social Psychology*, *95*(2), 183–184. <https://doi.org/10.1080/00224545.1975.9918701>
- Lanfear, C. C. (2018). *Disorder in the neighborhood: A large-scale field experiment on disorder, norm violation, and pro-social behavior* [Master's thesis, University of Washington]. Research Works Archive of the University of Washington. <http://hdl.handle.net/1773/40974>
- Lenga, M. R., & Kleinke, C. L. (1974). Modeling, anonymity, and performance of an undesirable act. *Psychological Reports*, *34*(2), 501–502. <https://doi.org/10.2466/pr0.1974.34.2.501>
- Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world?. *Journal of Economic Perspectives*, *21*(2), 153–174. <https://doi.org/10.1257/jep.21.2.153>
- List, J. A. (2007). Field experiments: A bridge between lab and naturally occurring data. *Advances in Economic Analysis & Policy*, *5*(2), 1–47. <https://doi.org/10.2202/1538-0637.1747>
- List, J. A., & Momeni, F. (2017). *When corporate social responsibility backfires: Theory and evidence from a natural field experiment* (Working paper 24169). National Bureau of Economic Research. <https://www.nber.org/papers/w24169.pdf>

- List, J. A., & Momeni, F. (2020). Leveraging upfront payments to curb employee misbehavior: Evidence from a natural field experiment. *European Economic Review*, *130*, 103601. <https://doi.org/10.1016/j.euroecorev.2020.103601>
- McNees, P., Gilliam, S. W., Schnelle, J. F., & Risley, T. R. (1980a). Controlling employee theft through time and product identification. *Journal of Organizational Behavior Management*, *2*(2), 113–119. https://doi.org/10.1300/J075v02n02_04x
- McNees, M. P., Kennon, M., Schnelle, J. F., Kirchner, R. E., & Thomas, M. M. (1980b). An experimental analysis of a program to reduce retail theft. *American Journal of Community Psychology*, *8*(3), 379–385. <https://doi.org/10.1007/BF00894349>
- Nagin, D. S., Rebitzer, J. B., Sanders, S., & Taylor, L. J. (2002). Monitoring, motivation, and management: The determinants of opportunistic behavior in a field experiment. *American Economic Review*, *92*(4), 850–873. <https://doi.org/10.1257/00028280260344498>
- Newman, C. V. (1979). Relation between altruism and dishonest profiteering from another's misfortune. *The Journal of Social Psychology*, *109*(1), 43–48. <https://doi.org/10.1080/00224545.1979.9933637>
- Okeke, E. N., & Godlonton, S. (2014). Doing wrong to do right? Social preferences and dishonest behavior. *Journal of Economic Behavior & Organization*, *106*, 124–139. <https://doi.org/10.1016/j.jebo.2014.06.011>
- Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy*, *115*(2), 200–249. <https://doi.org/10.1086/517935>
- Pierce, L., & Balasubramanian, P. (2015). Behavioral field evidence on psychological and social factors in dishonesty and misconduct. *Current Opinion in Psychology*, *6*, 70–76. <https://doi.org/10.1016/j.copsyc.2015.04.002>
- Pruckner, G. J., & Sausgruber, R. (2013). Honesty on the streets: A field study on newspaper purchasing. *Journal of the European Economic Association*, *11*(3), 661–679. <https://doi.org/10.1111/jeea.12016>
- Rabinowitz, F. E., Colmar, C., Elgie, D., Hale, D., Niss, S., Sharp, B., & Sinclitico, J. (1993). Dishonesty, indifference, or carelessness in souvenir shop transactions. *The Journal of Social Psychology*, *133*(1), 73–79. <https://doi.org/10.1080/00224545.1993.9712120>
- Ramos, J., & Torgler, B. (2012). Are academics messy? Testing the broken windows theory with a field experiment in the work environment. *Review of Law & Economics*, *8*(3), 563–577. <https://doi.org/10.1515/1555-5879.1617>

- Robins, L. N. (1992). The role of prevention experiments in discovering causes of children's antisocial behavior. In J. McCord & R. E. Tremblay (Eds.), *Preventing antisocial behavior: Interventions from birth through adolescence* (pp. 3–18). Guilford Press.
- Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology, 45*, 181-196. <https://doi.org/10.1016/j.joep.2014.10.002>
- Schlüter, A., & Vollan, B. (2015). Flowers and an honour box: Evidence on framing effects. *Journal of Behavioral and Experimental Economics, 57*, 186–199. <https://doi.org/10.1016/j.socec.2014.10.002>
- Schneider, H. S. (2012). Agency problems and reputation in expert services: Evidence from auto repair. *The Journal of Industrial Economics, 60*(3), 406–433. <https://doi.org/10.1111/j.1467-6451.2012.00485.x>
- Shu, L. L., Mazar, N., Gino, F., Ariely, D., & Bazerman, M. H. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences, 109*(38), 15197–15200. <https://doi.org/10.1073/pnas.1209746109>
- Siniver, E., & Yaniv, G. (2018). Losing a real-life lottery and dishonest behavior. *Journal of Behavioral and Experimental Economics, 75*, 26–30. <https://doi.org/10.1016/j.socec.2018.05.005>
- Steinberg, J., McDonald, P., & O'Neal, E. (1977). Petty theft in a naturalistic setting: The effects of bystander presence. *The Journal of Social Psychology, 101*(2), 219-221. <https://doi.org/10.1080/00224545.1977.9924010>
- Thurber, S., & Snow, M. (1980). Signs may prompt antisocial behavior. *The Journal of Social Psychology, 112*(2), 309–310. <https://doi.org/10.1080/00224545.1980.9924336>
- Tracy, P. E., & Fox, J. A. (1989). A field experiment on insurance fraud in auto body repair. *Criminology, 27*(3), 589–603. <https://doi.org/10.1111/j.1745-9125.1989.tb01047.x>
- Tremblay, P., Bacher, J. L., Tremblay, M., & Cusson, M. (2000). Inflated claims of theft and tolerance threshold of insurers: Experimental analysis of situational deterrence. *Canadian Journal of Criminology-Revue Canadienne de Criminologie, 42*(1), 21–38.
- Weisburd, D. (2003). Ethical practice and evaluation of interventions in crime and justice: The moral imperative for randomized trials. *Evaluation Review, 27*(3), 336–354. <https://doi.org/10.1177/0193841X03027003007>

- Weisburd, D. (2005). Hot spots policing experiments and criminal justice research: Lessons from the field. *The ANNALS of the American Academy of Political and Social Science*, 599(1), 220–245. <https://doi.org/10.1177/0002716205274597>
- Widner, C. J. (1998). *Reducing and understanding petrified wood theft at Petrified Forest National Park* [Doctoral dissertation, Virginia Tech]. VTechWorks of Virginia Tech. <http://hdl.handle.net/10919/40250>
- Wortley, R., & McFarlane, M. (2011). The role of territoriality in crime prevention: A field experiment. *Security Journal*, 24(2), 149–156. <https://doi.org/10.1057/sj.2009.22>
- Yap, A. J., Wazlawek, A. S., Lucas, B. J., Cuddy, A. J., & Carney, D. R. (2013). The ergonomics of dishonesty: The effect of incidental posture on stealing, cheating, and traffic violations. *Psychological Science*, 24(11), 2281–2289. <https://doi.org/10.1177/0956797613492425>
- Yuchtman-Yaar, E., & Rahav, G. (1986). Resisting small temptations in everyday transactions. *The Journal of Social Psychology*, 126(1), 23–30. <https://doi.org/10.1080/00224545.1986.9713565>
- Zimbardo, P. G. (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In W. J. Arnold & D. Levine (Eds.), *Nebraska Symposium on Motivation* (Vol. 17, pp. 237–307). University of Nebraska Press.
- Zimny, G. H. (1961). *Method in experimental psychology*. Ronald Press Company. <https://doi.org/10.1037/14006-000>

CHAPTER III
MEASUREMENT BIAS IN SELF-REPORTS OF OFFENDING:
A SYSTEMATIC REVIEW OF EXPERIMENTS

Manuscript Published in:

Gomes, H. S., Farrington, D. P., Maia, Â., & Krohn, M. D. (2019). Measurement bias in self-reports of offending: A systematic review of experiments. *Journal of Experimental Criminology*, *15*(3), 313-339.

<https://doi.org/10.1007/s11292-019-09379-w>

Abstract

Objectives. Self-reported offending is one of the primary measurement methods in criminology. In this article, we aimed to systematically review the experimental evidence regarding measurement bias in self-reports of offending.

Methods. We carried out a systematic search for studies that (a) included a measure of offending, (b) compared self-reported data on offending between different methods, and (c) used an experimental design. Effect sizes were used to summarize the results.

Results. The 21 pooled experiments provided evidence regarding 18 different types of measurement manipulations which were grouped into three categories, i.e., Modes of administration, Procedures of data collection, and Questionnaire design. An analysis of the effect sizes for each experimental manipulation revealed, on the one hand, that self-reports are reliable across several ways of collecting data and, on the other hand, self-reports are influenced by a wide array of biasing factors. Within these measurement biases, we found that participants' reports of offending are influenced by modes of administration, characteristics of the interviewer, anonymity, setting, bogus pipeline, response format, and size of the questionnaire.

Conclusions. This review provides evidence that allows us to better understand and improve crime measurements. However, many of the experiments presented in this review are not replicated and additional research is needed to test further aspects of how asking questions may impact participants' answers.

Keywords: Bias; Delinquency; Experiment; Measurement; Methodology; Modes of administration; Offending; Question design; Self-reports; Systematic review

Introduction

The measurement of crime is at the heart of criminology. Every research question which includes a measurement of offending behavior is reliant on the quality of the measurement technique. Similarly, the validity of research findings is limited by the validity of the measurement itself. Traditionally, the most widely used methods of measuring crime are official records and self-reports of offending (SRO) (for reviews, see Gomes et al., 2018a; Thornberry & Krohn, 2000). Both measurements have their strengths and weaknesses, though there is evidence that self-report measures provide better estimates of the prevalence and mean frequency of delinquent behavior (e.g., Loeber et al., 2015).

SRO were first introduced in an attempt to overcome the limitations of official records of crime (Nye & Short, 1957; Porterfield, 1943). Since then, self-reports have become the most widely used technique in criminal behavior research, becoming “one of the most important innovations in criminological research in the 20th century” (Thornberry & Krohn, 2000, p. 34). However, the great number of studies on the validity of SRO seen in the 1960s and 1970s (e.g., Clark & Tiff, 1966; Farrington, 1973; Hardt & Peterson-Hardt, 1977; Kulik et al., 1968; Schore et al., 1979) decreased after the publication of the influential book *Measuring Delinquency* (Hindelang et al., 1981), which seemed to have established the validity of SRO once and for all (Jolliffe & Farrington, 2014).

Despite the scarcity of recent studies on the validity of SRO, psychological research on self-reports of sensitive behavior has increased remarkably (for reviews, see Schwarz, 1999; Tourangeau et al., 2000). Sensitive questions are commonly defined by an invasion of privacy, which may pose a threat of disclosure, and by the need for socially undesirable answers (Tourangeau & Yan, 2007). As a result, when faced with sensitive questions, participants tend to systematically underreport behaviors that are considered socially undesirable (e.g., Krumpal, 2013).

While attempting to improve the measurement accuracy of sensitive questions, researchers have been developing experiments using different measurement techniques and comparing their behavioral estimates. Since participants are expected to underreport sensitive information, researchers usually apply the “more is better” hypothesis, assuming that the procedure that provides the highest prevalence is the most accurate method (Tourangeau & Yan, 2007).

Despite generally accepted through the sensitive question literature, this assumption could be threatened by the possibility that some individuals may overreport some forms of deviant behavior. However, literature does seem to support that overreporting is a less prevalent problem than underreporting. Studies comparing official records and SRO (mainly arrests) show medium to high agreement between the two methods (e.g., Krohn et al., 2013; Piquero et al., 2014), although indicating

a higher frequency with SRO (e.g., Auty et al., 2015; Maxfield et al., 2000). Official records' databases may be incomplete, and this may overestimate the true amount of overreporting (Daylor et al., 2019). On a slightly different note, Clark and Tiff (1966) interviewed students with and without a polygraph in order to study the validity of SRO. Findings from this study showed that participants were three times more likely to underreport deviant behavior than to overreport. Therefore, for the purposes of examining bias in self-report techniques, we focus on underreporting of offending behavior.

Several aspects of data collection have been shown to minimize response bias and to improve the quality of participants' responses to sensitive questions. For instance, evidence suggests that privacy is an important aspect of disclosure. Ong and Weiss (2000), for example, found that students' reports of cheating in school were much higher in an anonymous condition (74%) compared to a confidential condition (25%). Similar results were obtained regarding substance use by postpartum women (Beatty et al., 2014) or undergraduate students' reports of sexual behavior (Durant et al., 2002). Like anonymity, many other variables seem to affect participants' willingness to report sensitive information, for example, setting effects, e.g., school vs. home (Biglan et al., 2004); bystander effects, e.g., the presence of a parent (Moskowitz, 2004); and response format, e.g., closed vs. open-ended questions (Tourangeau & Smith, 1996).

One key variable that has been shown to affect participants' responses is mode of administration. Research on mode effects is extensive and sometimes yields conflicting results. For example, while some studies found a higher prevalence of drug, cigarette, and alcohol use in self-administered modes (e.g., surveys), compared to other-administered modes (e.g., interviews) (Gribble et al., 1998, 2000), others found no significant differences in reports of alcohol use (e.g., Sobell & Sobell, 1981) or cigarette smoking (e.g., Moskowitz, 2004). Other studies even found higher reports of alcohol use in interviews compared to self-administered modes (Cutler et al., 1988; Rehm & Spuhler, 1993). Despite the apparently conflicting results, literature reviews suggest that modes of administration affect self-reports (Richman et al., 1999) and that the benefit of self-administration increases as a function of item sensitivity (Turner & Miller, 1997) and the recency of the behavior (Tourangeau & McNealey, 2003).

Unfortunately, research on sensitive questions commonly includes questions about income, voting, sexual behaviors, and drug use (Tourangeau & Yan, 2007), and only very rarely are self-reports of offensive behavior included. Kleck and Roberts (2012), for example, reviewed experiments on mode effects of self-reports of delinquent behavior and, from a total of 27 studies, only 6 included measures of offending behavior; "most findings in this area pertain to illegal drug use, and it is possible they do not apply to other kinds of criminal behavior" (Kleck & Roberts, 2012, p. 438). Considering the

abovementioned definition of item sensitivity, surveys of offending behavior should be considered as highly sensitive; people naturally try to conceal their offenses, which often involve feelings of guilt and shame, and participants might fear potential incriminating consequences of their reports. Therefore, while sensitive questions research should more often include items about offending behavior, knowledge derived from item sensitivity research should be considered with caution by crime researchers and results should be replicated and further explored within criminological experiments.

In this article, we systematically review findings regarding potential sources of bias in collecting data on SRO. In this review, we rely only on experimental studies that compared estimates of offending from different methods of data collection, in order to gather evidence on measurement techniques, where differences are caused by the data collection method itself and the potential for confounding variables is minimized. From this systematic review of experiments, we intend to summarize the available information about the best ways of collecting SRO.

Methods

Search strategy

In order to maximize the number of experiments included in this systematic review, the literature search was developed in four steps. In a first step, we carried out a systematic search for experiments conducted until June 2018 by entering selected keywords into 30 data bases, i.e., Scopus, EBSCOhost (Anthropology Plus, Bibliography of Asian Studies, British Education Index, Business Source Ultimate, Child Development and Adolescent Studies, Criminal Justice Abstracts, eBook Collection (EBSCOhost), Education Abstracts (H.W. Wilson), Educational Administration Abstracts, ERIC, Global Health, GreenFILE, Library, Information Science and Technology Abstracts, PsycARTICLES, PsycINFO, Russian Academy of Sciences Bibliographies, Teacher Reference Center); Elsevier (ScienceDirect); Wiley InterScience; Web of Science (Web of Science Core Collection, Current Contents Connect, Derwent Innovations Index, Korean Journal Database, Medline, Russian Science Citation Index, and Scielo Citation Index); ProQuest; Ethos.

The literature search was carried out using the following keywords: (“self-report” or “self-reported” or “self-reporting” or “self-interview” or “self-interviewing” or “self-administered” or “self-administration”) and (antisocial* or delinquen* or crim* or offend* or devian* or violen* or aggressi* or arrest* or convict*) and (bias* or missing* or nonrespons* or “under-report” or “over-report” or underreport* or overreport*) and (experiment*).

Second, we searched the reference lists of all the relevant studies found in the systematic search. In a third step, taking into account the relevant studies found in the two previous procedures, we carried out a citation search using the google scholar search engine. In a last step, we contacted 6 experts in the field of self-reported offending and requested information about any experiments on measurement bias in self-reported offending, which were then included in our findings.

Inclusion criteria

In order to be included in the present systematic review, studies had to meet the eligibility criteria that are described below. This review included all published and unpublished studies, reported in English, French, Spanish, or Portuguese, that met all the three criteria.

1. The study included a measure of offending

In this review, we intended to gather relevant information specifically about the collection of self-report data on offending. Therefore, and because of its variability in the legal status across countries and between states in the USA, illegal drug use was not included. Moreover, we considered only studies which included items of offending that are typically included in delinquency research. Therefore, as an example, experiments that included exclusively bullying items (e.g., Baly & Cornell, 2011; Chan et al., 2005; Huang & Cornell, 2015) were not included in our review. However, experiments on sensitive question or risk behaviors which included typical items on delinquency questionnaires were included in this review; for example, Turner et al. (1998) studied sensitive behaviors such as sexual behavior, drug use, and violence; in this review, we considered only the offending items (i.e., threatened to hurt someone, carried a gun, in physical fight, pulled knife or gun on someone, and carried a knife or razor).

2. The study compared self-reported data on offending between different methods of data collection

This criterion allows for a between-method pairwise comparison of the prevalence and frequency of offending, thus showing which method yielded higher reports. As in the previous criteria, we were interested in gathering information about the most common methods of measurement in delinquency/criminology research. Therefore, indirect methods of measuring behavior, where it is not possible to know which individual admitted SRO making it impossible to investigate their characteristics or predictors, such as the Item Count Technique (Wolter & Laier, 2014), Unmatched Count Technique

(Dalton et al., 1994),¹ and Randomized Response Technique (Wolter & Preisendörfer, 2013)² were not included in this review.

3. The study used an experimental design

Only studies with random assignment of participants to the experimental conditions were included in this review. The criterion of random allocation of participants ensures that the findings included in this review are unlikely to be caused by confounding variables and allows for a direct comparison of the results presented in this review with the findings from the research on sensitive questions.

Search for eligible studies

As illustrated in Figure 4, our systematic search resulted in a total of 312 studies. The elimination of duplicates revealed a total of 183 different studies, of which 105 studies lacked a measure of self-reported offending, 53 lacked comparisons of SRO between different methods of data collection and, finally, 18 studies did not follow an experimental design. From these final studies, one doctoral thesis was unavailable, even after contacting the authors (Grysmen & Johnson, 2010), and could not be included in this review. The final six studies that met our three eligibility criteria were then used for the reference list search. This search resulted in 221 referenced studies, but 218 studies were excluded for failing to meet the eligibility criteria or not reporting in the included languages.

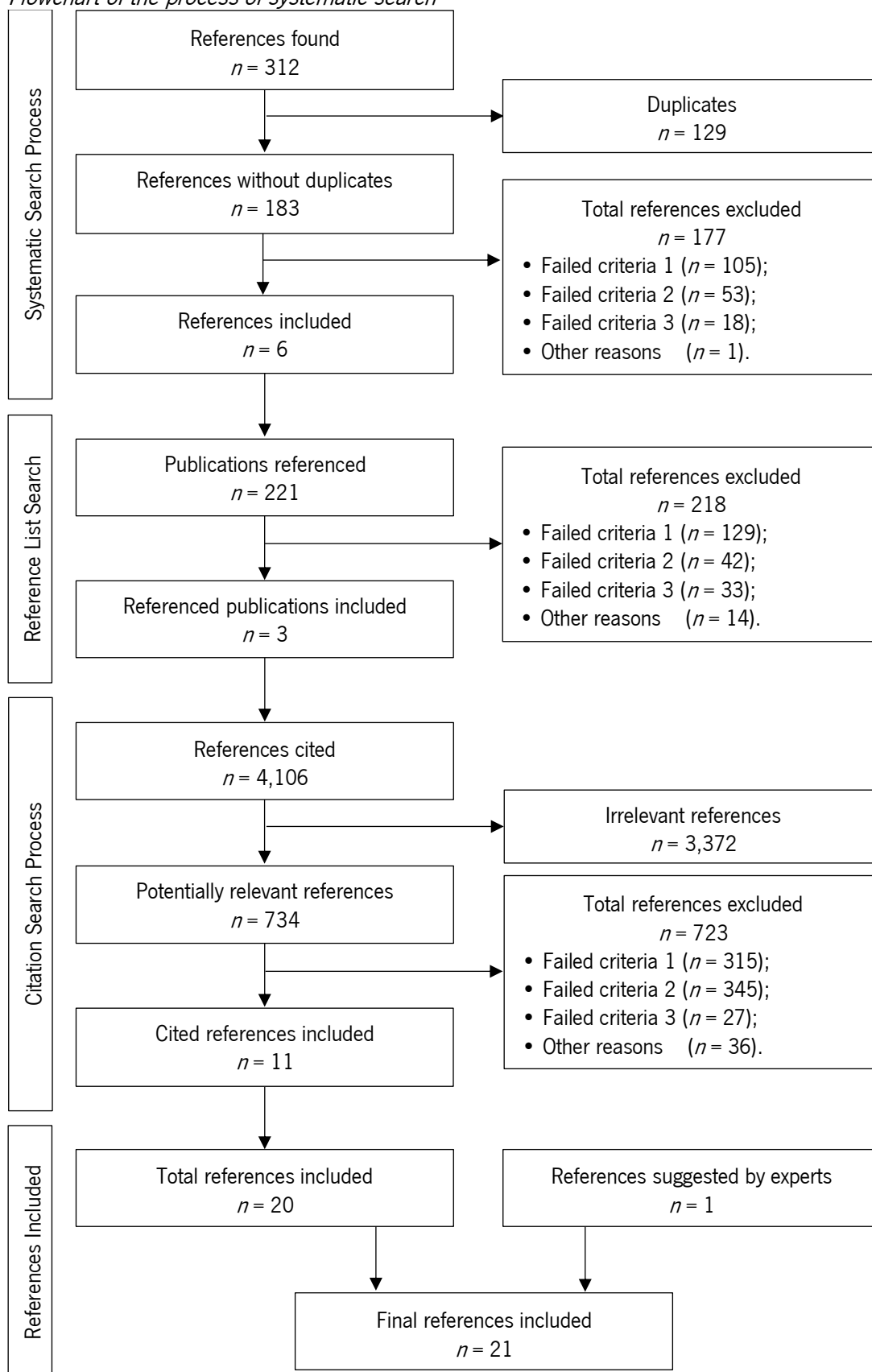
The nine relevant studies that were found in the first two steps of the search strategy were then used for a citation search, which resulted in a total of 4,106 new references. An initial title search resulted in the exclusion of 3,372 irrelevant studies and, from the 734 potentially relevant studies, 11 met the three eligibility criteria. Finally, one additional study suggested by experts in the field was included in this review. The total number of studies included in the present review comprised 21 experiments.

¹ “Item count technique” or “Unmatched count technique” are methods to reduce response bias, in which participants are randomly divided into at least two groups. The control group receives a list of questions without the sensitive item while the experimental group receives the same questions including the sensitive item. The prevalence estimate is calculated by the subtraction of the mean sum of the control group from the mean sum of the experimental group (Wolter & Laier, 2014).

² “Random response technique” is a method to reduce response bias, in which participants are presented with a pair of questions, one sensitive and one innocuous. Participants use a randomization device, such as a dice or a coin, to either give a predetermined answer (e.g., yes or no) or to answer the sensitive question truthfully. A prevalence estimation is possible from knowing the probability of the predetermined outcome (Wolter & Preisendörfer, 2013).

Figure 4

Flowchart of the process of systematic search



Analysis

As described above, the experiments included in the present systematic review focused on multiple topics of crime measurement. Experimental procedures are limited to comparisons of SRO in different measurement techniques (i.e., experimental manipulations). In order to provide comparable information regarding the magnitude of different measurements, we estimated odds ratio (OR) effect sizes for each manipulation by evaluating the difference in the odds of SRO within each measurement manipulation (e.g., interviews vs. questionnaire). A single reviewer coded the overall offending prevalence for each measurement manipulation and for each recall period, while a second reviewer double-coded 10 out of the 21 studies. This showed complete consistency between the OR calculations.

The original experiments reported findings of offending prevalence, reported mainly as percentages of offenders, with different recall periods (i.e., lifetime prevalence, past-year prevalence, and past 30-day prevalence). Along with offending prevalence reports, two studies also reported mean offending frequencies (Baier, 2017; Hindelang et al., 1981). In order to prevent non-independence issues arising from having multiple ORs from different recall periods, we have considered only offending prevalence reports and calculated combined, weighted mean effects (Borenstein et al., 2009). Effect sizes were calculated using the Comprehensive Meta-Analysis (CMA) software, version 3 (Borenstein et al., 2014).

Since the results in this study are presented as offending weighted mean prevalence, taking into consideration the “more is better” assumption, the measurement technique resulting in the highest prevalence was assumed to be the best measurement (i.e., providing the closest estimate to the true amount of offending behavior). Therefore, we estimated OR effect sizes in order to provide information regarding the odds of self-reporting offending in condition A relative to the odds of self-reporting offending in condition B. Since we are dealing with comparisons of very few studies and heterogeneity cannot be reliably estimated, the method of choice for evidence synthesis was to use the random effects model (Bender et al., 2018).

Results

Table 6 summarizes the descriptive information about the 21 studies included in this systematic review. These experiments were carried out mainly in the USA (61.9%, $k=13$), followed by several European countries, i.e., Germany, Netherlands, Switzerland, and Finland (33.3%, $k=7$), and one experiment in India. Regarding the participants, 16 studies focused on adolescents (76.2%), while 3

focused on undergraduates (14.3%) and 2 on adults (9.5%). This, in turn, reveals that the large majority of studies focused on school students (76.2%), and only two studies included sentenced participants (9.5%).

All 21 experiments studied variations in SRO caused by measurement manipulations. However, both outcome and experimental manipulation varied considerably throughout the studies. As for the measures of offending, nine studies considered measures of delinquent behavior, five experiments focused on risk behaviors, two on sensitive topics, and two on health indicators, which all included items of offending; one study looked at intimate partner violence, another studied offending frequency (i.e., lambda), and one resorted to measures of sexual aggression. As for recall periods, 12 studies included lifetime prevalence, 16 studies 12-month prevalence, and 3 studies considered 30-day prevalence of offending. Items of self-reported offensive behaviors varied from low seriousness offenses such as graffiti drawing, shoplifting, or illegal downloading, to serious and violent offenses such as vehicle theft, serious assault, or sexual aggression.

Regarding the experimental manipulations, most of the experiments included in this review studied the effect of modes of administration (66.7%, $k=14$), followed by the design of the questionnaire ($k=3$), the effect of anonymity ($k=2$), and the supervision of data collection ($k=2$). The remaining six manipulations were carried out once in each experiment; they are disclosure of information, setting of data collection, in-person follow-up, characteristics of the interviewer, reference period, and bogus pipeline (i.e., a procedure where participants are led to believe they are being monitored by a device, in order to increase honesty in self-reporting; Strang & Peterson, 2020). These 21 experiments accounted for a total of 18 different types of measurement manipulations, resulting in a total of 33 independent effect sizes, which were grouped into three categories: Modes of administration, Procedures of data collection, and Questionnaire design.

Table 6*Descriptive information on studies in the systematic review*

Study	Country	Sample	Topic	Recall Period	Outcome measure
Baier (2017)	Germany	2,643 adolescent students	Mode of administration	Lifetime prevalence Year prevalence	Delinquent behavior (Violence; Damage to property; Shoplifting; Spraying graffiti; Selling pirate copies).
Beebe et al. (1998)	USA	368 adolescent students	Mode of administration	Lifetime prevalence Year prevalence	Delinquent behavior (Involved in gang; Ran away from home; Damaged property; Beat person up; Stolen something).
Beebe et al. (2006)	USA	610 adolescent patients	Mode of administration and Disclosure	Year prevalence	Delinquent behavior (Beat someone up; Carried weapon).
Brener et al. (2006)	USA	4,506 adolescent students	Mode of administration and Setting	Year prevalence 30-day prevalence	Delinquent risk behavior (Drove after drinking alcohol; Carried a weapon; Carried a gun; Physical fight; Dating violence).
Eaton et al. (2010)	USA	5,227 adolescent students	Mode of administration	Year prevalence 30-day prevalence	Delinquent risk behavior (Drove after drinking alcohol; Carried a weapon; Carried a gun; Carried a weapon on school property; Physical fight; Physical fight on school property; Dating violence).
Enzmann (2013)	Germany	1,629 adolescent students	Questionnaire design	Lifetime prevalence Year prevalence	Delinquent behavior (Vandalism; Shoplifting; Burglary; Bicycle theft; Car theft; Car break; Snatching; Carrying weapons; Extortion; Group fight; Assault; Drug dealing).
Hamby et al. (2006)	USA	160 undergraduate students	Mode of administration and Questionnaire design	Year prevalence	Partner violence Perpetration (Psychological aggression; Physical assault; Sexual coercion; Injury).

Hindelang et al. (1981)	USA	13,842 adolescents	Mode of administration and Anonymity	Lifetime prevalence Year prevalence	Delinquent behavior (69 items grouped into Contacts with the criminal justice system; Serious crimes; General delinquency; Drug offenses; and School and family).
Horney & Marshall (1992)	USA	700 convicted male offenders	Questionnaire design	Year prevalence	Magnitude of self-reported offending frequency, i.e. Lambda (Burglary, robbery, theft, auto-theft, forgery, fraud, assault, and drug deals).
King et al. (2012)	USA	245 adolescents patients	In-person follow-up	Year prevalence	Risk factors for suicidal behavior (including aggressive/delinquent behavior)
Kivivuori et al. (2013)	Finland	924 adolescent students	Supervision	Lifetime prevalence Year prevalence	Delinquent behavior (graffiti drawing, vandalism at school, vandalism elsewhere, shoplifting, stealing at school, motor vehicle theft, other theft, breaking and entering, fighting, beating up someone, robbery, drunken driving, illegal downloading)
Knapp and Kirk (2003)	USA	352 undergraduate students	Mode of administration	Lifetime prevalence	Sensitive questions (Have you ever written on a restroom wall?, Have you ever used someone else's credit card (number) without their permission?, and Have you ever been in jail?)
Krohn et al. (1974)	USA	321 undergraduate students	Mode of administration and Interviewer	Year prevalence	Delinquent behavior (Drunken driving, Fighting, Petty theft, Grand larceny, Property damage, and Illegal entry)
Lucia et al. (2007)	Switzerland	1,203 adolescent students	Mode of administration and Reference period	Lifetime prevalence Year prevalence	Delinquent behavior (Driving without license, Shoplifting [more than €35], Shoplifting [less than €35], Breaking into a car, Harassing somebody in the street, Theft at school, Theft at home, Fare dodging, Vehicle theft, Theft of an object from a vehicle, Assault, Threats with gun/knife, Racket [extortion], Robbery, Arson, Selling soft drugs, Selling hard drugs, Graffiti, Vandalism, Theft from the person)
Potdar and Koenig (2005)	India	900 male undergraduate students and 600	Mode of administration	Lifetime prevalence	Risk behavior (Carrying a weapon/gun and Engaged in abusive, violent behavior after drinking)

		male residents in slums.			
Strang and Peterson (2020)	USA	93 young, community men	Bogus Pipeline	Lifetime prevalence	Sexual aggression (Verbal coercion tactics, Drugs and alcohol tactics, and Force tactics of sexual assault)
Trapl et al. (2013)	USA	275 adolescent students	Mode of administration	Lifetime prevalence	Sensitive behaviors (Shoplifting)
Turner et al. (1998)	USA	1,672 adolescent males	Mode of administration	Year prevalence 30-day prevalence	Risk behavior (Threatened to hurt someone; Carried a gun; In physical fight; Pulled knife or gun on someone; and Carried a knife or razor)
van de Looij-Jansen et al. (2006)	Netherlands	704 adolescent students	Anonymity	Lifetime prevalence	Health indicators (Aggressive behavior; Vandalism and stealing; Violent delinquent behavior; and Carrying a weapon)
van de Looij-Jansen and de Wilde (2008)	Netherlands	532 adolescent students	Mode of administration	Year prevalence	Health indicators (Aggressive behavior; Vandalism and stealing; and Carrying a weapon)
Walser and Killias (2012)	Switzerland	1,197 adolescent students	Supervision	Lifetime prevalence Year prevalence	Delinquent behavior (Assault; Group fight; Robbery; Sexual assault; Burglary; Shoplifting; Bicycle theft; Other theft; Vandalism; Carrying a weapon; Drug dealing; and Any delinquency)

Table 7 summarizes the results of these experiments, organized by measurement manipulations. For each manipulation, we provided information regarding the OR effect sizes (i.e., OR, 95% confidence intervals, Z statistics, and *p* value). Because most comparisons are made with few cases, additionally to ORs, we also reported the number of statistically significant differences found in individual item comparisons (when available). Since this review includes results from several different manipulations, Table 7 provides information on the experimental manipulation under analysis (experimental condition A vs. experimental condition B). Considering the calculation of OR effect sizes to be the odds of reporting offending behavior in condition A divided by the odds of reporting offending in condition B, an OR > 1 indicates higher reports in condition A, while an OR < 1 indicates higher reports in condition B, and an OR = 1 indicates a null effect. For example, in the first line of Table 7, we present the comparison of personal interview (i.e., condition A) vs. self-administered questionnaire (i.e., condition B) (Krohn et al., 1974); an OR = 0.70 indicates that the odds of reporting deviant behavior in the interview (i.e., condition A) were decreased by 30% relative to the questionnaire (i.e., condition B).

Table 7

Main findings of experiments in the systematic review

Study	Comparison (<i>p</i> < .05)	OR	95% CI	<i>z</i>	<i>p</i>
Modes of administration					
Personal Interview (PI) vs. Self-Administered Questionnaire (SAQ) (k = 3)					
Krohn et al. (1974)	0 of 6	0.70	[0.34, 1.45]	-0.96	.336
Hindelang et al. (1981)	-	0.97	[0.92, 1.04]	-0.90	.398
Potdar and Koenig (2005)	0 of 2	0.83	[0.36, 1.91]	-0.45	.656
Random model		0.97	[0.92, 1.03]	-0.95	.341
Personal Interview (PI) vs. Audio Computer-Assisted Self-Interview (ACASI) (k = 1)					
Potdar and Koenig (2005)	> PI (1 of 4)	1.23	[0.84, 1.80]	1.05	.293
Self-Administered Questionnaire (SAQ) vs. Computer-Assisted Self-Interview (CASI) (k = 10)					
Beebe et al. (1998)	>SAQ (2 of 5)	1.42	[0.91, 2.20]	1.55	.122
Knapp and Kirk (2003)	0 of 3	1.11	[0.59, 2.09]	0.33	.742
Beebe et al. (2006)	0 of 2	1.06	[0.65, 1.71]	0.22	.823
Brener et al. (2006)	>CASI (2 of 5)	0.84	[0.70, 0.99]	-2.06	.040
Hamby et al. (2006)	> CASI (1 of 4) > SAQ (1 of 4)	0.93	[0.49, 1.77]	-0.21	.835
Lucia et al. (2007)	> CASI (2 of 40) > SAQ (5 of 40)	1.12	[0.80, 1.56]	0.66	.507

van de Looij-Jansen and de Wilde (2008)	> CASI (1 of 3)	0.81	[0.59, 1.10]	-1.36	.174
Eaton et al. (2010)	> CASI (5 of 7)	0.90	[0.77, 1.04]	-1.42	.157
Trapl et al. (2013)	0 of 1	1.10	[0.58, 2.10]	0.30	.764
Baier (2017)	> SAQ (1 of 10)	0.94	[0.71, 1.24]	-0.47	.635
Random Model		0.92	[0.84, 1.01]	-1.85	.064
Self-Administered Questionnaire (SAQ) vs. Audio Computer-Assisted Self-Interview (ACASI) (k = 3)					
Turner et al. (1998)	> ACASI (4 of 5)	0.69	[0.51, 0.93]	-2.42	.015
Potdar and Koenig (2005)	-	1.05	[0.47, 2.36]	0.12	.902
Trapl et al. (2013)	0 of 1	1.14	[0.60, 2.19]	0.41	.685
Random Model		0.82	[0.59, 1.14]	-1.20	.232
Computer-Assisted Self-Interview (CASI) vs. Audio Computer-Assisted Self-Interview (ACASI) (k = 1)					
Trapl et al. (2013)	0 of 1	1.04	[0.54, 1.99]	0.11	.914
Self-Administered Questionnaire (SAQ) vs. Automated Touch-Tone Telephone (TACASI) (k = 1)					
Knapp and Kirk (2003)	0 of 3	1.18	[0.72, 1.93]	0.66	.510
Computer-Assisted Self-Interview (CASI) vs. Telephone Audio Computer-Assisted Self-Interview (TACASI) (k = 1)					
Knapp and Kirk (2003)	0 of 3	1.06	[0.54, 2.07]	0.17	.863
Procedures of Data Collection					
Supervision by teachers vs. Supervision by researchers (k = 2)					
Walser and Killias (2012)	> teacher (2 of 22)	1.04	[0.83, 1.31]	0.35	.726
Kivivuori et al. (2013)	> research (2 of 26)	0.87	[0.58, 1.31]	-0.66	.508
Random Model		1.00	[0.82, 1.22]	-0.02	.981
Non-anonymous vs. Anonymous (k = 2)					
Hindelang et al. (1981)	-	0.98	[0.92, 1.04]	-0.71	.481
van de Looij-Jansen et al. (2006)	> Anonym. (3 of 4)	0.67	[0.51, 0.88]	-2.89	.004
Random Model		0.83	[0.58, 1.20]	1.00	.319
No-Disclosure vs. Disclosure (k = 1)					
Beebe et al. (2006)	> No Discl. (1 of 2)	1.69	[0.99, 2.88]	1.93	.053
Home setting vs. School setting (k = 1)					
Brener et al. (2006)	> school (5 of 5)	0.75	[0.63, 0.89]	-3.27	.001
'Conservative' interviewer vs. 'Hip' interviewer (k = 1)					
Krohn et al. (1974)	> 'Hip' (2 of 6)	0.54	[0.27, 1.08]	-1.75	.080
No in-person follow-up vs. In-person follow-up (k = 1)					
King et al. (2012)	0 of 1	0.62	[0.36, 1.06]	-1.76	.079
Bogus pipeline (BPL) vs. Control group (k = 1)					

Strang and Peterson (2020)	> BPL (2 of 8)	2.18	[0.82, 5.81]	1.55	.121
Questionnaire design					
Response Format: 2-options vs. 7-options (k = 1)					
Hamby et al. (2006)	> 7-options (2 of 4)	1.19	[0.63, 2.25]	0.52	.602
Long vs. Short questionnaire (k = 1)					
Enzmann (2013)	> Short (5 of 24)	0.89	[0.74, 1.06]	-1.31	.192
Standard vs. Month-by-month reporting (k = 1)					
Horney and Marshall (1992)	0 of 8	0.98	[0.69, 1.39]	-0.13	.900
Reference Period: "12 months" vs. "Since October 2003" (k = 1)					
Lucia et al. (2007)	0 of 20	1.04	[0.62, 1.74]	0.15	.878

Note. The "Comparison ($p < .05$)" column shows the number of statistically significant differences found in individual item comparisons (when available). > = higher estimates, e.g. "> PI (1 of 2)" = 1 of 2 item comparisons presented significantly higher estimates of self-reported offending in the Personal Interview.

Modes of administration

In the first category, we included all the experimental manipulations regarding the methods through which participants provide their answers to the offending questions. In this review, experiments considered the following: (a) personal interviews (PI), where questions are delivered in face-to-face interviews and answers are provided orally to an interviewer; (b) self-administered questionnaires (SAQ), where participants are given a paper-and-pencil questionnaire which they complete on their own; (c) computer-assisted self-interviews (CASI), where participants are given a questionnaire on a computer screen which they complete on their own directly onto a computer; (d) audio computer-assisted self-interview (ACASI), where questionnaires are presented on a computer screen and participants can listen to audio records of the questions and provide their answers directly onto the computer; and (e) telephone audio computer-assisted self-interview (TACASI), where participants are contacted via telephone, listen to audio records of the questions, and provide their answers on the telephone which are recorded via automated software.

PI vs. SAQ

Three studies compared results of SRO collected under PI and SAQ (Hindelang et al., 1981; Krohn et al., 1974; Potdar & Koenig, 2005). The pooled effect sizes presented virtually null ORs, slightly in favor of SAQ but with no statistical significance. The overall analysis under a random model suggested

no significant differences between data collected with these two methods (OR = 0.97, 95% CI [0.92, 1.03], $z = -0.95$, $p = .341$).

PI vs. ACASI

Only one study compared PI and ACASI (Potdar & Koenig, 2005). Results mainly favored PI (OR = 1.23, 95% CI [0.84, 1.80], $z = 1.05$, $p = .293$), though it did not reach statistical significance ($p > .05$).

SAQ vs. CASI

The analysis of SAQ vs. CASI was the most replicated comparison in the present review, with 10 studies (Baier, 2017; Beebe et al., 1998, 2006; Brener et al., 2006; Eaton et al., 2010; Hamby et al., 2006; Knapp & Kirk, 2003; Lucia et al., 2007; Trapl et al., 2013; van de Looij-Jansen & de Wilde, 2008). An analysis of the individual effect sizes showed that 5 comparisons favored CASI, though only one reached statistical significance with an OR of 0.84 (Brener et al., 2006), while of the 5 comparisons favoring SAQ none reached statistical significance. On average, the mean effect slightly favored CASI over SAQ (OR = 0.92, 95% CI [0.84, 1.01], $z = -1.85$, $p = .064$), though with only marginal significance ($p < .10$).

SAQ vs. ACASI

Three studies provided comparisons of offending behavior collected with SAQ or ACASI (Potdar & Koenig, 2005; Trapl et al., 2013; Turner et al., 1998). One out of the three ORs presented statistically significant results in favor of the ACASI mode (OR = 0.69, $p = .015$). Considering random effects, the average effect size showed an OR = 0.82 favoring ACASI but with no statistical significance (OR = .82, 95% CI [0.59, .136], $z = -1.20$, $p = .232$).

CASI vs. ACASI

Trapl et al. (2013) conducted the sole experiment comparing SRO obtained through CASI and ACASI. Despite participants reporting slightly higher estimates of lifetime shoplifting under the CASI mode of data collection, results of this experiment showed a nonsignificant OR effect size (OR = 1.04, 95% CI [0.54, 1.99], $z = 0.11$, $p = .914$).

SAQ vs. TACASI

Knapp and Kirk (2003) carried out the unique experiment that compared SAQ and TACASI. Results showed slightly higher estimates of offending in the SAQ mode of administration, though with no statistical significance (OR = 1.18, 95% CI [0.72, 1.93], $z = 0.66$, $p = .510$).

CASI vs. TACASI

Similar to the previous results, the experimental comparison between CASI and TACASI (Knapp & Kirk, 2003) showed a nonsignificant effect size (OR = 1.06, 95% CI [0.54, 2.07], $z = 0.173$, $p = .863$).

Procedures of data collection

The second category of manipulations takes into account different procedures applied in the data collection that might influence the participants' SRO. This category accounts for seven out of the total 18 manipulations, which included manipulations in Supervision of data collection ($k = 2$), Anonymity ($k = 2$), Characteristics of the Interviewer ($k = 1$), Setting of data collection ($k = 1$), Disclosure of information ($k = 1$), In-person follow-up ($k = 1$), and Bogus pipeline ($k = 1$).

Supervision

Two studies compared supervision by the participants' teacher with supervision by the researchers during the completion of the questionnaire with CASI methodology (Kivivuori et al., 2013; Walser & Killias, 2012). In general, results showed slightly higher estimates in the condition where participants were supervised by researchers, though not reaching statistical significance. On average, random effects showed no statistically significant differences between the two methods (OR = 1.00, 95% CI [0.82, 1.22], $z = -0.02$, $p = .981$).

Anonymity

From the pooled experiments, two studies focused on the issue of anonymity in SRO. Hindelang et al. (1981) used both anonymous/non-anonymous questionnaires and anonymous/non-anonymous interviews (where contact between interviewer and interviewee was prevented by a screen). Results showed no statistically significant differences, with an OR of 0.98 ($p = .481$). In the experiment of van de Looij-Jansen et al. (2006), participants received questionnaires with their names on them (i.e., confidential group) vs. questionnaires with no identifying information (i.e., anonymous condition). In this case, results showed higher SRO in the anonymous condition (OR = 0.67, $p = .004$). The average effect

size favored anonymous procedures, showing a reduced odds by 17% of reporting offending behavior in the non-anonymous condition, though with no statistically significant effects (OR = 0.83, 95% CI [0.58, 1.20], $z = -1.00$, $p = .319$).

Disclosure

Beebe et al. (2006) conducted an experiment studying the effect of disclosure of self-reported information. This experiment compared results of two groups. In one group, participants were told that their responses would only be seen by the researchers and in a second group, participants were told that a summary report would be given to their health care provider. Findings showed an increased odds by 69% of reporting offending behavior in the no-disclosure condition, though statistical significance reached only a marginal level (OR = 1.69, 95% CI [0.99, 2.88], $z = 1.93$, $p = .053$).

Setting

Brener et al. (2006) developed an experiment to test differences between data collection at home vs. data collection at school. Results considerably favored data collection at schools, with a reduced odds of reporting offending behavior by 25% in a home setting (OR = 0.75, 95% CI [0.63, 0.89], $z = -3.27$, $p = .001$).

Characteristics of the interviewer

Krohn et al. (1974) carried out an experiment to test the hypothesis that the characteristics of the interviewer might influence the reports of offending. The two experimental conditions included interviewers with a conservative appearance, dressed formally and closely trimmed hair (i.e., “conservative” interviewers) vs. a group of interviewers casually dressed and with long hair (i.e., “hip interviewers”). Findings showed that the odds of reporting delinquent behavior decreased by 46% with the “conservative” interviewer, though the statistical test revealed to be only marginally significant, i.e., $p < .10$ (OR = 0.54, 95% CI [0.27, 1.08], $z = -1.75$, $p = .080$).

In-person follow-up

King et al. (2012) conducted the unique experiment comparing self-reports of aggressive/delinquent behavior of adolescent patients seeking medical emergency services who were randomly allocated to two groups. The control group had no in-person follow-up, but in the experimental group, participants were told about a subsequent session of in-person follow-up where they would receive

feedback on their answers. Results from this experiment showed a decreased odds of self-reports by approximately 38% in the control group (i.e., no in-person follow-up), and once again, z statistics showed only marginally significance at a level of $p < .10$ (OR = 0.62, 95% CI [0.36, 1.06], $z = -1.76$, $p = .079$).

Bogus pipeline

Finally, Strang and Peterson (2020) carried out an experiment to test the effects of a bogus pipeline in reporting sexual aggressive behavior. In the control group, participants were attached to a physiological measurement device and were told that it was to “determine the level of anxiety prior to starting the questionnaire.” In the bogus pipeline group, participants were attached to the same physiological measurement device and were told it was “similar to a polygraph or lie detector test” and “that the machine was being attached to encourage honest responding.” Overall, despite non-significant results from z statistics, findings showed an increased odds ratio of 2.18 of reporting sexual aggression (including verbal coercion, use of drugs and alcohol tactics, and force) in the bogus pipeline condition (OR = 2.18, 95% CI [0.82, 5.81], $z = 1.55$, $p = .121$). Moreover, individual item comparisons revealed that men in the bogus pipeline condition showed 6.5 times greater odds of reporting illegal sexual assault (OR = 6.49, 95% CI [1.78, 23.69], $z = 2.83$, $p < .01$).

Questionnaire design

In the third category, we grouped the experimental manipulations of the design of the questionnaire itself. This category accounts for four out of the total 18 manipulations, which included manipulations in response format ($k = 1$), response format and follow-up questions ($k = 1$), Month-by-month reporting ($k = 1$), and reference periods ($k = 1$).

Response format

One study focused on the response format (Hamby et al., 2006). In this experiment, self-reports of partner violence perpetration were given in two different formats: (a) a dichotomous response format (i.e., yes and no) and (b) a 7-category response format (i.e., once, twice, 3 to 5 times, 6 to 10 times, 11 to 20 times, more than 20 times, and never). The average effect size showed nonsignificant effects of the response manipulation (OR = 1.19, 95% CI [0.63, 2.25], $z = 0.52$, $p = .602$). However, results varied considerably according to the types of crimes. Self-reports of psychological aggression (OR = 0.80, 95% CI [0.31, 2.04]) and physical assault (OR = 0.77, 95% CI [0.41, 1.46]) were slightly higher in the

dichotomous condition, but not statistically significant ($p > .05$). For self-reports of sexual coercion (OR = 3.58, 95% CI [1.34, 9.58]) and injury (OR = 3.35, 95% CI [1.03, 10.89]), results were significantly higher in the 7-option response condition ($p < .05$).

Response format and follow-up questions

Enzmann (2013) developed a cross-sectional experiment testing a shorter version of the ISRD-2 questionnaire. The two experimental conditions were as follows: (a) a standard ISRD-2 questionnaire (i.e., long version), with five follow-up questions for each offending item, and a no-yes response pattern; (b) a short version of the ISRD-2 questionnaire, with only one follow-up question, and a yes-no response pattern. The effect size showed a slight decrease in chances of reporting delinquent activity in the long version by 11%, though without statistical significance (OR = 0.89, 95% CI [0.74, 1.06], $z = -1.31$, $p = .192$). However, individual item comparison showed statistically significant higher reports in the short version in 5 out of 24 comparisons.

Standard vs. Month-by-month reporting

Horney and Marshall (1992) carried out an experiment comparing standard interviewing methods in the RAND Second Inmate Survey (Chaiken & Chaiken, 1982) and a Month-by-month reporting interview to measure Lambda (i.e., individual offending frequency). Results showed little difference between the two methods (OR = 0.98, 95% CI [0.69, 1.39], $z = -0.13$, $p = .900$).

Reference period

Finally, Lucia et al. (2007) conducted the only experiment found in the present systematic review that attempted to study the potential effects of different instructions regarding the recall period. In this experiment, authors manipulated the instructions about the reference period: (a) "During the last 12 months," and (b) "Since the school vacation of October 2003" (which corresponded to a 12-month period). The results showed similar estimates of delinquent behavior in both conditions (OR = 1.04, 95% CI [0.62, 1.74], $z = 0.15$, $p = .878$).

Discussion

Despite the wide use of the self-report methods in criminology, many researchers have shared their concerns about the quality of this methodology and how several contextual features may impact

participants' SRO. However, much of the research in the field of criminological methodology has been focused on the comparison between offending data collected through self-reports and official records (Gomes et al., 2018a), which tells us little about how to improve the methods of obtaining offending information. In this review, we carried out a systematic search for experiments testing potential sources of bias in collecting SRO in order to summarize the available information about measurement bias in criminology, providing evidence to improve data collection of SRO.

We found that, contrary to other fields of sensitive questions (e.g., Richman et al., 1999; Tourangeau & Yan, 2007), experimental research on SRO is very scarce. The total 21 pooled experiments aimed to study 18 different potential measurement biases, which in turn resulted in many one-study experimental manipulations. However, the summarized available information in this review provides relevant information regarding the best practices of data collection, the stability of data throughout different methods, and points to directions for future research. Present findings were grouped into three categories (i.e., modes of administration, procedures of data collection, and questionnaire design) and are discussed below.

Modes of administration

Considering the first category, experiments included in this systematic review compared seven different pairs of administration methods. Evidence revealed general similarity in the results collected through the multiple modes of administration. The evidence suggests that, for the study of SRO, personal interviews, paper-and-pencil or computer questionnaires, with or without audio, in person or by the telephone, provide similar results. However, these results should be interpreted very carefully; evidence is based on only few studies and, in some cases, carried out several decades ago, showing that more research is clearly needed.

One clear example of this is the findings referring to the personal interviews, which were inconsistent with the sensitive questions literature that has shown that, because people are required to report sensitive information face-to-face to a third person, PI is usually seen as a weaker measurement mode, which tends to decrease the odds of reporting sensitive behavior (e.g., Gribble et al., 1998, 2000). Out of the three experiments considering PI, two studies were developed more than 30 years ago (Krohn et al., 1974; Hindelang et al., 1981). Since then, much has changed in regard to the use of self-report questionnaires, computers, among many other aspects; and the relationship between individuals and face-to-face interviews may have changed. On the other hand, the most recent experiment considering SRO collected with PI was carried out in India (Potdar & Koenig, 2005), which may add a confounding

cultural aspect that we are not aware of. Furthermore, recent projects seem to provide contradicting results. In a recent presentation, Gomes et al. (2018b) presented preliminary results of an experimental study which compared PI to SAQ and CASI, where self-administered modes resulted in higher scores of SRO.

Similarly, our results on the effect of audio of modes of administration seem to contradict the general findings in the self-report literature. While findings from research on sensitive questions and substance use generally report the benefits of audio, both in overcoming illiteracy and eliciting higher reports (e.g., Thornberry & Krohn, 2000; Tourangeau & Smith, 1996), present findings on SRO comparing reports collected under both SAQ and CASI to ACASI found no evidence of benefits from audio. However, one of the three studies comparing SAQ and ACASI reported an overall decreased chances of reporting offending behavior by about 31% in the SAQ condition, providing evidence for the significant advantages of audio (Turner et al., 1998). Therefore, the results are not clear about the impact of audio in SRO and more research is needed.

Regarding the comparison between SAQ and CASI, a total of 10 experiments reported results which suggested overall no statistically significant different results between the two methods. However, the overall OR slightly favored CASI, showing an 8% reduced odds of reporting offending under the SAQ condition, with marginal significance ($p = .064$). An individual experiment overview suggests a considerable variability in the results. Out of the total 10 experiments, five comparisons slightly favored SAQ. Although no individual OR favoring SAQ reached statistical significance, individual item comparisons reported 7 out of 51 (14%) statistically significant higher reports under SAQ. On the other hand, from the five experimental comparisons favoring CASI, one reached a statistically significant OR of 0.84 (Brener et al., 2006), and individual item comparisons presented a total of 10 out of 29 (34%) individual item comparisons significantly favoring the CASI mode. Therefore, despite the results showing overall no significant difference between these two methods, there seems to be some evidence favoring computer-assisted methods over paper-and-pencil. Future research should carry out more research in this subject matter and, on the other hand, further analyze these results trying to better understand the impact of modes of administration on SRO, for example, in order to explore for potential moderators, such as recall periods or types of offenses.

Procedures of data collection

Taking into consideration the second category, a total of nine experiments provided evidence regarding potential biases derived from the procedures applied in the data collection. Despite the limited

number of experiments, we collected data regarding seven pairwise comparisons of types of procedures. Results demonstrated that completion of a questionnaire at school environments seems preferable to completion at home, though only a single study focused on this matter (Brener et al., 2006). In this study, all five individual items of offending presented statistically significant higher reports in the school condition, and the overall OR showed a 25% decreased likelihood of self-reports in home settings. This result is consistent with the self-report literature and has been reported in previous quasi-experimental studies (e.g., Cops et al., 2016). On the contrary, we found no evidence that supervision by teachers or supervision by research staff impacts youth SRO. In the same way, the experiment looking at the effect of bogus pipeline (i.e., where participants are attached to a physiological measurement device that they believe detects lies) showed non-significant results, though the odds of reporting offending behavior in the bogus pipeline condition increased by 118%. On the other hand, despite the overall OR for the two experiments focusing on the effect of anonymity presented non-significant results, one experiment presented evidence favoring SRO in anonymous conditions, showing a reduced odds of 37% of reporting offending behavior in the only confidential condition (van de Looij-Jansen et al., 2006).

The three remaining experiments on the topic of procedures of data collection showed ORs with marginal statistical significance ($p < .1$). In the case of disclosure to third parties, some evidence was collected favoring collection procedures when reported information is not disclosed to third parties, though OR was slightly over the statistical significance threshold ($p = .053$; Beebe et al., 2006). In the same way, the interviewer characteristics seemed to have a marginally significant impact on participants' reports of offending behavior, showing a 46% reduced chances of reports collected by a formally dressed interviewer, when compared to a casually dressed interviewer. Finally, there was marginally significant evidence of decreased odds by 32% of reporting offending behavior in procedures where adolescent patients seeking medical emergency services were not screened after the completion of the questionnaire.

Questionnaire design

In the last category, the four pooled experiments provided information regarding four different manipulations of questionnaire design. As a summary of results, we found no evidence that SRO vary as a function of Standard vs. Month-by-month reporting and Reference period. In other words, we found no evidence that SRO are subject to telescoping, i.e., a memory distortion in which participants report events that occurred prior to the recall period (e.g., Loftus & Marburger, 1983). Furthermore, the experimental comparison on response format provided a non-significant effect size. However, 2 out of 4 individual offending items showed significantly higher scores in the 7-option response format, rather than

dichotomous response options. Response format of self-report offending questionnaire clearly needs more research in the future.

Finally, despite the overall OR not reaching statistical significance, the analysis of item comparisons in the experiment developed by Enzmann (2013) showed significantly higher reports of offending in 5 out of 24 items in the short version. Findings suggesting a slight increased odds of reporting offending behavior with a short questionnaire with fewer follow-up questions and a yes-no response pattern are consistent with the self-report literature, i.e., longer questionnaires may cause more fatigue, driving participants to answer negatively (Krosnick & Presser, 2010). In the same way, follow-up questions may discourage participants from answering positively to items, in order to answer fewer questions, and a yes-no response pattern may increase positive answers because of response order effects (for a discussion, see Enzmann, 2013). However, these results need to be carefully considered because the effects of questionnaire size (short vs. long), number follow-up questions (1 question vs. 5 questions), and response order (yes-no vs. no-yes) are confounded in the two experimental manipulations, and future experiments should try to disentangle these effects and explore the potential isolated effect of each of these variables.

Limitations

In this article, we reviewed the available experimental evidence regarding measurement bias in SRO, in order to provide evidence to improve data collection in the study of offending behavior. In doing this, we have conceptualized offending behavior as a broad concept which includes several types of criminal and offending behavior, which varied from sexual aggression (Strang & Peterson, 2020), to delinquent behavior (e.g., Walser & Killias, 2012), ever being in jail (Knapp & Kirk, 2003), etc. We have also considered results referring to different recall periods, such as lifetime prevalence (e.g., Knapp & Kirk, 2003), past-year prevalence (e.g., Beebe et al., 2006), and past 30-day prevalence (Turner et al., 1998). This variability, both in types of offenses and recall periods, results in an unstandardized dependent variable which may introduce bias in our results. Future experiments on SRO should focus on standardized measures of offending in order to produce comparable results.

One of the primary conclusions of our systematic review is the need for more experimental research on the topic of measurement bias in criminological studies. The total number of studies pooled in this review was 21 experiments, which is very small, especially compared to the research developed in other areas of self-report methodology (e.g., sensitive questions). Furthermore, the total 21 experiments focused on 18 different types of measurement manipulations, resulting in many measurement

manipulations based on one and two studies, which does not allow for solid conclusions. Finally, systematic reviews can be subject to publication bias, which may impact its representativeness by overestimating studies easily available, such as those reporting statistically significant results (Wilson, 2009). In the present review, because of the small number of effect sizes contributing to each measurement manipulation that was tested, we did not carry out a publication bias analysis. However, we did include evidence from unpublished studies in order to provide the widest average possible of the available evidence in the literature.

General conclusion

Findings from this review are twofold. On the one hand, most experimental comparisons included in this article showed no statistically significant differences in the prevalence of SRO behavior. This result suggests that the self-reported offending methodology generally yields consistent and stable results throughout multiple modes of administration, procedures of data collection, and questionnaire designs. On the other hand, we collected experimental evidence suggesting that SRO are, at least to some extent, subject to measurement bias resulting from mode effects, procedure effects, and design effects. Since criminological knowledge is so widely dependent on data collected through the self-report methodology, understanding that participants' self-reports may vary as a function of such a wide array of factors may call into question the validity and reliability of research conclusions. However, it is not reasonable to simply conclude that there is a lack of reliability of self-report methods, nor is it "sufficient to attach warning labels to reports of self-reported delinquency, pointing to the possibility that differences in methods may result in different estimates of the amount of crime" (Enzmann, 2013, p. 149). As in any other scientific field, criminology researchers should focus their efforts on understanding the biasing factors and to what extent they impact participants' self-reports, and thus improve the quality of crime measurements.

Despite the evident need for more replication and experimental studies in this field, we have tried to compile what we consider to be the key takeaway points from this systematic review. Taking into consideration the data analyzed in this study, we found no evidence for mode effects. Therefore, studies using different modes of administration to collect offending data seem to provide similar results and are generally comparable. However, there seemed to be some evidence supporting the benefits of audio presentation. Also, we found no evidence that Supervision by teachers vs. Supervision by researchers impacts participants' reports. Therefore, researchers who are interested in studying offending using self-

report measures should consider to include audio presentation in their projects and prioritize anonymous data collections in school environments, rather than at participants' homes.

Finally, this review included experiments testing biasing effects from very different aspects. However, most of these experimental tests occurred in single experiments, and further replication is surely needed. Furthermore, the experiments included in this review covered only a limited number of the aspects that concern self-report researchers, and many other research questions remain unanswered. For example, does the sex of the interviewer/supervisor influence the participants' reports of offending? Is the prevalence of offending under variety scales different from those under questionnaires with follow-up questions? Does the size of the questionnaire (number of questions) impact SRO? Does it matter to have the offending items at the beginning or at the end of the questionnaire? Does paying the participants have an impact on SRO? Does having different recall periods matter? And how do these biasing effects interact with each other? Are different participants differently affected by these factors? Further research is crucial and randomized experiments are very important in answering these questions and in determining the reliability of methods of collecting SRO.

References

- Auty, K. M., Farrington, D. P., & Coid, J. W. (2015). The validity of self-reported convictions in a community sample: Findings from the Cambridge Study in Delinquent Development. *European Journal of Criminology, 12*(5), 562-580. <https://doi.org/10.1177/1477370815578198>
- Baier, D. (2017). Computer-assisted versus paper-and-pencil self-report delinquency surveys: results of an experimental study. *European Journal of Criminology, 15*(4), 385-402. <https://doi.org/10.1177/1477370817743482>
- Baly, M. W., & Cornell, D. G. (2011). Effects of an educational video on the measurement of bullying by self-report. *Journal of School Violence, 10*(3), 221-238. <https://doi.org/10.1080/15388220.2011.578275>
- Beatty, J. R., Chase, S. K., & Ondersma, S. J. (2014). A randomized study of the effect of anonymity, quasi-anonymity, and certificates of confidentiality on postpartum women's disclosure of sensitive information. *Drug and Alcohol Dependence, 134*, 280-284. <https://doi.org/10.1016/j.drugalcdep.2013.10.016>
- Beebe, T. J., Harrison, P. A., Mcrae, J. A., Anderson, R. E., & Fulkerson, J. A. (1998). An evaluation of computer-assisted self-interviews in a school setting. *Public Opinion Quarterly, 62*(4), 623-632. <https://doi.org/10.1086/297863>
- Beebe, T. J., Harrison, P. A., Park, E., McRae, J. A., Jr., & Evans, J. (2006). The effects of data collection mode and disclosure on adolescent reporting of health behavior. *Social Science Computer Review, 24*(4), 476-488. <https://doi.org/10.1177/0894439306288690>
- Bender, R., Friede, T., Koch, A., Kuss, O., Schlattmann, P., Schwarzer, G., & Skipka, G. (2018). Methods for evidence synthesis in the case of very few studies. *Research Synthesis Methods, 9*(3), 382-392. <https://doi.org/10.1002/jrsm.1297>
- Biglan, M., Gilpin, E. A., Rohrbach, L. A., & Pierce, J. P. (2004). Is there a simple correction factor for comparing adolescent tobacco-use estimates from school-and home-based surveys? *Nicotine and Tobacco Research, 6*(3), 427-437. <https://doi.org/10.1080/14622200410001696592>
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. John Wiley & Sons. <https://doi.org/10.1002/9780470743386>
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2014). *Comprehensive meta-analysis* (Version 3). Biostat.
- Brener, N. D., Eaton, D. K., Kann, L., Grunbaum, J. A., Gross, L. A., Kyle, T. M., & Ross, J. G. (2006). The association of survey setting and mode with self-reported health risk behaviors among high

- school students. *Public Opinion Quarterly*, 70(3), 354–374.
<https://doi.org/10.1093/poq/nfl003>
- Chaiken, J. M., & Chaiken, M. R. (1982). *Varieties of criminal behavior*. Rand Corporation.
- Chan, J. H., Myron, R., & Crawshaw, M. (2005). The efficacy of non-anonymous measures of bullying. *School Psychology International*, 26(4), 443–458.
<https://doi.org/10.1177/0143034305059020>
- Clark, J. P., & Tifft, L. L. (1966). Polygraph and interview validation of self-reported deviant behavior. *American Sociological Review*, 31(4), 516-523. <https://doi.org/10.2307/2090775>
- Cops, D., De Boeck, A., & Pleysier, S. (2016). School vs. mail surveys: Disentangling selection and measurement effects in self-reported juvenile delinquency. *European Journal of Criminology*, 13(1), 92–110. <https://doi.org/10.1177/1477370815608883>
- Cutler, S. F., Wallace, P. G., & Haines, A. P. (1988). Assessing alcohol consumption in general practice patients - A comparison between questionnaire and interview (findings of the Medical Research Council's general practice research framework study on lifestyle and health). *Alcohol and Alcoholism*, 23(6), 441–450. <https://doi.org/10.1093/oxfordjournals.alcalc.a044844>
- Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior. *Personnel Psychology*, 47(4), 817-829.
<https://doi.org/10.1111/j.1744-6570.1994.tb01578.x>
- Daylor, J. M., Blalock, D. V., Davis, T., Klauberg, W. X., Stuewig, J., & Tangney, J. P. (2019). Who tells the truth? Former inmates' self-reported arrests vs. official records. *Journal of Criminal Justice*, 63, 49-57. <https://doi.org/10.1016/j.jcrimjus.2019.04.002>
- Durant, L. E., Carey, M. P., & Schroder, K. E. (2002). Effects of anonymity, gender, and erotophilia on the quality of data obtained from self-reports of socially sensitive behaviors. *Journal of Behavioral Medicine*, 25(5), 439–467. <https://doi.org/10.1023/A:1020419023766>
- Eaton, D. K., Brener, N. D., Kann, L., Denniston, M. M., McManus, T., Kyle, T. M., Roberts, A. M., Flint, K. H., & Ross, J. G. (2010). Comparison of paper-and-pencil versus Web administration of the Youth Risk Behavior Survey (YRBS): Risk behavior prevalence estimates. *Evaluation Review*, 34(2), 137-153. <https://doi.org/10.1177/0193841X10362491>
- Enzmann, D. (2013). The impact of questionnaire design on prevalence and incidence rates of self-reported delinquency: Results of an experiment modifying the ISRD-2 questionnaire. *Journal of Contemporary Criminal Justice*, 29(1), 147-177.
<https://doi.org/10.1177/1043986212470890>

- Farrington, D. P. (1973). Self-reports of deviant behavior: Predictive and stable? *Journal of Criminal Law and Criminology*, *64*(1), 99-110. <https://doi.org/10.2307/1142661>
- Gomes, H. S., Maia, Â., & Farrington, D. P. (2018a). Measuring offending: Self-reports, official records, systematic observation and experimentation. *Crime Psychology Review*, *4*(1), 26-44. <https://doi.org/10.1080/23744006.2018.1475455>
- Gomes, H. S., Maia, Â., & Farrington, D. P. (2018b, November 14-17). *Method effects in measuring self-reported offending: an experimental comparison of personal, paper-and-pencil, and computer assisted interviews* [Paper presentation]. 74th American Society of Criminology Annual Meeting (ASC), Atlanta, GA, U.S.
- Gribble, J. N., Miller, H. G., Cooley, P. C., Catania, J. A., Pollack, L., & Turner, C. F. (2000). The impact of T-ACASI interviewing on reported drug use among men who have sex with men. *Substance Use & Misuse*, *35*(6-8), 869-890. <https://doi.org/10.3109/10826080009148425>
- Gribble, J. N., Rogers, S. M., Miller, H. G., & Turner, C. R. (1998). Measuring AIDS-related behaviors in older populations: Methodological issues. *Research on Aging*, *20*(6), 798-821. <https://doi.org/10.1177/0164027598206009>
- Grysmar, B., & Johnson, C. (2010). *Effects of value affirmation on drug use disclosure in patients entering a community mental health center* (Publication No. 3431796) [Doctoral dissertation, Hofstra University]. ProQuest Dissertations and Theses Global.
- Hamby, S., Sugarman, D. B., & Boney-McCoy, S. (2006). Does questionnaire format impact reported partner violence rates?: An experimental study. *Violence and Victims*, *21*(4), 507-518. <https://doi.org/10.1891/0886-6708.21.4.507>
- Hardt, R. H., & Peterson-Hardt, S. (1977). On determining the quality of the delinquency self-report method. *Journal of Research in Crime and Delinquency*, *14*(2), 247-259. <https://doi.org/10.1177/002242787701400210>
- Hindelang, M. J., Hirschi, T., & Weis, J. G. (1981). *Measuring delinquency*. Sage.
- Horney, J., & Marshall, I. H. (1992). An experimental comparison of two self-report methods for measuring lambda. *Journal of Research in Crime and Delinquency*, *29*(1), 102-121. <https://doi.org/10.1177/0022427892029001006>
- Huang, F. L., & Cornell, D. G. (2015). The impact of definition and question order on the prevalence of bullying victimization using student self-reports. *Psychological Assessment*, *27*(4), 1484-1493. <https://doi.org/10.1037/pas0000149>

- Jolliffe, D., & Farrington, D. P. (2014). Self-reported offending: Reliability and validity. In G. Bruinsma, & D. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice* (pp. 4716-4723). Springer. https://doi.org/10.1007/978-1-4614-5690-2_648
- King, C. A., Hill, R. M., Wynne, H. A., & Cunningham, R. M. (2012). Adolescent suicide risk screening: the effect of communication about type of follow-up on adolescents' screening responses. *Journal of Clinical Child & Adolescent Psychology, 41*(4), 508-515. <https://doi.org/10.1080/15374416.2012.680188>
- Kivivuori, J., Salmi, V., & Walser, S. (2013). Supervision mode effects in computerized delinquency surveys at school: Finnish replication of a Swiss experiment. *Journal of Experimental Criminology, 9*(1), 91-107. <https://doi.org/10.1007/s11292-012-9162-z>
- Kleck, G., & Roberts, K. (2012). What survey modes are most effective in eliciting self-reports of criminal or delinquent behavior? In L. Gideon (Ed.), *Handbook of survey methodology for the social sciences* (pp. 417-439). Springer. https://doi.org/10.1007/978-1-4614-3876-2_24
- Knapp, H., & Kirk, S. A. (2003). Using pencil and paper, Internet and touch-tone phones for self-administered surveys: Does methodology matter? *Computers in Human Behavior, 19*(1), 117-134. [https://doi.org/10.1016/S0747-5632\(02\)00008-0](https://doi.org/10.1016/S0747-5632(02)00008-0)
- Krohn, M. D., Lizotte, A. J., Phillips, M. D., Thornberry, T. P., & Bell, K. A. (2013). Explaining systematic bias in self-reported measures: Factors that affect the under-and over-reporting of self-reported arrests. *Justice Quarterly, 30*(3), 501-528. <https://doi.org/10.1080/07418825.2011.606226>
- Krohn, M. D., Waldo, G. P., & Chiricos, T. G. (1974). Self-reported delinquency: A comparison of structured interviews and self-administered checklists. *Journal of Criminal Law and Criminology, 65*(4), 545-553. <https://doi.org/10.2307/1142528>
- Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 263-313). Emerald.
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity, 47*(4), 2025-2047. <https://doi.org/10.1007/s11135-011-9640-9>
- Kulik, J. A., Stein, K. B., & Sarbin, T. R. (1968). Disclosure of delinquent behavior under conditions of anonymity and nonanonymity. *Journal of Consulting and Clinical Psychology, 32*(5, Pt1), 506-509. <https://doi.org/10.1037/h0026260>
- Loeber, R., Farrington, D. P., Hipwell, A. E., Stepp, S. D., Pardini, D., & Ahonen, L. (2015). Constancy and change in the prevalence and frequency of offending when based on longitudinal self-reports

- or official records: Comparisons by gender, race, and crime type. *Journal of Developmental and Life-Course Criminology*, *1*(2), 150–168. <https://doi.org/10.1007/s40865-015-0010-5>
- Loftus, E. F., & Marburger, W. (1983). Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition*, *11*(2), 114–120. <https://doi.org/10.3758/BF03213465>
- Lucia, S., Herrmann, L., & Killias, M. (2007). How important are interview methods and questionnaire designs in 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*(1), 39-64. <https://doi.org/10.1007/s11292-007-9025-1>
- Maxfield, M. G., Weiler, B. L., & Widom, C. S. (2000). Comparing self-reports and official records of arrests. *Journal of Quantitative Criminology*, *16*(1), 87–110. <https://doi.org/10.1023/A:1007577512038>
- Moskowitz, J. M. (2004). Assessment of cigarette smoking and smoking susceptibility among youth: telephone computer-assisted self-interviews versus computer-assisted telephone interviews. *Public Opinion Quarterly*, *68*(4), 565–587. <https://doi.org/10.1093/poq/nfh040>
- Nye, F. I., & Short, J. F. (1957). Scaling delinquent behavior. *American Sociological Review*, *22*(3), 326–331. <https://doi.org/10.2307/2088474>
- Ong, A. D., & Weiss, D. J. (2000). The impact of anonymity on responses to sensitive questions. *Journal of Applied Social Psychology*, *30*(8), 1691–1708. <https://doi.org/10.1111/j.1559-1816.2000.tb02462.x>
- Piquero, A. R., Schubert, C. A., & Brame, R. (2014). Comparing official and self-report records of offending across gender and race/ethnicity in a longitudinal study of serious youthful offenders. *Journal of Research in Crime and Delinquency*, *51*(4), 526–556. <https://doi.org/10.1177/0022427813520445>
- Porterfield, A. L. (1943). Delinquency and its outcome in court and college. *American Journal of Sociology*, *49*(3), 199–208. <https://doi.org/10.1086/219369>
- Potdar, R., & Koenig, M. A. (2005). Does audio-CASI improve reports of risky behavior? Evidence from a randomized field trial among young urban men in India. *Studies in Family Planning*, *36*(2), 107-116. <https://doi.org/10.1111/j.1728-4465.2005.00048.x>
- Rehm, J., & Spuhler, T. (1993). Measurement error in alcohol consumption: The Swiss health survey. *European Journal of Clinical Nutrition*, *47*(Suppl 2), S25-S30.

- Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*(5), 754-775. <https://doi.org/10.1037/0021-9010.84.5.754>
- Schore, J., Maynard, R., & Piliavin, I. (1979). *The accuracy of self-reported arrest data*. Mathematica Policy Research.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93-105. <https://doi.org/10.1037/0003-066X.54.2.93>
- Sobell, L. C., & Sobell, M. B. (1981). Effects of three interview factors on the validity of alcohol abusers' self-reports. *American Journal of Drug and Alcohol Abuse, 8*(2), 225-237. <https://doi.org/10.3109/00952998108999127>
- Strang, E., & Peterson, Z. D. (2020). Use of a bogus pipeline to detect men's underreporting of sexually aggressive behavior. *Journal of Interpersonal Violence, 35*(1-2), 208-232. <https://doi.org/10.1177/0886260516681157>
- Thornberry, T. P., & Krohn, M. D. (2000). The self-report method for measuring delinquency and crime. In D. Duffee (Ed.), *Measurement and analysis of crime and justice* (pp. 33-84). U.S. National Institute of Justice.
- Tourangeau, R., & McNeeley, M. E. (2003). Measuring crime and crime victimization: Methodological issues. In J. V. Pepper, & C. V. Petrie (Eds.), *Measurement problems in criminal justice research: Workshop summary* (pp. 10-42). National Academy Press.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511819322>
- Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60*(2), 275-304. <https://doi.org/10.1086/297751>
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-883. <https://doi.org/10.1037/0033-2909.133.5.859>
- Trapl, E. S., Taylor, H. G., Colabianchi, N., Litaker, D., & Borawski, E. A. (2013). Value of audio-enhanced handheld computers over paper surveys with adolescents. *American Journal of Health Behavior, 37*(1), 62-69. <https://doi.org/10.5993/AJHB.37.1.7>

- Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science*, *280*(5365), 867–873. <https://doi.org/10.1126/science.280.5365.867>
- Turner, C. F., & Miller, H. G. (1997). Monitoring trends in drug use: strategies for the 21st century. *Substance Use and Misuse*, *32*(14), 2093–2103. <https://doi.org/10.3109/10826089709035621>
- van de Looij-Jansen, P. M., & de Wilde, E. J. (2008). Comparison of web-based versus paper-and-pencil self-administered questionnaire: Effects on health indicators in Dutch adolescents. *Health Services Research*, *43*(5p1), 1708-1721. <https://doi.org/10.1111/j.1475-6773.2008.00860.x>
- van de Looij-Jansen, P. M., Goldschmeding, J. E., & de Wilde, E. J. (2006). Comparison of anonymous versus confidential survey procedures: Effects on health indicators in Dutch adolescents. *Journal of Youth and Adolescence*, *35*(4), 652-658. <https://doi.org/10.1007/s10964-005-9027-0>
- Walser, S., & Killias, M. (2012). Who should supervise students during self-report interviews? A controlled experiment on response behavior in online questionnaires. *Journal of Experimental Criminology*, *8*(1), 17-28. <https://doi.org/10.1007/s11292-011-9129-5>
- Wilson, D. B. (2009). Missing a critical piece of the pie: Simple document search strategies inadequate for systematic reviews. *Journal of Experimental Criminology*, *5*, 429–440. <https://doi.org/10.1007/s11292-009-9085-5>
- Wolter, F., & Laier, B. (2014). The effectiveness of the item count technique in eliciting valid answers to sensitive questions: An evaluation in the context of self-reported delinquency. *Survey Research Methods*, *8*(3), 153-168. <https://doi.org/10.18148/srm/2014.v8i3.5819>
- Wolter, F., & Preisendörfer, P. (2013). Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods and Research*, *42*(3), 321–353. <https://doi.org/10.1177/0049124113500474>

CHAPTER IV

HOW SENSITIVE ARE SELF-REPORTS OF OFFENDING?:
THE IMPACT OF RECALL PERIODS ON QUESTION SENSITIVITY

Manuscript Submitted:

Gomes, H. S., Farrington, D. P., Krohn, M. D., & Maia, Â. (2021b). *How sensitive are self-reports of offending?: The impact of recall periods on question sensitivity* [Manuscript submitted for publication].

School of Psychology, University of Minho.

HOW SENSITIVE ARE SELF-REPORTS OF OFFENDING?: THE IMPACT OF RECALL PERIODS ON QUESTION SENSITIVITY

Abstract

Although research on sensitive topics has produced a large body of knowledge on how to improve the quality of self-reported data, little is known regarding the sensitivity of offending questions, and much less is known regarding how topic sensitivity is affected by recall periods. In this study, we developed a multi-dimensional assessment of item sensitivity in order to assess and rank the sensitivity of offending and drug use items. Second, to explore the impact of recall period on respondents' perceptions of question sensitivity, we have experimentally compared questions with different time frames (i.e., lifetime, past-year, and past-month). Our results provided a ranking of sensitivity of offending and drug use questions. Furthermore, the experimental manipulation showed that questions about recent time frames were rated as more sensitive than questions covering a longer period of time. The present findings allow future methodological research on offending behavior to control for question sensitivity. Also, this study shows that recall periods impact respondents' perceptions of question sensitivity.

Keywords: Sensitive questions; Social desirability; Measurement error; Offending; Recall periods

Introduction

The study of criminal and deviant behavior is heavily reliant on the self-report methodology (Gomes et al., 2018; Thornberry & Krohn, 2000). However, despite widely shared concerns about the validity of offending data provided by self-reported methods, experimental research trying to assess and improve the quality of self-reports of offending (SRO) is very scarce (Gomes et al., 2019). Survey methodologists have long shown how multiple factors (e.g., question wording, response format, etc.) affect respondents' answers (e.g., Schwarz, 1999), especially in regards to sensitive topics (e.g., Tourangeau et al., 2000). Unfortunately, survey methodologists only very rarely include offending items in their experiments. Therefore, in our study, we have developed a measure of item sensitivity in order to estimate how sensitive are SRO and drug use. Furthermore, since offending questions are usually asked in reference to lifetime, past-year, and/or past-month prevalence, we have experimentally manipulated the recall period of each item in order to test the impact of time frame on participants' evaluations of item sensitivity.

Self-reports of sensitive questions

When responding to self-report questions about behavior, participants have to understand the question, remember, add, average, and/or combine the information in order to provide a valid response (Tourangeau et al., 2000). All these comprehension, memory, and response processes create multiple opportunities for measurement error (see Schwarz, 1999). Adding to these generic self-report issues, researchers who are interested in studying sensitive topics have to deal with the fact that, when reporting undesirable behaviors, participants may tend to deliberately distort and edit their answers in order to avoid disclosing potentially embarrassing or incriminating information (Tourangeau & Yan, 2007).

Evidence supporting the motivated misreporting hypothesis is well established in the literature (see Tourangeau & Yan, 2007). First, multiple research has shown how participants tend to systematically underreport socially undesirable behavior, such as food intake (e.g., Wehling & Lusher, 2019), drug use (e.g., Palamar et al., 2021), and many other sensitive topics (for a review see Krumpal, 2013). Second, measurement procedures that increase respondents' motivation to report the truth (e.g., bogus pipeline), as well as procedures that reduce motivation to misreport (e.g., indirect measurement or self-administration), impact participants' reports of sensitive behavior but show little to no effects on less sensitive topics (Tourangeau & Yan, 2007).

For example, experiments on modes of administration show that participants tend to report a higher prevalence of sensitive behaviors (such as illegal drug use or risky sexual behavior) in self-administered conditions compared to face-to-face interviews, but show no mode effects on low sensitivity topics (e.g., questions on job satisfaction and personality scales) (Gnambs & Kaspar, 2015; Richman et al., 1999). The benefits of self-administration are usually explained by the respondents' increased sense of privacy and confidentiality in self-administered settings, compared to personal interview conditions where respondents have to report their behavior to a third person. On the other hand, responses on non-sensitive information are less affected by self-administration because there is no motivation to conceal (Tourangeau et al., 2000). These findings suggest that mode of administration effects result from a motivated process of respondents' editing their answers in a socially desirable way, mostly when they report their answers to a third person.

Further, several studies noted that the benefits of self-administration in the reporting of sensitive behaviors tend to be larger for more recent time frames than for more distant ones (Tourangeau & McNealey, 2003; Tourangeau et al., 2000; Tourangeau & Yan, in press). For example, in the studies carried out by Turner et al. (1992) and by Schober et al. (1992), the benefits of self-administration in the respondents' disclosure of drug use (i.e., higher reports of drug use in self-administered conditions) are lowest for lifetime, higher for past-year, and highest for past-month prevalence. In light of the previous argument, these findings suggest that asking someone to report recent socially undesirable behaviors is a more sensitive question than asking someone to report the same behavior over a longer period of time. Respondents may feel less threatened or embarrassed to report sensitive behavior in the distant past, than disclosing these practices over a recent time frame. However, the impact of recent time frames on question sensitivity has not yet been demonstrated.

Definition of sensitive questions

Tourangeau and colleagues (Tourangeau et al., 2000; Tourangeau & Yan, 2007) described three aspects that make a sensitive topic (i.e. intrusiveness, threat of disclosure, and social desirability). Intrusiveness refers to questions on inappropriate, out-of-bounds (i.e., "taboo") topics. In this sense, the question itself is intrusive, and people may see it as an invasion of privacy, regardless of what the socially acceptable answer might be. Second, threat of disclosure refers to participants' concerns about the potential consequences of their answers being disclosed to a third party. Third, social desirability refers to the extent to which a question requires socially unacceptable or undesirable answers (Tourangeau & Yan, 2007). Some previous studies attempted to assess item sensitivity (e.g., Bradburn et al., 1979;

Fortier et al., 2020; Holbrook et al., 2003; Sudman & Bradburn, 1974). However, these evaluations of topic sensitivity usually focus on only one aspect of sensitivity and lack an assessment of the main dimensions of topic sensitivity, namely, Intrusiveness, Threat of disclosure, and Social desirability (Tourangeau & Yan, 2007).

Present study

This study has two main objectives. First, we intend to assess and rank the question sensitivity of offending and drug use items. Although it might be assumed that questions on illegal behavior are sensitive, we do not know how sensitive these questions really are. Furthermore, we do not know which offending items are the most sensitive within an offending questionnaire.

The present study will provide a ranking of the sensitivity of offending questions, which will allow future researchers to control for the effect of question sensitivity in their methodological studies with offending variables. Second, we aim to explore the impact of recall period on respondents' perceptions of question sensitivity. Survey questions on offending, similarly to other behavioral measures, are usually asked in reference to either lifetime, past-year, and/or past-month prevalence. However, we are unaware of any study that has explored the impact of recall periods on question sensitivity. Therefore, this study provides a contribution to the study of sensitive questions by testing this hypothesis.

Methods

Sample and study design

This study was conducted in Portugal with a sample of 269 university students. A total of 20 participants failed to complete the questionnaire and were removed. The final sample was composed of 249 university students (89.6% females, $n = 223$), mostly Portuguese nationals (90.8%, $n = 226$), aged between 17 and 51 years ($M = 22.74$, $SD = 6.60$). Participants were recruited both through institutional e-mailing and in exchange for class credit.

Mean comparisons showed no statistically significant sex differences in the reports of females and males on question sensitivity for the behavioral variables in the study (i.e., offending, contact with the police, and drug use), with the exception of the sexual behavior question in which females ($M = 4.07$, $SD = 1.38$) reported that this question was more sensitive than did male participants ($M = 3.32$, $SD = 1.48$) ($t(247) = -2.60$, $p < .05$).

Measures

Sensitive behavioral items

Participants reported their evaluations of sensitivity for 23 behavior items; 15 items on offending, one item regarding past contacts with the police, six items referring to drug use, and one item on sexual behavior (i.e., sexual intercourse with someone) (see Table 8). These behavioral items were selected from the International Self-Report Delinquency 3 questionnaire (ISR3; Enzmann et al., 2018; Martins et al., 2015), with the exception of tobacco, derbisol (a fictitious drug), and the sexual behavior item, which were added by our team.

Measures of question sensitivity

We have created three questions designed to assess the three dimensions of question sensitivity proposed by Tourangeau and Yan (2007). Regarding the first dimension, i.e. Intrusiveness, after each sensitive behavioral item participants were asked “Do you think this question is too personal?” (from 1-“Nothing personal at all” to 7-“Very personal”). Regarding Threat of disclosure, we asked “Imagine your answer is YES. Would you feel uncomfortable if other people [colleagues, parents, friends, etc.] could see your answer to this question?” (from 1-“Nothing uncomfortable at all” to 7-“Very uncomfortable”). Finally, for Social desirability, participants responded to “Do you think other people answer honestly and truthfully to this question?” (from 1-“Completely false” to 7-“Completely true”). Average sensitivity scores were computed for each behavioral item, so that higher values represented higher topic sensitivity.

Procedures

This study was carried out online using Qualtrics software during July and November of 2019. After completing a brief socio-demographic questionnaire, participants were invited to rate the sensitivity of selected behavioral items. Participants did not respond if they had themselves practiced any of these behaviors. Initial instructions indicated that they would be presented with behavioral items typically used in anonymous and confidential scientific studies and that we were only interested in their opinion regarding these items. The 23 behavioral items were presented in a random order in three blocks, each corresponding to one dimension of question sensitivity (Tourangeau & Yan, 2007), where respondents provided their sensitivity ratings for every behavioral item. Ethical approval was provided by the Institutional Review Board of the University of Minho.

Experimental design

We have manipulated the recall period for behavioral items. For every block, the recall period of the behavioral items was randomly selected. Behavioral items were presented either in a lifetime (e.g., “Have you ever in your life stolen a bicycle?”), past-year (e.g., “In the last 12 months, have you stolen a bicycle?”), or past-month (e.g., “In the last 30 days, have you stolen a bicycle?”) prevalence format.

Data analysis

Regarding our first objective, we used average scores to rank the behavioral items from the least to the most sensitive topics. We used one-way ANOVAs with Gabriel’s post-hoc test to explore the impact of recall periods on respondents’ assessments of behavioral items’ sensitivity. Statistical analyses were carried out using SPSS v27 software (IBM SPSS, Chicago, IL).

Results

Table 8 presents the results of respondents’ evaluations of sensitivity for each item, organized from the lowest to the highest sensitivity question. Within the offending items, Illegal downloading and Group fight were the least sensitive questions, while Robbery and Assault scored as the most sensitive offending questions. Within drug use questions, Alcohol and Tobacco ranked as the least sensitive items, while Ecstasy/LSD/amphetamines and Heroin/cocaine/crack scored as the most sensitive questions. Inter-dimensional comparisons show that behavioral items scored similarly throughout the three dimensions. With the exception of the question on sexual behavior that ranked as the most sensitive question in the Intrusiveness dimension and, at the same time, ranked as one of the least sensitive questions on Threat of disclosure and Social desirability dimensions.³

Recall periods

Results regarding the impact of recall periods on question sensitivity are described in Table 9. The manipulation of time periods had no effect on Intrusiveness. As for the two remaining dimensions of sensitivity (i.e., Threat of disclosure and Social desirability), findings for overall offending, drug use, and sexual behavior, as well as contact with the police in the Social desirability dimension, showed that respondents rated recent time frames (i.e., 12-month or 30-day periods) as statistically more sensitive

³ In a pilot study with students from an American university, we replicated this study and found results very similar to those reported here (see Table 10).

than the same questions regarding lifetime prevalence. Figure 5 illustrates the results of the impact of time frame on the respondents' ratings of question sensitivity for the offending items.

Table 8

Average question sensitivity of behavioral items (Portuguese experiment)

	Intrusiveness		Threat of disclosure		Social desirability		Question Sensitivity	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Offending (overall mean)	4.16	1.67	5.62	1.00	4.45	1.15	4.75	0.93
Illegal downloading	2.40	1.60	2.14	1.64	1.94	1.48	2.16	1.23
Group fight	3.22	1.75	4.56	1.66	2.95	1.53	3.58	1.18
Graffiti	3.26	1.76	4.38	1.79	3.34	1.68	3.66	1.30
Carrying a weapon	4.22	1.94	5.14	1.89	4.12	1.69	4.49	1.37
Vandalism	3.93	1.90	5.61	1.46	4.16	1.67	4.57	1.22
Bike theft	3.98	1.90	5.72	1.45	4.12	1.79	4.61	1.26
Shoplifting	4.36	1.95	6.00	1.40	4.64	1.62	5.00	1.18
Stealing from a person	4.37	2.01	6.08	1.35	4.67	1.59	5.04	1.18
Animal cruelty	4.24	2.06	6.27	1.26	4.83	1.74	5.11	1.17
Stealing from a car	4.45	1.99	6.21	1.29	4.88	1.59	5.18	1.12
Car theft	4.63	2.02	6.45	1.17	5.32	1.58	5.47	1.12
Burglary	4.75	2.09	6.45	1.12	5.43	1.67	5.54	1.18
Drug sales	4.90	2.02	6.28	1.30	5.56	1.45	5.58	1.10
Robbery	4.81	2.03	6.59	1.00	5.39	1.46	5.60	1.02
Assault	4.94	2.02	6.47	1.18	5.45	1.54	5.62	1.09
Contact with police	4.42	2.00	6.11	1.37	4.55	1.56	5.03	1.13
Drug use (overall mean)	3.66	1.43	4.83	1.23	3.73	1.14	4.07	0.91
Alcohol	2.24	1.44	2.11	1.53	1.73	1.24	2.03	1.01
Tobacco	2.34	1.49	3.42	2.08	1.96	1.41	2.57	1.18
Cannabis/marijuana/hash	4.04	1.85	5.23	1.88	3.98	1.79	4.42	1.40
Derbisol	4.02	1.89	5.58	1.74	4.52	1.62	4.71	1.23
Ecstasy/LSD/amphetamines	4.58	1.92	6.01	1.51	4.80	1.59	5.13	1.17
Heroin/cocaine/crack	4.64	1.94	6.25	1.36	4.99	1.62	5.29	1.12
Sexual behavior	5.16	1.79	4.16	2.19	2.65	1.52	3.99	1.40

Discussion

This is the only experimental study that we are aware of that explores the impact of recall periods on question sensitivity. In doing so, we have developed an assessment of topic sensitivity, which allowed, first, an evaluation and rank of the sensitivity of offending and drug use behavioral questions. Second, we

tested the impact of time frames within behavioral questions on the respondents' perceptions of question sensitivity.

Table 9

Mean comparisons of question sensitivity by recall period

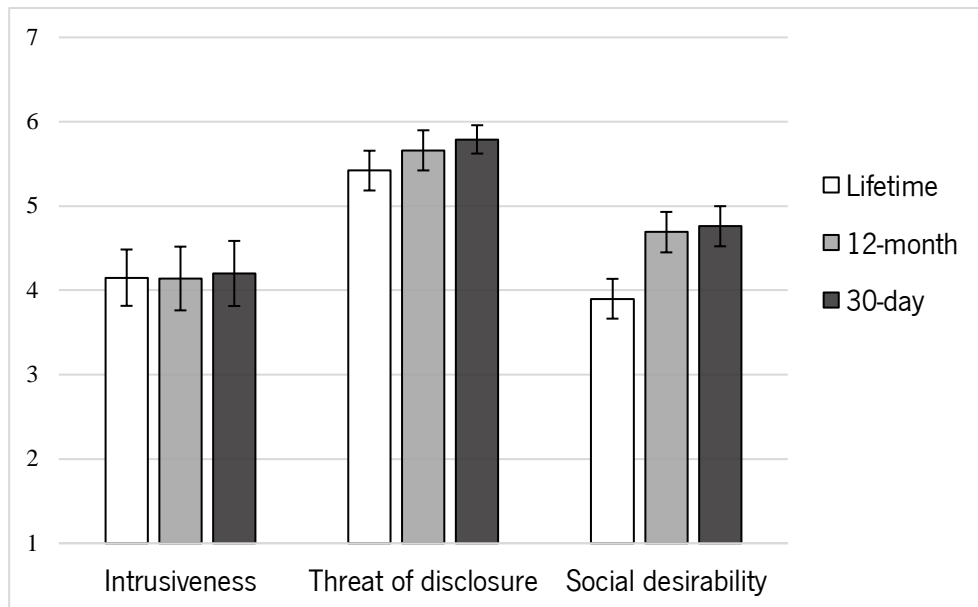
	Intrusiveness						<i>F</i> (2)	<i>p</i>	η^2
	Lifetime (<i>n</i> = 82)		12-month (<i>n</i> = 87)		30-day (<i>n</i> = 80)				
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Offending (overall)	4.15 _a	1.51	4.14 _a	1.76	4.20 _a	1.73	0.03	.968	.00
Contact with the police	4.43 _a	1.85	4.48 _a	2.09	4.34 _a	2.06	0.11	.895	.00
Drug use (overall)	3.62 _a	1.33	3.72 _a	1.51	3.63 _a	1.44	0.13	.877	.00
Sexual behavior	5.04 _a	1.69	5.32 _a	1.94	5.11 _a	1.72	0.58	.561	.01
	Threat of disclosure						<i>F</i> (2)	<i>p</i>	η^2
	Lifetime (<i>n</i> = 85)		12-month (<i>n</i> = 83)		30-day (<i>n</i> = 81)				
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Offending (overall)	5.42 _a	1.09	5.66 _{ab}	1.08	5.79 _b	.75	3.03	.050	.02
Contact with the police	5.82 _a	1.67	6.24 _a	1.28	6.28 _a	1.05	2.91	.056	.02
Drug use (overall)	4.50 _a	1.22	4.97 _b	1.19	5.04 _b	1.21	4.99	.007	.04
Sexual behavior	3.53 _a	2.11	4.22 _{ab}	2.11	4.77 _b	2.19	6.97	.001	.05
	Social desirability						<i>F</i> (2)	<i>p</i>	η^2
	Lifetime (<i>n</i> = 83)		12-month (<i>n</i> = 82)		30-day (<i>n</i> = 84)				
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Offending (overall)	3.90 _a	1.07	4.69 _b	1.08	4.76 _b	1.09	16.22	<.001	.12
Contact with the police	4.05 _a	1.53	4.59 _{ab}	1.58	5.01 _b	1.44	8.48	<.001	.06
Drug use (overall)	3.38 _a	1.08	3.85 _b	1.18	3.96 _b	1.10	6.47	.002	.05
Sexual behavior	2.37 _a	1.51	2.61 _{ab}	1.52	2.96 _b	1.49	3.25	.041	.03

Note. Each subscript letter denotes a subset of behavioral categories which illustrate the results of the post-hoc analysis; different letters represent statistically significant differences between columns.

Our findings provide an evaluation of topic sensitivity for offending and drug use questions that allows future methodological research to control for the effect of question sensitivity. Furthermore, most offending items scored higher on sensitivity than the sexual behavior item, covering a topic that is often referred to as highly sensitive. This finding is consistent with our initial expectation that some questions on offending behavior are perceived as highly sensitive.

Figure 5

Average scores of sensitivity for offending items by recall period (Error bars are 95 percent confidence intervals)



On a different aspect, offending and drug use questions behaved very similarly throughout the three dimensions. In other words, items scoring as high sensitivity in one dimension also scored high in the remaining dimensions, and vice-versa. However, the item about sexual behavior showed a different trajectory. The sexual behavior item was ranked as the most sensitive questions in Intrusiveness, but ranked as one of the less sensitive questions in the dimensions of Threat of disclosure and Social desirability, slightly above the smoking tobacco question. These findings suggest that the same question might be perceived as highly sensitive on one dimension but have low sensitivity on a different dimension. We do not know which aspects of topic sensitivity have more effect on the quality of participants' reports and more research on topic sensitivity is needed.

Regarding the manipulation of recall periods, findings showed that asking questions about sensitive behavior over longer periods of time are generally regarded as less sensitive than asking the same questions for more recent time frames. In both dimensions of Threat of disclosure and Social desirability, respondents consistently reported that recent time frames (i.e., past-year and/or past-month) were more sensitive than asking the same questions regarding lifetime prevalence of behavior. It is possible that respondents feel less threatened in disclosing sensitive behavior that might have happened in the distant past. Conversely, respondents might feel shame or fear potentially incriminating consequences of reporting recent illegal behavior.

These results are consistent with previous methodological experiments that found higher benefits of self-administration for recent time frames than for more distant ones (e.g., Schober et al., 1992, Turner et al., 1992). These findings consistently show that item sensitivity increases with recency of the behavior, and survey researchers should take that into account when asking sensitive questions. Bradburn et al. (2004), for example, suggest that, since questions about current behavior are more threatening, questionnaires about socially undesirable behavior should start with lifetime questions, rather than starting with questions about current behavior.

On the other hand, the dimension of Intrusiveness was not affected by recall periods. In other words, respondents described questions about recent offending and drug use as intrusive as questions about offending and drug use over the lifetime. These results might be understood under the definition of Intrusiveness, where the topic of the question itself is sensitive, regardless of the circumstances and whether the respondent has or not practiced the behavior referred to in the question (Tourangeau et al., 2000). Therefore, respondents might feel that questions on sensitive topics are none of the researcher's business independently of the time frame.

In conclusion, the present study shows that question sensitivity is affected by recall periods. Questions about recent behavior are perceived by respondents as more sensitive than questions about behavior that might have happened over a longer period of time. Considering that question sensitivity affects the quality of participants' reports (Tourangeau & Yan, 2007), behavioral reports over recent time frames may be subject to increased measurement error, such as deliberate misreporting. Further research is needed to better understand how recall periods affect the quality of self-reports of behavior.

Table 10*Average question sensitivity of behavioral items for the American pilot study (n = 43)*

	Intrusiveness		Threat of disclosure		Social desirability		Question Sensitivity	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Offending (overall mean)	3.53	1.64	5.00	1.14	4.14	0.80	4.17	0.86
Illegal downloading	2.19	1.28	1.98	1.35	1.97	0.96	2.04	0.88
Group fight	3.12	1.73	3.93	1.85	3.28	1.17	3.41	1.10
Graffiti	2.86	1.71	3.55	1.86	2.95	1.21	3.10	1.14
Carrying a weapon	3.19	1.76	3.48	2.05	2.87	1.26	3.16	1.22
Vandalism	3.51	1.84	5.12	1.70	4.00	1.28	4.16	1.19
Bike theft	3.28	1.88	4.69	1.77	3.92	1.18	3.91	1.09
Shoplifting	3.42	1.78	5.12	1.76	3.74	1.39	4.05	1.03
Stealing from a person	3.49	1.83	5.17	1.74	4.08	1.38	4.20	1.08
Animal cruelty	3.88	2.09	6.38	1.27	5.49	1.79	5.17	1.36
Stealing from a car	3.63	1.98	5.36	1.66	4.13	1.28	4.33	1.11
Car theft	3.86	1.98	5.76	1.54	4.97	1.51	4.79	1.11
Burglary	3.81	1.97	5.83	1.43	4.90	1.37	4.78	1.08
Drug sales	4.21	2.07	5.88	1.58	5.03	1.50	4.98	1.12
Robbery	4.26	2.06	6.36	1.21	5.28	1.62	5.22	1.07
Assault	4.26	2.17	6.33	1.24	5.46	1.57	5.27	1.26
Contact with police	4.09	2.14	5.48	1.77	3.87	1.52	4.48	1.22
Drug use (overall mean)	3.02	1.45	4.28	1.28	3.47	0.89	3.56	0.88
Alcohol	2.05	1.23	1.81	1.21	1.56	0.60	1.84	0.73
Tobacco	2.23	1.57	2.76	1.65	1.85	1.16	2.29	0.86
Cannabis/marijuana/hash	3.00	1.51	3.74	2.18	2.67	1.39	3.14	1.11
Derbisol	3.07	1.89	4.90	2.12	4.28	1.73	4.01	1.38
Ecstasy/LSD/amphetamines	3.51	1.94	5.76	1.65	4.46	1.43	4.52	1.26
Heroin/cocaine/crack	4.07	2.04	6.10	1.46	5.13	1.56	5.05	1.17
Sexual behavior	4.16	1.84	3.52	2.24	2.36	1.06	3.44	1.12

References

- Bradburn, N. M., Sudman, S., Blair, E., Locander, W., Miles, C., Singer, E., & Stocking, C. (1979). *Improving interview method and questionnaire design: Response effects to threatening questions in survey research*. Jossey-Bass.
- Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design—for market research, political polls, and social and health questionnaires*. Jossey-Bass.
- Enzmann, D., Kivivuori, J., Marshall, I. H., Steketee, M., Hough, M., & Killias, M. (2018). *A global perspective on young people as offenders and victims: First results from the ISRD3 study*. Springer.
- Fortier, J., Stewart-Tufescu, A., Salmon, S., Davila, I. G., MacMillan, H. L., Gonzalez, A., Mathews, B., Struck, S., Taillieu, T., & Afifi, T. O. (2020). What type of survey research questions are identified by adults as upsetting? A focus on child maltreatment. *Child Abuse & Neglect, 109*, 104764. <https://doi.org/10.1016/j.chiabu.2020.104764>
- Gomes, H. S., Farrington, D. P., Maia, Â., & Krohn, M. D. (2019). Measurement bias in self-reports of offending: A systematic review of experiments. *Journal of Experimental Criminology, 15*(3), 313-339. <https://doi.org/10.1007/s11292-019-09379-w>
- Gomes, H. S., Maia, Â., & Farrington, D. P. (2018). Measuring offending: Self-reports, official records, systematic observation and experimentation. *Crime Psychology Review, 4*(1), 26-44. <https://doi.org/10.1080/23744006.2018.1475455>
- Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods, 47*(4), 1237-1259. <https://doi.org/10.3758/s13428-014-0533-4>
- Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly, 67*, 79-125. <https://doi.org/10.1086/346010>
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity 47*, 2025-2047. <https://doi.org/10.1007/s11135-011-9640-9>
- Martins, P., Mendes, S., & Fernandez-Pacheco, G. (2015, September 2-5). *Cross-cultural adaptation and online administration of the Portuguese Version of ISRD3* [Paper presentation]. 15th Annual Conference of the European Society of Criminology, Porto, Portugal.

- Palamar, J. J., Salomone, A., Rutherford, C., & Keyes, K. M. (2021). Extensive underreported exposure to ketamine among electronic dance music party attendees. *Journal of General Internal Medicine, 36*(1), 235-237. <https://doi.org/10.1007/s11606-020-05672-x>
- Richman, W., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*, 754-775. <https://doi.org/10.1037/0021-9010.84.5.754>
- Schober, S. E., Caces, M. F., Pergamit, M. R., & Branden, L. (1992). Effect of mode of administration on reporting of drug use in the National Longitudinal Survey. In C. F. Turner, J. T. Lessler, & J. C. Gfroerer (Eds.), *Survey measurement of drug use: Methodological studies* (pp. 267-276). National Institute on Drug Abuse.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93-105. <https://doi.org/10.1037/0003-066X.54.2.93>
- Sudman, S., & Bradburn, N. M. (1974). *Response effects in surveys: A review and synthesis*. Aldine.
- Thornberry, T. P., & Krohn, M. D. (2000). The self-report method for measuring delinquency and crime. In D. Duffee (Ed.), *Criminal Justice* (pp. 33-84). U.S. National Institute of Justice.
- Tourangeau, R., & McNeeley, M. E. (2003). Measuring crime and crime victimization: Methodological issues. In J. V. Pepper & C. V. Petrie (Eds.), *Measurement problems in criminal research: Workshop summary* (pp. 10-42). National Academies Press.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859-883. <https://doi.org/10.1037/0033-2909.133.5.859>
- Tourangeau, R., & Yan, T. (in press). Reporting issues in surveys of drug use. *Substance Use and Misuse*.
- Turner, C. F., Lessler, J. T., Devore, J. W. (1992). Effects of mode of administration and wording on reporting of drug use. In C. F. Turner, J. T. Lessler, & J. C. Gfroerer (Eds.), *Survey measurement of drug use: Methodological studies* (pp. 177-219). National Institute on Drug Abuse.
- Wehling, H., & Lusher, J. (2019). People with a body mass index ≥ 30 under-report their dietary intake: A systematic review. *Journal of Health Psychology, 24*(14), 2042-2059. <https://doi.org/10.1177/1359105317714318>

CHAPTER V

THE IMPACT OF MODES OF ADMINISTRATION ON SELF-REPORTS OF OFFENDING:
EVIDENCE FROM TWO METHODOLOGICAL EXPERIMENTS

Manuscript Submitted:

Gomes, H. S., Farrington, D. P., Krohn, M. D., Cunha, A., Jurdi, J., Sousa, B., Morgado, D., Hoft, J., Hartsell, E., Kassem, L., & Maia, Â. (2021). *The impact of modes of administration on self-reports of offending: Evidence from two methodological experiments* [Manuscript submitted for publication].

School of Psychology, University of Minho.

THE IMPACT OF MODES OF ADMINISTRATION ON SELF-REPORTS OF OFFENDING:
EVIDENCE FROM TWO METHODOLOGICAL EXPERIMENTS

Abstract

Objectives: Current knowledge about the causes of offending behavior is heavily reliant on self-reports of offending (SRO). Offending behavior is a highly sensitive topic, and thus may be subject to multiple biasing factors, such as modes of administration. However, methodological research on the impact of modes of administration on SRO is very scarce. Further, the existing evidence conflicts with the general knowledge about responding to sensitive questions. The failure to identify mode effects on SRO threatens the qualities of this methodology and may result in misleading conclusions. This study aimed to test the impact of mode effects on participants' willingness to disclose offending behavior.

Methods: In an attempt to test whether SRO are affected by modes of administration, we carried out two methodological experiments, with a 2 (modes of administration: interviewer-administered vs. self-administered) × 2 (modes of data collection: paper-and-pencil vs. computer interviews) factorial design.

Results: Both experiments consistently showed an increased odds of reporting offending behavior in self-administered surveys, compared with face-to-face interviews. However, these findings were only statistically significant in experiment 1. These experiments provided evidence that data collection with computer-assisted and paper-and-pencil surveys provide comparable estimates of offending, only slightly favoring higher results in paper-and-pencil modes.

Conclusions: The present findings demonstrate that modes of administration affect the respondents' willingness to report their own offenses. Whenever possible, self-administered modes of administration should be preferable over face-to-face interviews.

Keywords: Sensitive questions; Modes of administration; Self-administration; Computer-assisted; Offending; Delinquency

Introduction

Self-reports of offending (SRO) have come a long way since their early stages in the 1950s, where only a few, minor types of delinquent behaviors were included (Thornberry & Krohn, 2000). Skepticism over the utility of these methods compelled criminologists to develop a large body of research on the validity and reliability of SRO (e.g., Farrington, 1973; Huizinga & Elliott, 1986; Jolliffe et al., 2003; Piquero et al., 2014), making self-reports one of the most widely used methods in the study of offending behavior (Gomes et al., 2018). Current knowledge about the prevalence and causes of offending, as well as risk and protective factors for juvenile delinquency, are almost exclusively reliant on the self-report methodology (Cops et al., 2016; Thornberry & Krohn, 2000). However, little is known about the impact of measurement biases, such as the ones caused by modes of administration and questionnaire format, on the reported rates of offending and data quality.

In a recent systematic review of methodological experiments using SRO, Gomes et al. (2019) found 21 experiments that explored a total of 18 different manipulations of potential biases relating to modes of administration, procedures of data collection, and questionnaire design. In this study, contrary to the large body of research on sensitive questions (e.g., Gnamb & Kaspar, 2015; Richman et al., 1999; Tourangeau & Yan, 2007), the methodological experiments on SRO failed to show any evidence of the benefits of self-administration over face-to-face interviews. The lack of evidence for mode effects on SRO led influential studies on crime measurement to conclude that self-reports are valid and stable over different modes of administration (e.g., Thornberry & Krohn, 2000). However, offending behavior is a highly sensitive topic (Gomes et al., 2021) and, unless there are specific features of criminal behavior, the disclosure of offending should be subject to mode effects, at least to the same extent as other types of sensitive behaviors.

In the case that SRO are, in fact, affected by modes of administration, the failure to identify these mode effects will lead researchers to apply unstandardized measurement methods, resulting in biased outcomes and, ultimately, misleading conclusions about offending behavior. In the present study, we have developed two methodological experiments, the first carried out in Portugal and the second a replication study carried out in Florida, with a 2 (modes of administration: interviewer-administered vs. self-administered) \times 2 (modes of data collection: paper-and-pencil vs. computer interviews) factorial design, in order to test whether or not SROs are affected by modes of administration.

Sensitive questions

Sensitive topics in survey research can be defined as intrusive, posing a threat of disclosure, and eliciting socially desirable answers (Tourangeau et al., 2000; Tourangeau & Yan, 2007). An intrusive question can be construed as an inappropriate invasion of privacy. In this sense, the question itself is intrusive, independently of the participant's truthful response. The dimensions of threat of disclosure and social desirability, on the other hand, are a product of the participant's past experience and the perceived likelihood of their answers becoming known to other parties. A question on bicycle theft, for example, is nonconsequential for participants who have never committed such behavior, even if their answers were to become known to other people outside the research study. Participants who have stolen a bicycle, on the other hand, may experience feelings of shame, guilt, or fear of criminal consequences, and thus refrain from providing a truthful answer to this question. As a result, respondents to sensitive questions may tend to systematically underreport their socially undesirable behavior (Tourangeau et al., 2000).

Evidence for the tendency to underreport sensitive behavior is well documented in the literature. For example, Liber and Warner (2018) compared data from cigarette-tax collections and nationwide surveys and concluded that respondents consistently underreport cigarette consumption over time. Giguère et al. (2019) used biomarkers of recent semen exposure among female sex workers in early antiretroviral treatments and concluded that respondents often underreport unprotected sexual intercourse. Studies using biomarkers to determine substance use (provided from blood, urine, saliva, or hair samples) show that respondents consistently underestimate consumption, such as alcohol (e.g., Kabashi et al., 2019; Littlefield et al., 2017; Vinikoor et al., 2018) and other drugs (e.g., Gerdtz et al., 2020; Palamar et al., 2021). Clark and Tiffit (1966) used the polygraph as an external criterion for SRO and found evidence of underreporting of deviant behaviors. Further, studies using indirect measures consistently result in higher rates of reporting sensitive behavior than in direct questioning (Druckman et al., 2015; Kirtadze et al., 2018), including reports of offending behavior (e.g., Wolter & Laier, 2014). Because respondents to sensitive questions tend to underreport their socially undesirable behavior, survey researchers have explored methods to overcome the effects of question sensitivity. For example, measurement methods that provide anonymity and confidentiality to a respondent consistently result in higher rates of sensitive behavior (Bradburn et al., 2004).

The systematic bias of reporting higher rates of sensitive behavior in less threatening measurement conditions, where the motivation to provide socially desirable answers is reduced, cannot be explained by chance, memory faults, or the usual reporting error in survey bias (e.g., Schwarz, 1999). Rather, this evidence is consistent with the deliberate misreporting hypothesis (Bradburn et al., 1979;

Tourangeau et al., 2000). According to the idea of deliberate distortion, respondents to sensitive questions deliberately edit their answers in order to avoid the embarrassment or consequences of admitting such behaviors. As a consequence, survey researchers have created the 'more is better' assumption, in which measurement conditions that result in higher estimates of a socially undesirable behavior are assumed to be the most accurate (Tourangeau & Yan, 2007). This assumption is especially useful in behaviors where there is no gold standard to which self-reported information can be compared, such as offending behavior.

Modes of administration

One key variable that has repeatedly been shown to affect participants' disclosure of sensitive behavior is modes of administration (Richman et al., 1999; Tourangeau & Yan, 2007). Mainly, self-administration of a questionnaire, in contrast to interviewer-administered modes, results in a steep effect in increasing participants' willingness to report sensitive behavior (Sudman & Bradburn, 1974). In face-to-face interviews, participants are requested to disclose their sensitive behavior to a third person (i.e., the interviewer). This is expected to affect participants' perceptions of confidentiality and anonymity, as well as social desirability, causing the above-described mode effects (Schwarz et al., 1991). Methodological experiments have provided evidence that self-administration causes increased rates of reporting multiple types of sensitive behavior, such as undesirable academic attributes (Kreuter et al., 2008), disclosure of non-heterosexual identity (Robertson et al., 2018), number of sexual partners (Jobe et al., 1997), suicidal ideation (Lee et al., 2019), and drug use (e.g., Aquilino, 1994; Butler et al., 2009; Schober et al., 1992; Turner et al., 1992). Tourangeau and Yan (2007) reviewed the survey methodological research on sensitive topics and concluded that respondents are more likely to disclose socially undesirable behaviors in self-administered conditions. Further, Tourangeau and Yan (in press) found that self-administration, in comparison to face-to-face interviews, resulted in an increase of reports of illicit drug use by 30%.

Survey research is increasingly transitioning from traditional paper-and-pencil questionnaires to computer-assisted modes of data collection. Computerized surveys are cheaper, they eliminate the need for printed questionnaires, data are automatically stored in databases and thus reduce data entry error, and computers allow for more complex branching questionnaires with skip questions, etc. (Lucia et al., 2007). Additionally, authors have suggested that computer-assisted modes increase perceived anonymity (e.g., Trau et al., 2013), raising the question of whether computerized modes of data collection impact participants' willingness to disclose sensitive behavior. The research on this particular question is fairly

inconsistent. Some researchers have found no evidence of mode effects caused by modes of data collection (e.g., Bates & Cox, 2008; Beebe et al., 2006; Knapp & Kirk, 2003). Further, the meta-analysis carried out by Dodou and de Winter (2014) found no differences in social desirability between paper-and-pencil and computer-assisted surveys.

On the other hand, some methodological experiments have found higher rates of disclosure in paper-and-pencil questionnaires (e.g., Beebe et al., 1998), while others have found results in the opposite direction, indicating higher reports of sensitive behavior in computer-assisted modes (e.g., Brener et al., 2006). Richman et al. (1999) carried out a meta-analysis and found 61 experiments comparing results obtained in computer-assisted and paper-and-pencil questionnaires (a total of 673 effect sizes). They concluded that, within self-administered modes, computer-assisted surveys resulted in a higher prevalence of sensitive behavior disclosure. More recently, Gnambis and Kaspar (2015) focused on methodological experiments comparing self-administered disclosure in paper-and-pencil and computer-assisted modes of data collection (39 studies and 460 effect sizes). These authors found that computer-assisted surveys resulted in an increased odds of reporting sensitive behavior, especially for highly sensitive topics.

The impact of modes of administration on self-reports of offending

Criminal behavior is a highly sensitive topic. Offenders naturally try to conceal their illegal behavior, and they may feel ashamed or regret their delinquent practices. The disclosure of offending behavior not only causes embarrassment and socially desirable answers, but offenders may also fear potential criminal consequences (Thornberry & Krohn, 2000). Gomes et al. (2021) developed an assessment of question sensitivity based on the three-dimensional definition proposed by Tourangeau and Yan (2007). Gomes et al. (2021) showed that most offending questions scored higher on topic sensitivity than a question about sexual behavior, especially the more serious and violent offenses which participants rated as very highly sensitive. For all these reasons, SRO are expected to be subject to reporting bias, at least to the same extent as other types of sensitive questions.

Unfortunately, methodological research on the response biases of SRO is very scarce. Gomes et al. (2019) systematically reviewed methodological experiments exploring potential response biases in the collection of SRO. In this review, the comparison between self-administered surveys using paper-and-pencil and computer-assisted modes of data collection was the most replicated manipulation within the SRO methodological literature ($k = 10$). Results were very inconsistent. Five experiments found evidence showing higher reports of offending in paper-and-pencil conditions, while the other five experiments

showed higher disclosure in computer-assisted modes. However, similar to previous reviews (Gnambs & Kaspar, 2015; Richman et al., 1999), the overall effect of modes of data collection on SRO showed that computer-assisted modes resulted in higher rates of reporting of sensitive behaviors, though this was only marginally significant.

As for the impact of modes of administration on SRO, Gomes et al. (2019) found a total of four experimental comparisons testing the effect of self-administration on respondents' disclosure of offending behavior. Three experiments compared face-to-face interviews with paper-and-pencil questionnaires and one of these studies also included a comparison between face-to-face interviews and audio-computer-assisted self-interview (ACASI). Results showed no significant effect of self-administration on participants' rates of reported offenses. These results disagree with the general evidence regarding self-reports of sensitive behavior (e.g., Tourangeau and Yan, 2007). However, it is worth considering that two of these studies were carried out more than 40 years ago (i.e., Hindelang et al., 1981; Krohn et al., 1974), and the third study was developed with the objective of testing mode effects on reports of risky behavior and only two types of offenses (i.e., carrying a weapon/gun and engaging in abusive/violent behavior after drinking) were included (Potdar & Koenig, 2005). These features may have limited the ability of these studies to find evidence of mode effects, and relying solely on these findings to conclude that SRO are not affected by modes of administration may be misleading. In sum, the question about what are the best practices to measure SRO is far from settled, and more methodological research using contemporary questionnaires of offending behavior is needed.

The present study

The aim of this study was to test whether SRO are affected by modes of administration and modes of data collection. The lack of evidence showing mode effects on SRO led influential reviews of crime measurement to conclude that modes of administration did not affect participants' willingness to report offending behavior (e.g., Thornberry & Krohn, 2000). However, if the disclosure of criminal behavior is, indeed, affected by modes of administration, similarly to the disclosure of other types of sensitive topics, then using unstandardized modes of administration may have resulted in biased conclusions about criminal behavior. Further, with the progressive transition into computerized modes of data collection, it is important to test the extent to which computer-assisted modes affect participants' reports of offending behavior in comparison to the traditional paper-and-pencil questionnaires.

In order to assess the impact of modes of administration and modes of data collection on SRO, we conducted a two-experiment replication study with two independent samples from different cultural

backgrounds (Portugal and U.S.). These experiments followed a 2 (modes of administration: interviewer-administered vs. self-administered) × 2 (modes of data collection: paper-and-pencil vs. computer interviews) factorial design in which participants were randomly assigned to one of the experimental conditions. Based on the findings in the literature about sensitive topics, we predicted that participants in the self-administered modes would report higher rates of offending behavior than participants in face-to-face interviews (Hypothesis 1); and that participants in computer-assisted modes of data collection would report higher rates of offending compared to participants assigned to the paper-and-pencil modes (Hypothesis 2).

Experiment 1: Method

Experiment 1: Participants

One hundred and eighty-one students from a large University in the North of Portugal, mostly female (90.6%, $n = 164$), aged between 18 and 50 years ($M = 20.57$, $SD = 3.66$), participated in this experiment in exchange for course credits.

Experiment 1: Design

The present study followed a 2 (modes of administration: interviewer-administered vs. self-administered) × 2 (modes of data collection: paper-and-pencil vs. computer-assisted) experimental design. The crossing of these manipulations resulted in four experimental conditions: paper-and-pencil interviewer-administered interviews (PAPI); computer-assisted interviewer-administered interviews (CAPI); paper-and-pencil self-administered questionnaires (SAQ); and computer-assisted self-administered questionnaires (CASI). Participants were randomly assigned to one of these survey methods and completed the same questionnaire.

Experiment 1: Instruments

Participants in this study completed a questionnaire composed of three main sections. First, we have included a section on socio-demographic information (e.g., sex, age, education, income, etc.). In the second section, participants were asked to complete questions about multiple sensitive behaviors, which included the offending behavior questionnaire. This included 15 questions on different types of deviant behavior (i.e., graffiti, vandalism, shoplifting, burglary, bicycle theft, car theft, illegal downloading, stealing

from a car, stealing from a person, carrying a weapon, robbery, group fight, assault, drug sales, and animal cruelty). Also, we have created four composite variables of offending: an overall offending variable (with all offending items with the exception of illegal downloading), and three levels of offending seriousness, that is minor offenses (i.e., shoplifting, carrying a weapon, vandalism, group fight, graffiti, and animal cruelty), property offenses (i.e., stealing from a person, bicycle theft, car theft, stealing from a car, burglary), and violent offenses (i.e., assault and robbery) (Siegmunt & Lukash, 2019). Both the socio-demographic section and the offending behavior questionnaire were drawn from the International Self-Report Delinquency 3 questionnaire (ISR3; Enzmann et al., 2018; Portuguese version by Martins et al., 2015). Behavioral questions followed the layout set by the ISR3 questionnaire, in which questions were asked referring to lifetime prevalence and, in case of positive responses, participants were referred to an open-ended follow-up question about past year incidence. Past-year offending prevalences were very low in this study and we focused our data analysis on lifetime offending.

In the third section of our questionnaire, we included measures of social desirability and participants' perceptions of privacy and anonymity. Social desirability was assessed using the Socially Desirable Response Set 5 (SDRS-5; Hays et al., 1989; Portuguese version by Pechorro et al., 2016). This is a five-item brief questionnaire (e.g., "I am always courteous even to people who are disagreeable"). Participants' perceptions of privacy and anonymity regarding their participation in this study were assessed using two ancillary questions ("I wish I could have taken the survey in a more private place" and "I am confident that the answers I gave in this survey will never be linked with my name", respectively) developed by Denniston et al. (2010). Independently of the experimental condition, all participants completed the third section of this questionnaire in a self-administered mode in order to reduce potential social desirability effects.

Experiment 1: Procedure

Participants were recruited through the platform of exchanging course credits for participation in psychological experiments. Further, the researcher made a presentation at the end of several classes in order to recruit more participants to participate in exchange for course credits. Participants enrolled in the experiment through a doodle calendar and met the researcher in a classroom. Students were randomly assigned to one of the four experimental conditions and completed the experiment individually in a classroom in the sole presence of the researcher. Ethical approval for this experiment was provided by the Portuguese university's Institutional Review Board. Data collection was carried out from March 2018 to May 2019.

In the classroom, the researcher obtained informed consents from the participants and explained that we were interested in studying how people responded to questionnaires about sensitive topics, that they would be answering questions on personal experiences such as offending, drug use, victimization, and sexual behavior, and that their participation in this experiment would take about 30 minutes. The researcher also stated that students' answers were anonymous and that their participation was confidential and voluntary. Respondents who were interested in participating in the experiment signed the informed consent, which was archived next to others in order to ensure the anonymity of participants.

Students were then randomly assigned to one of the four possible experimental conditions (i.e., PAPI, CAPI, SAQ, and CASI). In the personal interview conditions, the interviewer read the questions appearing either on the questionnaire (i.e., PAPI) or on a computer screen (i.e., CAPI) to the participants, and the interviewer ticked/entered the response provided by the participants. Interviews were carried out by five researchers (three females) which were randomly distributed to the participants. In the self-administered conditions, after providing the instructions, the researcher would step back and the exact same questions appeared either on a questionnaire (i.e., SAQ) or on a computer screen (i.e., CASI), and participants completed the survey on their own. The computer-assisted conditions were carried out using Qualtrics software with the same questions as in the paper-and-pencil conditions.

Data analysis was developed using descriptive statistics, logistic regression models to test the impact of modes of administration and modes of data collection on individual items, and negative binomial regression models to test the impact of mode effects on composite variables (i.e., offending variety, etc.). Taking into consideration that our hypotheses suggested relationships in one specific direction (e.g., higher reports of offending in self-administered conditions), all statistical analyses were carried out using one-tailed tests. In experiment 1, the socio-demographic characteristics of participants did not differ between the experimental conditions. However, experiment 2 showed differences in participants' sex and age between the experimental conditions. Therefore, we have included participants' sex and age as covariates in all our regression models; prevalences and means shown in tables throughout this article are estimated marginal means after controlling for the effect of age and sex. All statistical analyses were carried out using SPSS software.

Experiment 1: Results

Experiment 1: Descriptive analysis

Participants in this study were randomly assigned either to a face-to-face interview or to a self-administered survey condition, as well as either to a paper-and-pencil or to a computer-assisted mode of data collection. As illustrated in Table 11, the random allocation of participants within these experimental manipulations resulted in similar demographic characteristics. No statistically significant differences were found between these manipulations for participants' age and sex, interviewers' sex, economic status, and university class year. Further, the manipulation of both modes of administration and modes of data collection did not cause any significant effect on social desirability (Table 11). On the other hand, the ancillary question about perceived privacy showed that, despite responding in similar environmental conditions, a larger prevalence of respondents in computer-assisted modes, compared to participants in paper-and-pencil modes, wished that they had taken this survey in a more private place, and this difference was marginally significant. Further, participants responding to the sensitive questions in face-to-face interviews reported more confidence about the anonymity of this study than participants in self-administered modes; this difference was nearly significant.

Regarding general descriptive statistics of offending, 39.8% of participants ($n = 72$) reported committing at least one type of offense during their life-course. Regarding offending variety, the present sample showed a mean number of types of offending of 0.80 ($SD = 1.24$, $min = 0$, $max = 5$). Male participants reported higher offending variety ($M = 1.53$, $SD = 0.48$) than females ($M = 0.73$, $SD = 0.09$). These differences were statistically significant (Incidence Rate Ratio [IRR] = 2.11, $\chi^2(1) = 4.97$, $p = .026$). Age was positively correlated with offending variety ($r_s = .15$, $p = .048$). Regarding the individual types of offenses, no participants in this study reported burglary or car theft, while only one reported bicycle theft, robbery, or assault. Therefore, in the tables for experiment 1, we will not be illustrating these offenses.

Experiment 1: Modes of administration (Interview vs. Survey)

Table 12 illustrates the effect of modes of administration on the prevalence of offending behavior. Results for overall offending prevalence show that 35.3% of participants in the face-to-face interview condition reported at least one type of offending behavior during their lifetime, compared to a total of 44.7% prevalence of offenders in self-administered surveys. However, despite the almost 10 percentage point difference between the two groups, the results were not statistically significant (OR = 1.48, $\chi^2(1) = 1.50$, $p = .110$). Considering an item-by-item analysis, participants reported a higher prevalence of

offending in the survey mode compared to face-to-face interviews in seven out of the total 10 offending questions. These differences reached statistical significance for group fight, graffiti, and vandalism, and were nearly significant ($p < .1$) for shoplifting. Regarding the three offending questions with higher prevalence estimates in the interview mode, only the illegal downloading reached statistical significance.

Modes of administration impacted the overall variety of offending (Table 12), since the incidence rate of offending in survey modes of administration was significantly higher than in interview modes (IRR = 1.69, $\chi^2(1) = 5.07$, $p = .012$). These results were also found to be statistically significant for minor offenses (IRR = 1.81, $\chi^2(1) = 5.67$, $p = .009$). Regarding property offenses, participants in survey modes reported a higher rate of offending, but with a very low mean incidence ($M(\text{interview}) = 0.05$ vs. $M(\text{survey}) = 0.09$) and not reaching statistical significance. Finally, participants in this experiment reported zero prevalence of violent offenses.

Experiment 1: Modes of data collection (Paper-and-pencil vs. Computer-assisted)

The manipulation of modes of data collection showed a statistically significant impact on the prevalence of overall offending (Table 12), where a larger proportion of participants reported offending behavior in paper-and-pencil modes than in computer-assisted modes (OR = 1.90, $\chi^2(1) = 3.99$, $p = .023$). Item-by-item analysis showed that, out of the 10 items, eight offending questions presented higher reports in paper-and-pencil modes of data collection, though only shoplifting reached marginal statistical significance. The remaining two offending items (i.e., group fight and vandalism) favored computer-assisted modes, one of which was nearly statistically significant.

Table 12 also illustrates the effects of modes of data collection on the composite variables of offending. Findings showed consistently higher scores of offending variety in paper-and-pencil conditions, though none reached statistical significance.

Table 11*Demographic characteristics by experimental manipulations (experiment 1)*

	Modes of administration				Modes of data collection			
	Interview	Survey	Test	p	Computer	Paper	Test	p
No. of cases	100	81			87	94		
Age (M [SD])	20.27 (2.24)	20.96 (4.93)	$t_{(170)} = -1.23$.221	20.64 (4.17)	20.49 (3.09)	$t_{(170)} = 0.27$.790
Sex (%)								
Female	91.0	90.1	$\chi^2_{(1)} = 0.04$.841	92.0	89.4	$\chi^2_{(1)} = 0.36$.550
Interviewers' sex (%)								
Female	41.4	50.0	$\chi^2_{(1)} = 1.32$.251	47.7	43.0	$\chi^2_{(1)} = 0.39$.531
Economic status (%)								
Worse off	1.0	3.7	$\chi^2_{(2)} = 4.03$.134	3.4	1.1	$\chi^2_{(2)} = 1.32$.517
Equal	94.0	85.2			89.7	90.4		
Better off	5.0	11.1			6.9	8.5		
University grade (%)								
1 st year	30.3	30.9	$\chi^2_{(3)} = 1.97$.579	35.6	25.8	$\chi^2_{(3)} = 4.22$.238
2 nd year	22.2	17.3			17.2	22.6		
3 rd year	32.3	40.7			31.0	40.9		
4 th year	15.2	11.1			16.1	10.8		
Social desirability								
SDRS-5 (M [SD])	1.41 (1.88)	1.33 (1.84)	IRR = 0.94	.389	1.33 (1.77)	1.42 (1.95)	IRR = 1.07	.369
Privacy (%)								
"I wish I could have taken the survey in a more private place."								
Strongly agree/Agree	6.1	4.0	OR = 0.65	.276	8.0	2.3	OR = 0.28	.058
Anonymity (%)								
"I am confident that the answers I gave in this survey will never be linked with my name."								
Strongly agree/Agree	91.8	84.2	OR = 0.48	.065	87.6	89.5	OR = 1.22	.342

Note. The statistical tests are negative binomial regression models for social desirability, and logistic regression models for privacy and anonymity, both with participants' sex and age as covariates.

Table 12

Experiment 1: Prevalence of offending and variety by modes of administration (left) and by modes of data collection (right)

	Modes of administration				Modes of data collection			
	Interview (<i>n</i> = 100)	Survey (<i>n</i> = 81)	OR/IRR	<i>p</i>	Computer (<i>n</i> = 87)	Paper (<i>n</i> = 94)	OR/IRR	<i>p</i>
Prevalence (%)								
Offending (overall)	35.3	44.7	1.48	.110	31.9	47.1	1.90	.023
Illegal downloading	97.0	78.8	0.12	<.001	88.7	89.5	1.08	.435
Group fight	4.1	15.4	4.21	.007	10.1	7.8	0.75	.289
Graffiti	15.5	26.7	1.99	.037	17.2	23.5	1.48	.153
Carrying a weapon	8.2	6.9	0.83	.378	7.5	7.9	1.06	.461
Vandalism	2.0	7.8	4.20	.425	6.7	2.3	0.32	.087
Shoplifting	14.5	22.3	1.70	.094	13.5	22.3	1.84	.068
Stealing from person	4.4	7.8	1.83	.169	3.8	8.0	2.18	.116
Animal cruelty	0.2	0.9	1.14	.115	0.2	0.8	3.22	.158
Stealing from a car	0	1.8	1.39	.499	0.7	0.8	0.98	.493
Drug sales	6.5	3.6	0.54	.193	4.6	5.9	1.29	.347
Variety (<i>M</i> [<i>SD</i>])								
Offending (overall)	0.59 (0.99)	1.01 (1.48)	1.69	.012	0.71 (1.10)	0.84 (1.31)	1.18	.235
Minor offenses	0.47 (0.84)	0.84 (1.30)	1.81	.009	0.58 (0.96)	0.68 (1.13)	1.19	.244
Property offenses	0.05 (0.23)	0.09 (0.32)	1.78	.156	0.05 (0.22)	0.08 (.31)	1.56	.219

Note. The statistical tests are logistic regression models for prevalence (i.e., OR) and negative binomial regression models for variety (i.e., IRR), both with participants' sex and age as covariates.

Experiment 1: Interaction effects

The negative binomial regression models testing the interaction effect of the two experimental manipulations (with participants' sex and age as covariates) on reports of offending variety is illustrated in Figure 6. Findings showed a nearly significant interaction effect for both overall variety of offending (IRR = 0.46, $\chi^2(1) = 2.67$, $p = .051$) and minor offenses (IRR = 0.46, $\chi^2(1) = 2.34$, $p = .063$). The interaction effect reached statistical significance for property offenses (IRR = 0.10, $\chi^2(1) = 2.81$, $p = .047$). These interaction findings showed that, within paper-and-pencil modes of data collection, reports of offending are only slightly higher in survey modes of administration. However, within computer-assisted modes of data collection, participants' reports of offending behavior in self-administered surveys are significantly higher than reports in face-to-face interviews.

Experiment 1: Discussion

In this experiment, we have tested the effects of modes of administration (i.e., face-to-face interview vs. self-administered survey) and modes of data collection (i.e., paper-and-pencil questionnaire vs. computer-assisted surveys) on SRO. The present findings showed that self-administration of the offending questionnaire resulted in an increased odds of disclosing offending behavior compared to the traditional face-to-face interviews. Regarding the manipulation of modes of data collection, the present findings showed an increased odds of reporting offending behavior in paper-and-pencil conditions compared to computer-assisted conditions. Further, the present study showed evidence of an interaction effect, where the benefits of self-administration over face-to-face interviews were only statistically significant in computer-assisted modes of data collection.

Experiment 1: Modes of administration (Interview vs. Survey)

In the present experiment, participants in self-administered conditions were more likely to report offending behavior than participants in face-to-face interviews. This effect was found for offending variety. Results showed that participants who were asked to complete the survey in a self-administered mode had a 69% increase in the rate of disclosing offending behavior compared to participants in interviewer-administered conditions. This result is in line with our first hypothesis and with the literature on sensitive questions, which showed that self-administered questionnaires yield higher estimates of sensitive behavior (e.g., Richman et al., 1999; Tourangeau & Yan, 2007).

However, contrary to the literature on modes of administration, the finding of mode effects in Experiment 1 was statistically significant despite the absence of differences in participants' social desirability or perception of privacy and anonymity. Social desirability was only slightly higher in face-to-face interviews, as was the wish to have taken the survey in a more private place (with no statistical significance). As for the participants' perception of anonymity while completing the questionnaire, respondents in face-to-face conditions reported higher confidence that their names would never be linked to their answers. Therefore, it seems that the benefits of self-administration in improving rates of disclosing offending behavior in this study go beyond the factors of social desirability, anonymity, and privacy.

On a different note, the question about illegal downloading was the only offending item that showed much higher reports in face-to-face interviews than in self-administered conditions. According to the results from Gomes et al. (2021), illegal downloading is a very low sensitivity topic, which would lead us to expect no mode effects. However, the present findings showed a strong effect in the opposite direction. One potential explanation for this is that, because illegal downloading is such a common practice among Portuguese young adults, the more socially desirable answer might be to admit that behavior in interview conditions. If the participant's perception about the expectations of the interviewer is that everybody has done it, then denying having ever downloaded music or films from the internet might be perceived as the most threatening answer. In other words, participants might feel more comfortable lying about having never done this behavior in the self-administered condition.

Experiment 1: Modes of data collection (Paper-and-pencil vs. Computer-assisted)

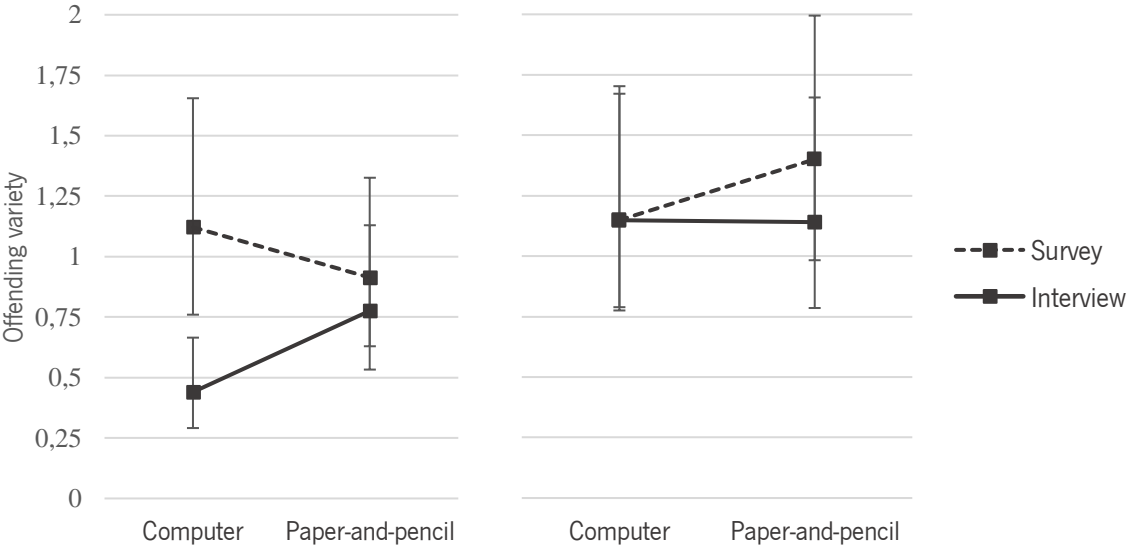
Contrary to our second hypothesis, modes of data collection caused an increased odds of reporting offending behavior by participants in the paper-and-pencil conditions compared to computer-assisted conditions. However, despite the consistently higher rates of offending disclosure in paper-and-pencil conditions, this effect was only statistically significant for the prevalence of offending. The present finding of higher reporting rates of offending in paper-and-pencil modes is contrary to the main body of evidence from the research on sensitive topics (e.g., Gnambs & Kaspar, 2015; Richman et al., 1999), as well as from studies including offending questions (Gomes et al., 2019). However, multiple studies have found similar results (e.g., Beebe et al., 1998; Knapp & Kirk, 2003), in which some questions result in higher rates of disclosure in paper-and-pencil modes, others in computerized modes, and others even showing no evidence of mode effects. This adds to the already inconsistent body of knowledge regarding

the effects of modes of data collection on participants' willingness to provide truthful answers, and more research on the moderators of this relationship is needed.

In the present experiment, participants reported very similar social desirability in both paper-and-pencil and computer-assisted modes of data collection. Also, no differences were found for participants' perceptions about anonymity, in which participants in computerized modes are just slightly less confident of the anonymity of this study. However, results for the respondents' perceptions about privacy showed that a higher proportion of participants in computer-assisted conditions wished that they had completed the survey in a more private place. This result is consistent with findings in the study of Denniston et al. (2010) that found less perceived privacy in computer-assisted modes compared to traditional paper-and-pencil modes.

Figure 6

Interaction effects of modes of administration and modes of data collection on lifetime offending variety (Experiment 1 on the left, Experiment 2 on the right; error bars are 90 percent confidence intervals)



Experiment 1: Interaction effects

The present findings revealed an interaction effect between modes of administration and modes of data collection in the reporting of offenses. Paper-and-pencil questionnaires resulted in very similar SRO scores for face-to-face interviews and self-administered surveys. On the other hand, respondents in computer-assisted modes reported much higher offending in self-administered conditions than

participants in face-to-face interviews (see Figure 6). In sum, despite participants in self-administered conditions reporting higher rates of offending throughout the two modes of data collection, the benefits of self-administration in increasing the willingness to report offending behavior was only statistically significant in the computer-assisted modes of administration. The interaction effect between these two aspects of the administration of questions needs more research.

Experiment 2: Method

Experiment 2 was carried out in the U.S., with students from a large university in central Florida. This experiment is a replication study, which allowed us to test whether the findings from experiment 1 would also be found in a different cultural context. Ethical approval for experiment 2 was provided by the Institutional Review Board of the U.S. University.

Experiment 2: Participants

One hundred and fifty-four students from a large University in central Florida, 63.0% female ($n = 97$), aged between 17 and 29 years ($M = 19.27$, $SD = 1.52$), participated in this experiment in exchange for course credits.

Experiment 2: Design, questionnaire, and procedure

In experiment 2, we used the English version of the exact same questionnaire as in experiment 1. Data collection was carried out from July 2019 to November 2019. Interviews were carried out by four researchers (three females).

Experiment 2: Results

Experiment 2: Descriptive analysis

Descriptive results for experiment 2 are illustrated in Table 13. Random assignment of participants resulted in statistically significant differences for participants' age and sex throughout the experimental manipulations. Participants in the survey conditions ($M = 19.01$, $SD = 1.24$) were significantly younger than participants in the face-to-face interviews ($M = 19.55$, $SD = 1.73$; $t(151) = 2.20$, $p = .030$). Further, the proportion of female participants (84.4%) in the computer-assisted modes

of data collection was significantly higher than in the paper-and-pencil conditions (41.6%; $\chi^2(1) = 30.33$, $p < .001$). No significant differences were found between the experimental conditions for the remaining socio-demographic characteristics (i.e., interviewers' sex, economic status, and university grade). In order to control for these differences in age and sex of participants, all models were tested using age and sex as covariates.

The experimental manipulations of modes of administration and modes of data collection showed no significant differences in social desirability and participants' perceived privacy (Table 13). As for the ancillary question regarding perceived anonymity, respondents in paper-and-pencil conditions reported higher confidence in the anonymity of their participation than did participants in computer-assisted modes (OR = 2.51, $\chi^2(1) = 3.19$, $p = .037$). As for the manipulation of modes of administration, despite no statistically significant differences being found, participants in face-to-face conditions reported higher perceived anonymity compared to participants in self-administered conditions.

Overall, 61% of participants ($n = 94$) in experiment 2 reported at least one type of offense. As for offending variety, the present sample showed a mean of offending of 1.27 ($SD = 1.48$, $min = 0$, $max = 9$). Male participants reported slightly higher offending variety ($M = 1.39$, $SD = 0.24$) than females ($M = 1.21$, $SD = 0.17$), though these differences were not statistically significant. Age was not correlated with lifetime offending ($r_s = .13$, $p = .100$). As for the individual types of offenses, zero participants in this study reported bicycle theft and car theft, while only one participant reported robbery. Therefore, in the tables for experiment 2, these offenses will not be illustrated.

Experiment 2: Modes of administration (Interview vs. Survey)

Results for the manipulation of modes of administration in experiment 2 are shown in Table 14. In this experiment, reports of prevalence of offending behavior were very similar throughout the two modes of administration. In the face-to-face interview conditions, 62.8% of participants reported committing at least one type of offending behavior compared to 60.3% of offenders in the self-administered survey condition. Item-wise analysis showed that, out of the total 10 offending questions, four items favored higher prevalence scores in survey conditions, one item showed the exact same mean prevalence in the two conditions, and five slightly favored higher reports in the face-to-face interviews, though none of these differences reached statistical significance.

Regarding the variety of offending, findings showed higher SRO scores in survey modes of administration than in face-to-face interviews (Table 14). However, these differences did not reach statistical significance for any of the composite variables considered.

Table 13*Demographic characteristics by experimental manipulations (experiment 2)*

	Modes of administration				Modes of data collection			
	Interview	Survey	Test	<i>p</i>	Computer	Paper	Test	<i>p</i>
No. of cases	75	79			77	77		
Age (M [SD])	19.55 (1.73)	19.01 (1.24)	$t_{(151)} = 2.20$.030	19.21 (1.66)	19.34 (1.38)	$t_{(151)} = -0.54$.587
Sex (%)								
Female	43.3	56.7	$\chi^2_{(1)} = 3.06$.080	84.4	41.6	$\chi^2_{(1)} = 30.33$	<.001
Interviewers' sex (%)								
Female	49.3	44.3	$\chi^2_{(1)} = 0.39$.532	50.6	42.9	$\chi^2_{(1)} = 0.94$.333
Economic status (%)								
Worse off	1.3	1.3	$\chi^2_{(2)} = 0.01$.999	1.3	1.3	$\chi^2_{(2)} = 0.01$.995
Equal	57.3	57.7			57.1	57.9		
Better off	41.3	41.0			41.6	40.8		
University grade (%)								
Freshman	33.3	44.9	$\chi^2_{(4)} = 4.17$.384	42.9	35.5	$\chi^2_{(4)} = 3.31$.507
Sophomore	25.3	23.1			26.0	22.4		
Junior	21.3	19.2			16.9	23.7		
Senior	20.0	11.5			41.7	58.3		
Non-degree	0	1.3			1.3	0		
Social desirability								
SDRS-5 (M [SD])	1.09 (1.52)	1.16 (1.61)	IRR = 1.07	.388	1.03 (1.55)	1.23 (1.79)	IRR = 1.19	.251
Privacy (%)								
"I wish I could have taken the survey in a more private place."								
Strongly agree/Agree	4.9	2.6	OR = 0.52	.232	3.2	4.1	OR = 1.28	.391
Anonymity (%)								
"I am confident that the answers I gave in this survey will never be linked with my name."								
Strongly agree/Agree	88.3	81.7	OR = 0.59	.131	79.2	90.5	OR = 2.51	.037

Note. The statistical tests are negative binomial regression models for social desirability, and logistic regression models for privacy and anonymity, both with participants' sex and age as covariates.

Table 14

Experiment 2: Prevalence of offending and variety by modes of administration (left) and by modes of data collection (right)

	Modes of administration				Modes of data collection			
	Interview (<i>n</i> = 75)	Survey (<i>n</i> = 79)	OR/IRR	<i>p</i>	Computer (<i>n</i> = 77)	Paper (<i>n</i> = 77)	OR/IRR	<i>p</i>
Prevalence (%)								
Offending (overall)	62.8	60.3	0.90	.381	66.3	56.7	0.67	.148
Illegal downloading	70.2	77.4	1.46	.159	76.4	71.3	0.77	.261
Group fight	4.7	5.3	1.13	.427	5.3	4.7	0.89	.435
Graffiti	11.1	8.3	0.73	.281	8.0	11.4	1.48	.241
Carrying a weapon	27.5	23.4	0.81	.287	27.1	23.7	0.84	.336
Vandalism	6.2	4.6	0.74	.331	2.5	8.5	3.63	.044
Shoplifting	27.1	36.7	1.56	.110	30.5	33.4	1.14	.367
Stealing from person	14.4	20.0	1.49	.188	15.9	18.6	1.21	.344
Animal cruelty	4.1	3.7	0.89	.438	4.0	3.8	0.94	.475
Burglary	1.2	1.2	0.97	.491	1.8	0.8	0.45	.289
Drug sales	12.9	14.5	1.15	.388	13.9	13.6	0.98	.481
Variety (<i>M</i> [<i>SD</i>])								
Offending (overall)	1.14 (1.58)	1.28 (1.75)	1.30	.295	1.15 (1.63)	1.27 (1.77)	1.11	.329
Minor offenses	0.85 (1.27)	0.87 (1.30)	1.03	.459	0.81 (1.26)	0.91 (1.37)	1.12	.334
Property offenses	0.17 (0.46)	0.24 (0.55)	1.37	.218	0.19 (0.49)	0.22 (0.54)	1.17	.349

Note. The statistical tests are logistic regression models for prevalence (i.e., OR) and negative binomial regression models for variety (i.e., IRR), both with participants' sex and age as covariates.

Experiment 2: Modes of data collection (Paper-and-pencil vs. Computer-assisted)

In regard to the manipulation of the modes of data collection (Table 14), the prevalence of overall offending was higher in computer (66.3%) than in paper-and-pencil (56.7%) modes, but with no statistical significance. Item-by-item analysis showed that, out of the 10 offending questions, four favored higher reports in paper-and-pencil modes of data collection, of which vandalism reached statistical significance ($OR = 3.63, \chi^2(1) = 2.90, p = .044$). The remaining six offending items favored higher reports in computer-assisted modes, none of which reached statistical significance.

The analysis for the effects of modes of data collection on offending variety showed higher offending scores in paper-and-pencil conditions compared to computer-assisted modes of data collection, though results were very similar and no statistically significant differences were found.

Experiment 2: Interaction effects

In the present experiment, results showed no evidence of an interaction effect between the two manipulated variables (i.e., modes of administration and modes of data collection) on SRO (Figure 6). While the rates of SRO were higher in survey conditions compared to face-to-face interviews, when using computerized modes, the difference was non-existent. On the other hand, when using paper-and-pencil modes of data collection, the benefit of self-administration over interviewer-administered modes is obvious, though statistically non-significant.

Experiment 2: Discussion

Experiment 2 is a replication study, in which we tried to reproduce the findings from the original experiment 1 with an independent sample from a different cultural background. In this experiment, we have followed the exact same procedures from the previous experiment and tested the effects of modes of administration (i.e., face-to-face interviews vs. self-administered surveys) and modes of data collection (i.e., paper-and-pencil questionnaires vs. computer-assisted surveys) on SRO. In experiment 2, despite the consistently higher rates of offending behavior in self-administered modes of administration, no statistically significant effects were found. Similarly, the present findings showed consistently higher reports of offending in paper-and-pencil modes of data collection compared to computer-assisted modes (with exception of overall prevalence), but with no statistical significance.

Experiment 2: Modes of administration (Interview vs. Survey)

In experiment 2, findings showed consistently higher reports of offending in self-administered conditions compared to face-to-face interviews. The only exceptions were the results for offending prevalence, which showed very similar results throughout the two methods, only slightly favoring face-to-face interviews. However, the benefit of self-administration on the participants' willingness to report higher rates of offending did not translate into statistically significant mode effects. This finding may be an indicator of the stability of the participants' reports of criminal behavior throughout different modes of administration. However, despite the lack of statistically significant results, the consistently higher disclosure of offending behavior in self-administered modes may indicate a pattern of response bias, which is consistent with our first hypothesis and with the literature on sensitive topics (e.g., Richman et al., 1999; Tourangeau & Yan, 2007).

Further, in contrast to the available evidence in the literature, the manipulation of modes of administration failed to reduce social desirability and increase participants' perceptions of privacy and anonymity in self-administered modes of administration. In the particular case of perceived anonymity, participants actually reported higher confidence in the anonymity assurances in the face-to-face interview conditions than in the self-administered modes (this difference was non-significant). This slight increase in the confidence of participants in the face-to-face interviews may explain the lack of significant effects favoring higher reports of offending caused by self-administration of the questionnaire. More research is needed to understand what aspects of modes of administration might affect participants' confidence in the study assurances about anonymity and confidentiality.

Experiment 2: Modes of data collection (Paper-and-pencil vs. Computer-assisted)

In regard to the manipulation of modes of data collection, despite the statistically non-significant effects, findings showed consistently higher reports of the variety of offending in paper-and-pencil modes. However, the prevalence of offending slightly favored higher reports in computer-assisted modes of data collection, though with no statistically significant results. In all comparisons, the only one that reached statistical significance was the prevalence of vandalism, in which the odds of reporting vandalism in paper-and-pencil modes was 3.62 times higher than in computer-assisted modes. Present findings of similar results slightly favoring higher reports in paper-and-pencil modes is contrary to our second hypothesis, as well as the most recent reviews of the literature on sensitive topics (e.g., Gnamb & Kaspar, 2015; Gomes et al., 2019). This is typical of the inconsistent findings regarding the effect of the transition from paper-

and-pencil to computerized modes of data collection. More research is needed to fully understand in which conditions computer-assisted modes might result in similar, higher, and sometimes lower reports of sensitive behavior.

The manipulation of modes of data collection also showed null effects on social desirability and participants' perceived privacy. However, a significant mode effect was observed on perceived anonymity. Participants in the paper-and-pencil conditions reported higher confidence that the answers would not be linked to their names, despite all conditions being anonymous. Some previous research has also found similar findings (e.g., Denniston et al., 2010), though others suggested that computer-assisted modes would increase participants' confidence about the study's anonymity and privacy and, in turn, result in an increased willingness of participants to provide truthful responses (e.g., Buchanan, 2000; Trau et al., 2013). This was not found in the present experiment, and more research on factors that affect the participants' confidence in the study is needed.

Experiment 2: Interaction effects

In the present experiment, we found no statistically significant evidence for the main effects of either modes of administration or modes of data collection on SRO. Similarly, no evidence of an interaction effect between these two manipulated variables was found. The small benefit of self-administration in increasing the disclosure of offending behavior, despite being nonsignificant, was more noticeable in the paper-and-pencil modes of data collection compared to computer-assisted modes.

General discussion

Self-reports are the most widely used measurement method in the study of offending behavior. Subject areas such as the study of the causes of delinquent behavior are heavily reliant on this methodology, making conclusions about delinquent behavior limited by the measurement technique. However, the lack of methodological research on SRO generates doubt about the quality of self-report measures, as well as the best ways to administer questions about offending behavior. This article provides evidence from two methodological experiments. Experiment 1 was conducted with undergraduate students from a Portuguese University and experiment 2 was a replication study with a sample of undergraduate students from a University in central Florida. In these experiments, we have tested the effects of modes of administration (i.e., face-to-face interviews vs. self-administered surveys) and modes of data collection (i.e., paper-and-pencil questionnaires vs. computer-assisted surveys) on SRO. In this

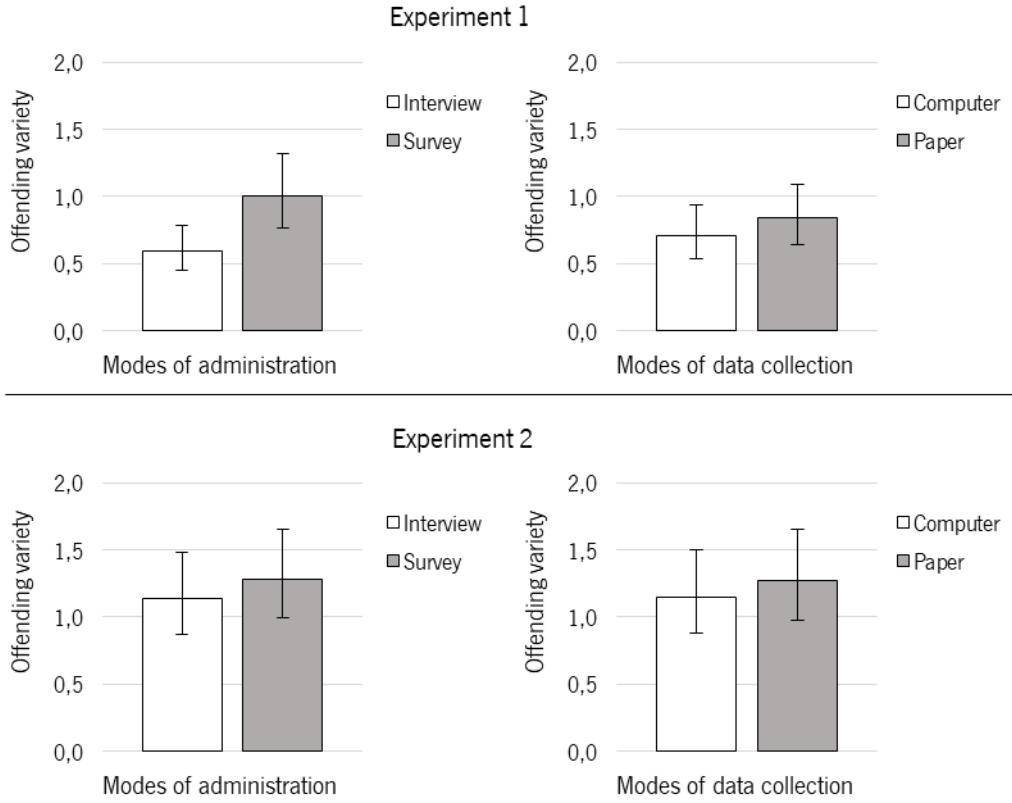
2x2 factorial design experiment, participants were randomly assigned to one of the four experimental conditions and were asked to disclose whether they have committed offending behavior during their lifetime.

Offending behavior is a highly sensitive topic that generates concern about socially desirable answers, poses the threat of responses being disclosed to other people outside of the study, or even fear of legal repercussions. Therefore, taking into consideration the evidence available in the literature on sensitive questions, self-administration of offending questionnaires is expected to result in higher rates of self-disclosed offending behavior compared to interviewer-administered conditions, where participants are requested to disclose their offending practices to a third person. An experimental approach is required to clearly demonstrate the impact of modes of administration on collecting such sensitive information (Tourangeau & Yan, 2007). However, out of the 21 methodological experiments on SRO reviewed by Gomes et al. (2019), only three studies compared offending results obtained by face-to-face interviews and self-administered surveys. Contrary to the evidence from the sensitive questions literature, these studies failed to find evidence of the benefits of self-administration on SRO. However, we must consider that two of these studies were carried out more than 40 years ago, and the third, most recent, study (Potdar & Koenig, 2005) was carried out to test mode effects on reports of risky behavior. Therefore, reliance on these findings to conclude that SRO are not affected by modes of administration may be misleading, and the question of what is the best practice to measure SRO still persists. In the present study, we aimed to provide evidence regarding the best practices of administering questions on offending, in order to improve the quality of SRO data.

In line with our initial hypothesis, the results from both experiments showed consistently higher rates of reported offending in self-administered conditions, compared with face-to-face interviews (see Figure 7 for a summary of findings). However, the increased odds of reporting offending behavior in self-administered surveys were only found to be statistically significant in experiment 1. Results in experiment 2 followed the same pattern of consistently higher disclosure of offending in the self-administered mode, though not reaching statistical significance. The evidence for the presence of mode effects found in these experiments is in line with the general literature on sensitive questions (e.g., Tourangeau & Yan, 2007; Richman et al., 1999). Requesting someone to disclose sensitive behavior to a third person, compared to completing a survey on their own, is expected to increase social desirability effects and, thus, influence participants' willingness to disclose embarrassing and criminal behavior (Bradburn et al., 1979; Tourangeau et al., 2000).

Figure 7

The effect of modes of administration (left) and modes of data collection (right) on overall offending variety (Experiment 1 on the top, Experiment 2 on the bottom; error bars are 90 percent confidence intervals)



However, in both experiments, participants reported very similar levels of social desirability and perceived privacy and anonymity in both modes of administration. Further, participants in this study reported higher perceived anonymity in the face-to-face interviews than in self-administered surveys. This finding seems to be contradictory to the deliberate misreporting hypothesis (Bradburn et al., 1979; Tourangeau et al., 2000), where self-administration is expected to provide greater confidence in the study’s assurances about anonymity. One potential explanation for this finding may be linked to our sample. University students may be used to completing surveys and may be aware of the ethical issues involved in carrying out research and be confident that the researcher will treat their answers carefully. Nevertheless, despite similar social desirability, anonymity, and privacy throughout the manipulated modes of administration, our results were still able to detect the presence of mode effects, in which participants in self-administered modes reported higher rates of offending than participants in face-to-face

interviews. More research to understand the mechanism through which self-administration causes an increased rate of reporting sensitive behavior is needed.

As for the manipulation of modes of data collection, results contrasted with our second hypothesis. According to the literature, we hypothesized that computerized modes would elicit higher rates of reporting offending behavior. However, compared to computer-assisted modes, paper-and-pencil conditions resulted consistently in higher reports of offending throughout the two experiments, though generally with no statistically significant differences (see Figure 7). The absence of statistically significant findings suggests that reports of offending may not be affected by modes of data collection. However, these results are to be treated with caution, since the findings in the literature for the comparison between paper-and-pencil and computer-assisted modes of data collection are inconsistent. In our experiments, results are in line with some previous research (e.g., Beebe et al., 2008; Knapp & Kirk, 2003), but failed to find the benefits of computerized modes in increasing the likelihood of reporting sensitive behavior found in the reviews of the literature (Gnambs & Kaspar, 2015; Gomes et al., 2019; Richman et al., 1999). Further, modes of data collection did not affect social desirability in both experiments. As for perceived privacy and anonymity, participants generally favored paper-and-pencil conditions. Despite previous research reporting similar results (e.g., Denniston et al., 2010), the present findings are contrary to the studies that have suggested that computer-assisted modes increase participants' confidence in the study's anonymity (e.g., Trau et al., 2013).

Finally, the results of experiment 1 showed the presence of an interaction effect between modes of administration and modes of data collection, in which the benefits of self-administration are only noticeable when data are collected via computer-assisted modes. However, this interaction effect failed to be replicated in experiment 2, in which computer-assisted personal interviews and computer-assisted self-interviews yielded very similar results. One reason that might explain this differential effect in the two experiments may be the differential perception of anonymity and privacy. While students from the Portuguese University showed similar results in paper-and-pencil and computer-assisted modes of data collection, students from the American University were much less confident about the anonymity of their participation when in computer-assisted modes, which might have caused them to refrain from reporting offending behavior. The potential interaction effects between these two important aspects of data collection is underexplored in the literature, and future research is needed to understand these relations.

Limitations

Some limitations of this study need to be discussed. First, samples in both experiments consisted of University students. The prevalence of offending among University students is expected to be low, especially for more serious types of offenses. Therefore, very large samples are needed in order to obtain significant results, and our samples may have been too small. The low prevalence of offending may have limited our capacity to detect mode effects because, in many cases, participants did not commit these behaviors. Also, taking into consideration that the serious offenses, as well as the most recent offenses, are regarded as the most sensitive questions (Gomes et al., 2021), the low prevalence of these serious and violent types of offenses made it impossible to test whether mode effects are higher for them. Second, our samples were mostly composed of female participants. This sample characteristic may affect the generalizability of the present findings. Also, similar to the previous limitation, offending behavior is less prevalent within female participants, which may limit even more our ability to detect the impact of mode effects. However, the fact that we have detected evidence for the beneficial effects of self-administration in the reports of offending behavior in these experiments is a strong indication that SRO questionnaires are affected by mode effects.

Third, the random allocation of participants in experiment 2 resulted in differences in participants' characteristics (i.e., sex and age). In order to control for potential confounding effects in the comparisons between experimental conditions, all analyses in this study were carried out with participants' sex and age as covariates. Future studies should consider these limitations, and carry out similar experiments with younger participants, from multiple backgrounds, in order to provide a larger variability of the offending variable. This would allow us to test for mode effects on more recent and more serious types of offenses, as well as to test whether the benefits of modes of administration increase with more sensitive offending questions.

Conclusions

Findings from this study showed that SRO behaviors are affected by modes of administration. Asking questions about offending behavior in self-administered conditions results in increased odds of participants' disclosure of offending behavior when compared to face-to-face interviews. Therefore, researchers using questionnaires to assess SRO should consider using self-administered modes of administration in order to increase measurement accuracy. As for the effect of modes of data collection, results from this study show that asking questions using paper-and-pencil questionnaires or computer-

assisted surveys resulted in mainly similar results, slightly favoring higher reports in paper-and-pencil modes. Further, through the two experiments in this study, participants in paper-and-pencil conditions reported higher perceived anonymity compared to computer-assisted modes of data collection. More research on the impact of modes of data collection on SRO is needed, especially considering the gradual transition into more computerized methods and the added advantages of computer-assisted modes in reducing costs, human resources, and overcoming the limitations caused by illiteracy.

References

- Aquilino, W. S. (1994). Interview mode effects in surveys of drug and alcohol use: A field experiment. *Public Opinion Quarterly*, *58*(2), 210-240. <https://doi.org/10.1086/269419>
- Bates, S. C., & Cox, J. M. (2008). The impact of computer versus paper-pencil survey, and individual versus group administration, on self-reports of sensitive behaviors. *Computers in Human Behavior*, *24*(3), 903-916. <https://doi.org/10.1016/j.chb.2007.02.021>
- Beebe, T. J., Harrison, P. A., Mcrae, J. A., Anderson, R. E., & Fulkerson, J. A. (1998). An evaluation of computer-assisted self-interviews in a school setting. *Public Opinion Quarterly*, *62*(4), 623-632. <https://doi.org/10.1086/297863>
- Beebe, T. J., Harrison, P. A., Park, E., McRae, J. A., Jr., & Evans, J. (2006). The effects of data collection mode and disclosure on adolescent reporting of health behavior. *Social Science Computer Review*, *24*(4), 476-488. <https://doi.org/10.1177/0894439306288690>
- Bradburn, N. M., Sudman, S., Blair, E., Locander, W., Miles, C., Singer, E., & Stocking, C. (1979). *Improving interview method and questionnaire design: Response effects to threatening questions in survey research*. Jossey-Bass.
- Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design-for market research, political polls, and social and health questionnaires*. Jossey-Bass.
- Brener, N. D., Eaton, D. K., Kann, L., Grunbaum, J. A., Gross, L. A., Kyle, T. M., & Ross, J. G. (2006). The association of survey setting and mode with self-reported health risk behaviors among high school students. *Public Opinion Quarterly*, *70*(3), 354-374. <https://doi.org/10.1093/poq/nfl003>
- Buchanan, T. (2000). Potential of the Internet for personality research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 121-140). Academic Press. <https://doi.org/10.1016/B978-012099980-4/50006-X>
- Butler, S. F., Villapiano, A., & Malinow, A. (2009). The effect of computer-mediated administration on self-disclosure of problems on the Addiction Severity Index. *Journal of Addiction Medicine*, *3*(4), 194-203. <https://doi.org/10.1097/ADM.0b013e3181902844>
- Clark, J. P., & Tiffit, L. L. (1966). Polygraph and interview validation of self-reported deviant behavior. *American Sociological Review*, *31*(4), 516-523. <https://doi.org/10.2307/2090775>

- Cops, D., Boeck, A., & Pleysier, S. (2016). School vs. mail surveys: Disentangling selection and measurement effects in self-reported juvenile delinquency. *European Journal of Criminology*, *13*(1), 92-110. <https://doi.org/10.1177/1477370815608883>
- Denniston, M. M., Brener, N. D., Kann, L., Eaton, D. K., McManus, T., Kyle, T. M., Roberts, A. M., Flint, K. H., & Ross, J. G. (2010). Comparison of paper-and-pencil versus Web administration of the Youth Risk Behavior Survey (YRBS): Participation, data quality, and perceived privacy and anonymity. *Computers in Human Behavior*, *26*(5), 1054-1060. <https://doi.org/10.1016/j.chb.2010.03.006>
- Dodou, D., & de Winter, J. C. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, *36*, 487-495. <https://doi.org/10.1016/j.chb.2014.04.005>
- Druckman, J. N., Gilli, M., Klar, S., & Robison, J. (2015). Measuring drug and alcohol use among college student-athletes. *Social Science Quarterly*, *96*(2), 369-380. <https://doi.org/10.1111/ssqu.12135>
- Enzmann, D., Kivivuori, J., Marshall, I. H., Steketee, M., Hough, M., & Killias, M. (2018). *A global perspective on young people as offenders and victims: First results from the ISRD3 study*. Springer.
- Farrington, D. P. (1973). Self-reports of deviant behavior: Predictive and stable?. *Journal of Criminal Law and Criminology*, *64*(1), 99-110. <https://doi.org/10.2307/1142661>
- Gerdtz, M., Yap, C. Y., Daniel, C., Knott, J. C., Kelly, P., & Braitberg, G. (2020). Prevalence of illicit substance use among patients presenting to the emergency department with acute behavioural disturbance: Rapid point-of-care saliva screening. *Emergency Medicine Australasia*, *32*(3), 473-480. <https://doi.org/10.1111/1742-6723.13441>
- Giguère, K., Béhanzin, L., Guédou, F. A., Leblond, F. A., Goma-Matsétsé, E., Zannou, D. M., Affolabi, D., Kékè, R. K., Gangbo, F., Bachabi, M., & Alary, M. (2019). Biological validation of self-reported unprotected sex and comparison of underreporting over two different recall periods among female sex workers in Benin. *Open Forum Infectious Diseases*, *6*(2), 1-6. <https://doi.org/10.1093/ofid/ofz010>
- Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, *47*(4), 1237-1259. <https://doi.org/10.3758/s13428-014-0533-4>

- Gomes, H. S., Farrington, D. P., Krohn, M. D., & Maia, Â. (2021). *How sensitive are self-reports of offending?: The impact of recall periods on question sensitivity* [Manuscript submitted for publication]. School of Psychology, University of Minho.
- Gomes, H. S., Farrington, D. P., Maia, Â., & Krohn, M. D. (2019). Measurement bias in self-reports of offending: A systematic review of experiments. *Journal of Experimental Criminology*, *15*(3), 313-339. <https://doi.org/10.1007/s11292-019-09379-w>
- Gomes, H. S., Maia, Â., & Farrington, D. P. (2018). Measuring offending: Self-reports, official records, systematic observation and experimentation. *Crime Psychology Review*, *4*(1), 26-44. <https://doi.org/10.1080/23744006.2018.1475455>
- Hays, R. D., Hayashi, T., & Stewart, A. L. (1989). A five-item measure of socially desirable response set. *Educational and Psychological Measurement*, *49*(3), 629-636. <https://doi.org/10.1177/001316448904900315>
- Hindelang, M. J., Hirschi, T., & Weis, J. G. (1981). *Measuring delinquency*. Sage.
- Huizinga, D., & Elliott, D. S. (1986). Reassessing the reliability and validity of self-report delinquency measures. *Journal of Quantitative Criminology*, *2*(4), 293-327. <https://doi.org/10.1007/BF01064258>
- Jobe, J. B., Pratt, W. F., Tourangeau, R., Baldwin, A. K., & Rasinski, K. A. (1997). Effects of interview mode on sensitive questions in a fertility survey. In L. Lyberg , P. Biemer , M. Collins , E. de Leeuw , C. Dippo , N. Schwartz , & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 311-329). John Wiley & Sons. <https://doi.org/10.1002/9781118490013.ch13>
- Jolliffe, D., Farrington, D. P., Hawkins, J. D., Catalano, R. F., Hill, K. G., & Kosterman, R. (2003). Predictive, concurrent, prospective and retrospective validity of self-reported delinquency. *Criminal Behaviour and Mental Health*, *13*(3), 179-197. <https://doi.org/10.1002/cbm.541>
- Kabashi, S., Vindenes, V., Bryun, E. A., Koshkina, E. A., Nadezhdin, A. V., Tetenova, E. J., Kolgashkin, A. J., Petukhov, A. E., Perekhodov, S. N., Davydova, E. N., Gamboa, D., Hilberg, T., Lerdal, A., Nordby, G., Zhang, C., & Bogstrand, S. T. (2019). Harmful alcohol use among acutely ill hospitalized medical patients in Oslo and Moscow: A cross-sectional study. *Drug and Alcohol Dependence*, *204*, 107588. <https://doi.org/10.1016/j.drugalcdep.2019.107588>
- Kirtadze, I., Otiashvili, D., Tabatadze, M., Vardanashvili, I., Sturua, L., Zabransky, T., & Anthony, J. C. (2018). Republic of Georgia estimates for prevalence of drug use: Randomized response techniques suggest under-estimation. *Drug and alcohol dependence*, *187*, 300-304. <https://doi.org/10.1016/j.drugalcdep.2018.03.019>

- Knapp, H., & Kirk, S. A. (2003). Using pencil and paper, Internet and touch-tone phones for self-administered surveys: Does methodology matter? *Computers in Human Behavior, 19*(1), 117-134. [https://doi.org/10.1016/S0747-5632\(02\)00008-0](https://doi.org/10.1016/S0747-5632(02)00008-0)
- Krohn, M. D., Waldo, G. P., & Chiricos, T. G. (1974). Self-reported delinquency: A comparison of structured interviews and self-administered checklists. *Journal of Criminal Law and Criminology, 65*(4), 545–553. <https://doi.org/10.2307/1142528>
- Liber, A. C., & Warner, K. E. (2018). Has underreporting of cigarette consumption changed over time? Estimates derived from US National Health Surveillance Systems between 1965 and 2015. *American Journal of Epidemiology, 187*(1), 113-119. <https://doi.org/10.1093/aje/kwx196>
- Littlefield, A. K., Brown, J. L., DiClemente, R. J., Safonova, P., Sales, J. M., Rose, E. S., Belyakov, N., & Rassokhin, V. V. (2017). Phosphatidylethanol (PEth) as a biomarker of alcohol consumption in HIV-infected young Russian women: Comparison to self-report assessments of alcohol use. *AIDS and Behavior, 21*(7), 1938-1949. <https://doi.org/10.1007/s10461-017-1769-7>
- Lucia, S., Herrmann, L., & Killias, M. (2007). How important are interview methods and questionnaire designs in 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*(1), 39-64. <https://doi.org/10.1007/s11292-007-9025-1>
- Martins, P., Mendes, S., & Fernandez-Pacheco, G. (2015, September 2-5). *Cross-cultural adaptation and online administration of the Portuguese Version of ISRD3* [Paper presentation]. 15th Annual Conference of the European Society of Criminology, Porto, Portugal.
- Palamar, J. J., Salomone, A., & Keyes, K. M. (2021). Underreporting of drug use among electronic dance music party attendees. *Clinical Toxicology, 59*(3), 185-192. <https://doi.org/10.1080/15563650.2020.1785488>
- Pechorro, P., Barroso, R., Silva, I., Marôco, J., & Gonçalves, R. A. (2016). Propriedades psicométricas da Escala de Respostas Socialmente Desejáveis-5 (SDRS-5) em jovens institucionalizados [Psychometric properties of the Socially Desirable Response Set 5 (SDRS-5) with institutionalized youth]. *Psicologia, 30*(1), 29-36. <https://doi.org/10.17575/rpsicol.v30i1.1065>
- Piquero, A. R., Schubert, C. A., & Brame, R. (2014). Comparing official and self-report records of offending across gender and race/ethnicity in a longitudinal study of serious youthful offenders. *Journal of Research in Crime and Delinquency, 51*(4), 526-556. <https://doi.org/10.1177/0022427813520445>

- Potdar, R., & Koenig, M. A. (2005). Does audio-CASI improve reports of risky behavior? Evidence from a randomized field trial among young urban men in India. *Studies in Family Planning, 36*(2), 107-116. <https://doi.org/10.1111/j.1728-4465.2005.00048.x>
- Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*(5), 754-775. <https://doi.org/10.1037/0021-9010.84.5.754>
- Schober, S. E., Caces, M. F., Pergamit, M. R., & Branden, L. (1992). Effect of mode of administration on reporting of drug use in the National Longitudinal Survey. In C. F. Turner, J. T. Lessler, & J. C. Gfroerer (Eds.), *Survey measurement of drug use: Methodological studies* (pp. 267–276). National Institute on Drug Abuse.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93-105. <https://doi.org/10.1037/0003-066X.54.2.93>
- Schwarz, N., Strack, F., Hippler, H. J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology, 5*(3), 193-212. <https://doi.org/10.1002/acp.2350050304>
- Sudman, S., & Bradburn, N. M. (1974). *Response effects in surveys: A review and synthesis*. Aldine Publishing Company.
- Thornberry, T. P., & Krohn, M. D. (2000). The self-report method for measuring delinquency and crime. In D. Duffee (Ed.), *Criminal Justice* (pp. 33-84). National Institute of Justice.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-883. <https://doi.org/10.1037/0033-2909.133.5.859>
- Tourangeau, R., & Yan, T. (in press). Reporting issues in surveys of drug use. *Substance Use and Misuse*.
- Trau, R. N., Härtel, C. E., & Härtel, G. F. (2013). Reaching and hearing the invisible: Organizational research on invisible stigmatized groups via web surveys. *British Journal of Management, 24*(4), 532-541. <https://doi.org/10.1111/j.1467-8551.2012.00826.x>
- Turner, C. F., Lessler, J. T., Devore, J. W. (1992). Effects of mode of administration and wording on reporting of drug use. In C. F. Turner, J. T. Lessler, & J. C. Gfroerer (Eds.), *Survey measurement of drug use: Methodological studies* (pp. 177–219). National Institute on Drug Abuse.

- Vinikoor, M. J., Zyambo, Z., Muyoyeta, M., Chander, G., Saag, M. S., & Cropsey, K. (2018). Point-of-care urine ethyl glucuronide testing to detect alcohol use among HIV-hepatitis B virus coinfecting adults in Zambia. *AIDS and Behavior, 22*(7), 2334-2339. <https://doi.org/10.1007/s10461-018-2030-8>
- Wolter, F., & Laier, B. (2014). The effectiveness of the item count technique in eliciting valid answers to sensitive questions. An evaluation in the context of self-reported delinquency. *Survey Research Methods, 8*(3), 153-168. <https://doi.org/10.18148/srm/2014.v8i3.5819>

INTEGRATIVE DISCUSSION

The present dissertation provides comprehensive reviews of the main measurement techniques for offending behavior, as well as experimental evidence to improve the accuracy of self-reports of offending (SRO). In this final section, we present an integrative discussion of the main findings and implications of our studies, as well as a reflection on the need for future studies in order to improve crime assessment. With this dissertation, we aimed to provide evidence on the existing methods used to assess participants' deviant behavior in real-life settings. We hope this work will encourage more researchers to consider the use of observation techniques, especially within field experimental designs, to study the causes of offending. Further, taking into consideration our studies on the potential factors that can influence the accuracy of data collected using SRO, we hope that researchers will carry out more methodological studies in order to build on and establish the guidelines for best practices in collecting self-disclosed information about delinquent behavior. Improving measurement methods will allow us to carry out more rigorous studies and develop solid knowledge about the factors influencing offending.

Major contributions

“Measurement is the basis of all science” (Ferraro & LaGrange, 1987, p. 70). Without rigorous assessments of offending behavior, knowledge about the causes and correlates of offending is seriously compromised. Taking into consideration that offending behavior is inherently difficult to measure (Osgood et al., 2002), as well as the lack of methodological research on crime measures (Jolliffe & Farrington, 2014), research evidence in the field of offending behavior is impaired due to uncertainty about the quality of offending assessments. With this dissertation, we aimed to assess the state of the art of crime measurement, as well as to improve the accuracy of SRO. In order to achieve these objectives, we started by reviewing the main techniques applied in the measurement of offending behavior.

Measures of offending behavior

In Chapter I, we have carried out a review of the main measurement methods of offending behavior (Gomes et al., 2018). The literature reviewed in this chapter provided a discussion about the main issues in crime assessment, as well as reflections on the advantages and limitations of these techniques. In this chapter, we concluded that the study of offending behavior is mainly based on three methodological techniques: official records, observation, and SRO. The evidence reviewed in this study suggests that some measurement methods provide more valid estimates of offending than others. Further, different methods often result in inconsistent (and sometimes contradictory) assessments of

offending behavior (e.g., Kazemian & Farrington, 2005). This is a serious concern that can lead to ambiguous conclusions, compromising the quality of research findings, theory testing, and policy decisions. However, the qualities of each measurement technique must be considered with caution, and it is not this study's conclusion that offending behavior should be measured using a single methodology. Each measurement technique has its own advantages and limitations, so the assessments must be tailored to the specific research questions under study.

Assessing criminal behavior using official records, for example, offers the advantage of precise information, such as exact dates of occurrence. This characteristic of officially recorded information is very useful, especially in fields of study such as criminal careers and offending through the life course (Farrington et al., 2003). Furthermore, official records of criminal behavior are legitimated by the procedures of formal institutions and, presumably, are not contaminated by the researchers' own biases (Klein, 1987). On the other hand, data provided by official records seriously underreport the true amount of offending behavior (Theobald et al., 2014). A large proportion of offenses are never reported to the police and, from the reported offenses, only a fraction result in arrests or convictions (Farrington & Jolliffe, 2004; Pepper & Petrie, 2003). Additionally, and even more concerning, some researchers have described how official records provide deeply biased information (Farrington et al., 2007). Different countries, jurisdictions, agencies, and courts vary considerably in policy and practice (Klein, 1987). Different types of offenses differ in their likelihood of being reported to the police or resulting in a conviction (Huizinga & Elliott, 1986). These biases may result in misleading conclusions about the variables affecting offending behavior, limiting the utility of official records in the study of the causes of offending (Hindelang et al., 1981). On the other hand, because highly serious crimes, such as homicides, are more likely to be reported to the police, official records may provide more accurate assessments of more serious offenses (Maxfield & Babbie, 2009).

Observation methods, on the other hand, are very powerful techniques that can provide rigorous assessments of offending and deviant behaviors. Observation techniques offer reliable, precise, and accurate information both about the offense itself, as well as its timing, duration, frequency, and the overall behavior of the offender (Buckle & Farrington, 1984, 1994; McCall, 1984). However, observing offending behavior can be very challenging because people naturally try to conceal their illegal practices. For these reasons, observations tap only a small, less serious proportion of the offending phenomenon and are not very well suited to assess the total distribution and patterns of offending (Thornberry & Krohn, 2000). Nevertheless, frequently occurring offenses, especially those committed in public, are feasible to be assessed using observation techniques (Buckle & Farrington, 1984). Further, observation techniques

can be very useful to test hypotheses in the real world, especially when applied within naturalistic field experimental designs (see Chapter II). However, when it is impossible to “observe the behavior taking place, self-reports of delinquent and criminal behavior would be the nearest data source to the actual behavior” (Thornberry & Krohn, 2000, p. 34).

The introduction of the self-report technique was a very important methodological innovation that has completely revolutionized the study of offending behavior (Krohn et al., 2012). Data provided from SRO played a determining role in our understanding of the causes and correlates of delinquent behavior, its prevalence and trajectories throughout the life-course, and greatly influenced policymaking in the juvenile justice system, etc. (Thornberry & Krohn, 2000). However, multiple factors can impact the quality of data obtained through SRO. For example, the accuracy of SRO is limited by the respondents’ memories and willingness to disclose their own offenses (Farrington et al., 2007). In turn, the willingness to report offending behavior can be affected by a number of factors, such as the presence of an interviewer (Tourangeau et al., 2000). Self-administered surveys allow respondents to disclose offending behavior without the biasing presence of an interviewer, but raise other issues such as the requirement of literacy of participants (e.g., Gribble et al., 2000). In Chapter III (Gomes et al., 2019), we have reviewed the potential sources of bias in collecting offending data using surveys.

Despite the considerable skepticism about the ability of participants to disclose relevant information about their own offending behavior, many studies have demonstrated the validity and reliability of SRO (e.g., Jolliffe & Farrington, 2014; Piquero et al., 2014). Additionally, SRO are fairly easy to implement, and researchers can collect offending data from large samples with relatively few resources. Furthermore, self-reports allow the collection of information that goes beyond the offending behavior itself, such as socio-economic status, peer delinquency, motivations, potential risk and protective factors, etc. (Farrington, 2001; Krohn et al., 2012). For these reasons, SRO are very well suited to the study of the etiology of offending behavior (Clark & Tifft, 1966; Thornberry & Krohn, 2000).

Observation methods within field experiments

In Chapter II (Gomes et al., 2021a), we have considered the findings from the previous chapter regarding the advantages of observation assessments within field experiments and aimed to systematically review the field experimental evidence on the causes of offending. Briefly, naturalistic field experimental designs have high internal and external validity, make it possible to test cause-and-effect relationships, provide unambiguous conclusions, and permit rigorous observations of unaware participants in real-life contexts (Farrington, 1979; Harrison & List, 2004). Despite the benefits of this

powerful research design, only a few naturalistic field experiments have been carried out to study the causes of offending (Farrington et al., 2020; Franzen & Pointner, 2013; Gomes et al., 2018). However, several field experiments have been conducted by behavioral economists in the study of stealing and monetary dishonesty (Farrington et al., 2020). Therefore, in Chapter II (Gomes et al., 2021a) we aimed to systematically review field experiments studying stealing or monetary dishonesty and reported since the previous review of Farrington (1979).

The systematic review of field experiments in Chapter II (Gomes et al., 2021a) provides the reader with detailed information about the field experimental designs used by psychologists, criminologists, and behavioral economists to assess offending and dishonest behavior in the real world. Also, this review describes the field methodologies implemented to assess multiple types of deviant behaviors, which we grouped into fraudulent/ dishonest behavior (e.g., insurance fraud, overcharging, etc.), stealing (e.g., theft of coins or apparently 'lost' wallets, etc.), keeping money (e.g., acceptance of a bribe, keeping extra change, etc.), and shoplifting. Further, in this chapter (Gomes et al., 2021a), we showed how field experiments can provide relevant information for theory testing in the study of offending. As previously carried out by Farrington (1979), we coded the findings from the reviewed field experiments into *benefits for the perpetrator* (e.g., financial gains), *costs for the self* (e.g., likelihood of apprehension), and *costs for the other* (e.g., suffering of the victim), in order to test fundamental features of the subjective expected utility (SEU) theoretical framework (see Farrington & Knight, 1980). In sum, consistently with the SEU theory's predictions, the reviewed studies provide rigorous real-world observational evidence that offending and dishonest behaviors are more likely when the *costs for the self* are low, when the *costs for the victim* are low, and when the *benefits for the self* are high.

Self-reports of offending behavior

In the first chapter (Gomes et al., 2018), we found that SRO are the most widely used measurement methods in the study of the causes of offending behavior. However, despite the relevance of the self-report methodology in the study of the etiology of crime, little is known about the best practices of asking questions about offending behavior. In order to compile the accumulated knowledge on how to collect information on offending behavior using self-reports, we have carried out the systematic review presented in Chapter III (Gomes et al., 2019). In this study, we have carried out a systematic review of methodological experiments testing potential sources of bias in collecting offending data through the self-report methodology. The findings from the reviewed experiments were summarized by the calculation of effect sizes using meta-analytical techniques. A total of 21 methodological experiments were pooled (33

independent effect sizes), which provided evidence about 18 different measurement manipulations that we have grouped into modes of administration, procedures of data collection, and questionnaire design.

The systematic review in Chapter III confirmed that methodological research on SRO is very scarce and that most of the measurement biases tested in these studies lacked replication (Gomes et al., 2019). However, these studies provide relevant information that can be used as guidelines to improve the accuracy of SRO. In summary, the reviewed methodological experiments showed no evidence that modes of administration impact SRO. According to these studies, self-reports provided generally stable estimates of offending behavior, whether participants' reports were collected using interviewer-administered or self-administered interviews (Hindelang et al., 1981; Krohn et al., 1974; Potdar & Koenig, 2005), self-administered surveys with or without audio recordings (Potdar & Koenig, 2005; Trapl et al., 2013), in person or by the telephone (Knapp & Kirk, 2003). Further, the reviewed studies provided no evidence that SRO are impacted by whether the completion of the questionnaire is supervised by teachers or by research staff (Kivivuori et al., 2013; Walser & Killias, 2012), by different definitions of the reference period (i.e., "12 months" vs. "Since October 2003", Lucia et al., 2007), or by telescoping (Horney & Marshall, 1992).

On the other hand, the reviewed methodological experiments provided evidence that self-administered surveys using computers resulted in slightly higher rates of offending than those collected using paper-and-pencil questionnaires (Brenner et al., 2006; van de Looij-Jansen & de Wilde, 2008; Eaton et al., 2010). Further, we found evidence that: SRO may vary as a function of the response format, where higher rates of offending were found using a 7-option response format compared to dichotomous response options (Hamby et al., 2006); completing the survey in school environments results in higher rates of offending than surveys completed at home (Brenner et al., 2006); the use of bogus pipeline methods increased the odds of reporting offenses (Strang & Peterson, 2020); anonymous conditions (van de Looij-Jansen et al., 2006), as well as when participants' reports are not disclosed to third parties (Beebe et al., 2006) and when there is no in-person follow-up (King et al., 2012), increased the respondents' willingness to report offending behavior; casually dressed interviewers resulted in higher offending rates compared to conservative-looking interviewers (Krohn et al., 1974); and, finally, short questionnaires with fewer follow-up questions and a yes-no response pattern resulted in an increased odds of reporting offending behavior (Enzmann, 2013).

The findings reviewed in Chapter III (Gomes et al., 2019) provide relevant information regarding the best practices of collecting data using SRO. Studies using self-reports to assess information about delinquent behavior should consider these findings in order to improve the accuracy of assessments and,

as a consequence, the validity and reliability of the studies' conclusions about the variables that affect offending behavior. Furthermore, researchers using SRO in their studies should consider including detailed procedural information regarding the modes of administration, procedures of data collection, and questionnaire design. This information would be important for the interpretation of conclusions, as well as to further study the impact of these aspects on participants' reports. On a different note, Chapter III highlighted the scarcity of methodological research on SRO. Considering that SRO are the most used measurement method in the study of the causes of offending, this scarcity of methodological research seriously compromises the studies' conclusions. As a consequence, the findings reviewed in Chapter III must be considered with caution because in many cases the findings were reported several decades ago and, in most cases, the results were not replicated. All this is evidence that more research on the best practices of collecting self-reported information about offending behavior is needed.

The most surprising finding from Chapter III was the lack of evidence for the impact of self-administration on respondents' reports of offending behavior. This finding is inconsistent with the large body of knowledge on sensitive questions (e.g., Gnambs & Kaspar, 2015; Richman et al., 1999; Tourangeau & Yan, 2007). Methodological experiments have repeatedly shown that participants' willingness to disclose sensitive information is reduced when their answers are provided to an interviewer, compared to self-administered survey conditions. In the present review (Gomes et al., 2019), we have pooled a total of three experiments comparing offending reports obtained either using personal interviews or self-administered surveys (i.e., Krohn et al., 1974; Hindelang et al., 1981; Potdar & Koenig, 2005). These studies reported very similar offending scores obtained using these two modes of administration. However, before concluding that SRO are not impacted by mode effects, we must consider that two of these experiments were carried out more than three decades ago (i.e., Krohn et al., 1974; Hindelang et al., 1981), and that the third study included only two types of offending behavior, i.e. carrying a weapon/gun and engaging in abusive/violent behavior after drinking (Potdar & Koenig, 2005). These features may compromise the ability of these studies to find mode effects. Therefore, the conclusion that current questionnaires of offending behavior are not affected by modes of administration based on these experiments may be misleading, and more methodological research is clearly needed.

Offending is a socially undesirable behavior that can generate feelings of guilt, shame, and fear of legal consequences. For these reasons, the disclosure of offending behavior is usually considered a sensitive topic. However, the lack of evidence for the impact of mode effects on SRO is inconsistent with the findings from methodological experiments using other types of sensitive questions. This led us to consider how sensitive questions about offending behavior really are. In Chapter IV (Gomes et al., 2021b),

we aimed to pre-test the sensitivity of the questions about offending behavior. In order to achieve this objective, we have created a multi-dimensional assessment of question sensitivity based on the three-dimensional definition (i.e., intrusiveness, threat of disclosure, and social desirability) proposed by Tourangeau and Yan (2007). This is the first study that we are aware of that assessed question sensitivity based on this multi-dimensional definition. Findings from this evaluation showed that most offending questions scored higher on topic sensitivity than a question on sexual behavior. This finding provides evidence that questions about offending behavior are highly sensitive. Furthermore, this assessment provided an evaluation of topic sensitivity for each offending question, allowing a ranking of offending questions from the least sensitive to the most sensitive offending topic. Future methodological research will benefit from these findings because they allow researchers to control for the effect of topic sensitivity within offending questionnaires.

In addition to the main objective of pre-testing the sensitivity of offending questions, in the experiment reported in Chapter IV we carried out an experimental manipulation to test whether the recall period affects the respondents' perceptions of question sensitivity. The questions about offending behavior presented a randomly selected reference period (i.e., lifetime, past-year, or past-month). This is the first study that we are aware of that tested the impact of recall periods on participants' perceptions of question sensitivity. Our findings provided evidence that disclosing information about offending behavior occurring over a distant time period is perceived as less sensitive than disclosing information about recent offending. These findings provide relevant information to the study of sensitive questions by showing that the recall periods within behavioral questions affect respondents' perceptions about question sensitivity which, in turn, may impact respondents' willingness to report sensitive information.

In Chapter IV (Gomes et al., 2021b), we have established that offending behavior is a highly sensitive topic. Therefore, according to the literature on sensitive questions (e.g., Gnamb & Kaspar, 2015; Richman et al., 1999; Tourangeau & Yan, 2007), one would expect that self-administration of offending surveys would result in higher rates of offending behavior than in personal interviews where the respondents have to disclose sensitive information to a third person. However, the available evidence reviewed in Chapter III (Gomes et al., 2021a) suggests that SRO are not influenced by mode effects, which led influential authors to conclude that self-reported rates of offending are generally stable over different modes of administration (e.g., Hindelang et al., 1981; Thornberry & Krohn, 2000). The failure to determine and account for the impact of modes of administration on SRO can compromise the conclusions of the studies using this methodology. Therefore, in order to understand the apparently contradictory results from the literature on sensitive questions and the experiments using SRO, as well as

to improve the accuracy of offending data collected using self-reports, we have carried out a methodological experiment presented in Chapter V (Gomes et al., 2021c).

In Chapter V (Gomes et al., 2021c), we presented an experimental study composed of two methodological experiments. The first experiment was carried out in Portugal and the second was a replication experiment carried out in Florida. These experiments used a 2 x 2 factorial design where we have randomly manipulated modes of administration (i.e., interviewer-administered vs. self-administered) and modes of data collection (i.e., paper-and-pencil vs. computer-assisted interviews). Our results provided solid evidence that the likelihood that respondents would admit the practice of offending behavior was higher in self-administered conditions than in face-to-face interviews. The present findings on the influence of mode effects on SRO are in line with the literature on sensitive questions (e.g., Bradburn et al., 1979; Tourangeau et al., 2000). On the other hand, the manipulation of modes of data collection resulted in similar scores on offending behavior obtained through paper-and-pencil and computer-assisted modes, only slightly favoring higher disclosure of offending in paper-and-pencil questionnaires. This result was inconsistent with the main findings from the literature on sensitive topics (Gnambs & Kaspar, 2015; Gomes et al., 2019; Richman et al., 1999), where a higher likelihood of disclosure is often found in computer-assisted modes of data collection. However, the literature on sensitive questions is very inconsistent in regard to the effect of computerization on participants' willingness to report sensitive information.

Implications

With the present dissertation, we aimed to address some of the main gaps in the literature regarding the assessment of offending behavior. In doing so, we have discussed key issues in crime measurement, we have provided comprehensive reviews on the main measurement techniques for offending behavior, and we have carried out methodological experiments in order to improve the accuracy of SRO. This is a methodological dissertation and, as a consequence, the direct implications of this work relate to the choice of research methods in attempting to improve the quality of research measurements and procedures in the study of offending behavior.

One of the main implications of the review of the literature presented in Chapter I (Gomes et al., 2018) is the identification that, not only do some measures of offending provide more valid results than others, the use of different crime assessments can lead to completely opposite conclusions (e.g., Farrington et al., 2007; Kazemian & Farrington, 2005). The discrepancy between the assessments of offending behavior often results in ambiguous conclusions, which can seriously undermine the quality of

research findings, as well as their ability to inform theory and policy. Rigorous assessments of the dependent variable are crucial for the development of an unambiguous body of knowledge. However, as we have discussed in Chapter I, each offending measure has its own advantages and disadvantages, which makes it very important for researchers to select the most appropriate assessment technique based on the research questions. In this dissertation, we provided a discussion about the qualities and limitations of each measurement, in order to inform researchers in their selection of the most appropriate method. However, a key implication of the discrepancy between offending measures is that, since no methodology provides a perfect assessment of the offending phenomenon, researchers should apply a mixed-methods approach and use multiple assessments to describe and explain offending (Farrington et al., 2007).

Regarding field observation techniques, the systematic review of field experiments described in Chapter II (Gomes et al., 2021a) shows that offending behaviors, such as fraud, stealing, and shoplifting are very suitable to be studied using rigorous observation measures. Chapter II identified multiple assessment techniques that allow the study of offending behavior with high measurement quality, providing researchers with a review of methods that can be applied to study offending behaviors. Furthermore, in the second chapter, we showed that dishonesty can be efficiently assessed in the field, which can be a relevant dependent variable that provides rigorous information for theory testing within criminology and the psychology of offending behavior. However, more research is needed to understand the relationship between offending and dishonesty. Finally, the review provided in Chapter II demonstrates the importance of the field experimental design in the study of offending behavior. Through naturalistic field experiments, researchers can test theoretical hypotheses and cause-and-effect relationships in real-life conditions where participants are unaware that their behavior is being scrutinized. In our systematic review (Gomes et al., 2021a), we have demonstrated how it is possible to test theory using field experiments (i.e., SEU). In conclusion, we hope Chapter II will motivate researchers to study offending and dishonest behavior using field experiments, in order to gradually improve our knowledge about the causes of deviant conduct.

In regard to implications for the research using self-report techniques to assess offending behavior, the study presented in Chapter III (Gomes et al., 2019) identifies the methodological experiments carried out with SRO, providing a summary of best practice guidelines in order to improve the quality of self-reported data. In sum, the reviewed methodological experiments suggest that SRO provides similar behavioral estimates through different modes of administration when the survey is supervised either by teachers or by research staff, as well as in questionnaires that use different definitions of the reference period. On the other hand, the methodological evidence reviewed in Chapter III suggests

that SRO can provide more accurate estimates of offending behavior when: a) using shorter questionnaires with fewer follow-up questions, b) using response formats with more options than the traditional dichotomous format; c) using anonymous conditions and where the respondents' answers are not disclosed to others or result in a follow-up; d) the data collection occurs in school environments (rather than in their homes); and e) the interviewers are dressed casually (rather than formally). Further, techniques such as the bogus pipeline can motivate participants to respond more honestly to questions about offending behavior.

Contrary to the methodological literature on sensitive questions (e.g., Richman et al., 1999; Tourangeau & Yan, 2007), the methodological experiments using SRO reviewed in Chapter III (Gomes et al., 2019) failed to demonstrate any evidence of the benefits of self-administration of surveys over personal interviews. Taking into account these incongruent results, in Chapter IV we have pre-tested the sensitivity of the questions about offending behavior (Gomes et al., 2021b). Two main implications for the study of SRO can be derived from this study. First, this study establishes that respondents perceive questions about offending behavior as highly sensitive. As a consequence, and similarly to other sensitive topics, SRO may be subject to editing processes such as motivated misreporting. Therefore, researchers using self-reports to assess offending behavior should consider the accumulated body of knowledge developed by survey researchers in asking sensitive questions (e.g., Bradburn et al., 1979; Sudman & Bradburn, 1974; Tourangeau et al., 2000; Tourangeau & Yan, 2007). Second, our experiment (Gomes et al., 2021b) shows that the respondents' perceptions about question sensitivity are influenced by the reference period. Respondents perceive questions about recent offending (i.e., during past-year and past-month) as more threatening than questions about lifetime offending. In turn, questions about recent offending can be more prone to motivated misreporting than questions about a distant past. This study identifies recall periods as a relevant variable in the study of truthful disclosure of sensitive information, and more research is needed to establish the best practices in this regard.

Finally, following the results from the previous chapters, we carried out the study described in Chapter V (Gomes et al., 2021c). In this study, we have developed two methodological experiments that provided evidence that SRO can be influenced by modes of administration and modes of data collection. Regarding the manipulation of modes of data collection, despite our results slightly favoring higher reports of offending behavior in paper-and-pencil conditions, the overall estimates obtained with paper-and-pencil and computer-assisted modes were fairly comparable. On the other hand, this study was successful in demonstrating that respondents disclose higher rates of offending behavior in self-administered conditions, compared to face-to-face interviews. These findings are in accordance with the literature on

sensitive questions, suggesting that participants are less willing to report their own offenses when answers are provided to a third person (i.e., the interviewer). Therefore, when possible, survey researchers should implement self-administered modes of administration to assess offending behavior, which is expected to result in more accurate estimations of the true amount of offending behavior.

Strengths and limitations

The present dissertation provides a comprehensive guide to the multiple issues, methodologies, and gaps in the literature regarding the assessment of offending behavior. Chapter I (Gomes et al., 2018) provides a broad review of the main measurement techniques for offending behavior. This study highlighted the main issues in measuring crime, as well as the advantages and limitations of crime measures. The discussion about the limitations of the techniques used in the assessment of offending behavior provided very relevant information about the topics that need more methodological research and served as a guide for the subsequent studies within this dissertation. As a limitation, despite acting as an introductory chapter, the study described in Chapter I is a narrative literature review. Therefore, the scope of the studies included in this chapter can be limited by the subjective nature of the selection process.

The systematic review of field experiments presented in Chapter II (Gomes et al., 2021a) discussed the main advantages of the field experimental design in the development of our knowledge about human behavior. Furthermore, this systematic review provided a summary of a wide range of methods used to evaluate offending and dishonesty in the real world. The field experiments reviewed in Chapter II consisted of a total of 106 samples from 44 different countries. This cultural diversity is an advantage that adds strength to the generalizability of the findings. Regarding potential limitations, in our review, we have included field experiments studying stealing and monetary dishonesty, while other types of deviant behaviors that can provide relevant information for the study of offending, such as littering (e.g., Ramos & Torgler, 2012) or illegal disposal of garbage (e.g., Dur & Vollaard, 2019), were not included. Further, the present review did not include field experiments that studied dishonest behavior that go beyond monetary gains, such as the field experiment carried out by Tobol et al. (2020) that observed passerby members of the public and evaluated dishonest use of masks around their chins in places where wearing a mask was mandatory due to the COVID-19 pandemic.

As for our systematic review of measurement biases in SRO presented in Chapter III (Gomes et al., 2019), our findings highlighted relevant aspects of the procedure and design of self-reports that can influence the quality of data on offending behavior. The studies reviewed in this chapter provide relevant information regarding the biasing factors in asking questions about offending behavior, and our systematic

review provides a valuable summary of these findings in order to improve the quality of the information obtained using SRO. However, the interpretation of these findings must be considered with caution. Methodological research using SRO is very scarce and most of the biasing factors explored in these experiments lacked replication. Also, due to the small number of effect sizes per experimental manipulation, we were not able to test for publication bias. Nevertheless, this review included unpublished methodological experiments in order to assess the best available information. Further, the broad conceptualization of offending behavior (e.g., sexual aggression, previous imprisonment, etc.), and the inclusion of reports from different recall periods (e.g., lifetime, past-year, or past-month), can result in unstandardized conceptions of the dependent variable and create doubt about the conclusions of this review.

In the face of the apparently inconsistent findings from methodological studies using SRO and other sensitive topics, we set out to develop the experimental studies described in Chapter IV (Gomes et al., 2021b) and Chapter V (Gomes et al., 2021c). The study described in Chapter IV took into account Hindelang et al.'s (1981) conclusion that the apparent lack of self-administration effects on SRO could be due to the non-threatening nature of disclosing information about offending behavior. We carried out an assessment of respondents' perceptions of the sensitivity of offending questions. Our findings established SRO as a highly sensitive topic and showed that recall periods impact respondents' perceptions about the questions' sensitivity (Gomes et al., 2021b). These findings supported the experiments described in Chapter V (Gomes et al., 2021c) and strengthen our argument to test the impact of mode effects on reports of offending. Results from our last study demonstrate that modes of administration affect respondents' willingness to report their own offenses. These findings provide very relevant information to improve the accuracy of SRO and reconcile the literature on SRO and sensitive questions. Furthermore, perhaps the main strength of these methodological studies (Gomes et al., 2021b, 2021c) is the fact that the original experiments carried out in Portugal were replicated in the US following the exact same procedures. The replication of our original experiments in a different cultural setting shows that our findings are robust.

As for the limitations of the experiments described in Chapter IV (Gomes et al., 2021b), the study conducted in the US with students from a university from central Florida presented a very small sample size ($N = 43$), which seriously limited the generalization of its findings. However, the US replication of the experiment described in Chapter IV was very useful as a pilot study that allowed us to check that the order of sensitivity of offending questions was very similar in these two different cultures. Regarding the last methodological study described in Chapter V (Gomes et al., 2021c), one of the main limitations is

concerned with our sample. In both the Portuguese and the American samples, the participants were university students (mainly females) which compromised the generalizability of our findings. Further, seeing that the prevalence of offending among university students was generally low, especially for more recent and more serious types of offenses, we were limited in our ability to detect mode effects. However, despite these limitations, our methodological experiments were able to provide evidence that SRO are subject to mode effects caused by self-administration of surveys. Finally, similarly to the literature on sensitive questions (e.g., Tourangeau & Yan, 2007), we have used the 'more is better' assumption and assume that higher rates of disclosure of offending indicate more honest reporting on the part of the respondents. However, there is a chance for dishonest overreporting of offending behavior, which would compromise our conclusions. Nevertheless, that would go against the large body of evidence on underreporting sensitive behaviors (e.g., Krumpal, 2013; Tourangeau et al., 2000) and, on the other hand, our results are in line with the literature on sensitive questions. Therefore, whenever possible, self-administered modes of administration should be preferable over face-to-face interviews.

Future studies

One of the main purposes of this doctoral dissertation was to identify the key issues and methodologies in measuring offending behavior. In doing so, we have identified multiple aspects of crime measurement that need more research. In this section, we will summarize some of the main needs for further research in crime measurement resulting from the work discussed in the previous chapters.

In regard to field experiments, the systematic review described in Chapter II (Gomes et al., 2021a) has discussed the large benefits of the field experimental design for the development of scientific knowledge. We hope that, by summarizing field experiments used to study stealing and monetary dishonesty, this work will inspire researchers to carry out more experiments in the real world to test hypotheses related to the causes of offending behavior. Furthermore, Chapter II shows that dishonesty is an easily assessable variable in real life that can provide relevant contributions to the study of offending behavior. Therefore, researchers interested in offending should consider including measures of dishonesty in their experiments. However, the link between offending and dishonesty needs more research in the future. Finally, the field experiments reviewed in Chapter II focused mainly on situational factors. In the future, researchers should consider manipulating other non-situational variables, such as impulsivity or peer factors (e.g., Defoe et al., 2019). In sum, we consider that future researchers should carry out more naturalistic field experiments to explore the causes of offending behavior.

As for the measurement qualities of SRO, in Chapter III (Gomes et al., 2019) we have summarized the available methodological experiments using questions about offending behavior and identified several aspects that need further research. Perhaps the most manifest topic in need of further research is concerned with the lack of evidence that self-administration leads to better estimates of offending behavior than face-to-face interviews. This led us to a series of experiments that we have described in Chapter IV (Gomes et al., 2021b) and Chapter V (Gomes et al., 2021c). In Chapter IV, we have developed an assessment of topic sensitivity and established that offending questions are highly sensitive. This evaluation allowed us to develop a ranking of offending questions from the least to the most sensitive item in the delinquency questionnaire. In order to further explore the hypothesis that the benefits of self-administration are higher for more sensitive topics (e.g., Tourangeau & Yan, in press), future methodological experiments may consider our findings and test the assumption that mode effects would be more evident for more sensitive offending questions. Further, the study described in Chapter IV showed the impact of recall period on participants' perceptions of question sensitivity. However, we do not know to what extent different recall periods impact respondents' willingness to disclose offending and other sensitive behaviors. More research on this topic is needed in order to gradually delimit the guidelines for best practices in asking sensitive questions.

In Chapter V (Gomes et al., 2021c), we have demonstrated that, similarly to other types of sensitive questions, SRO are subject to mode effects. In other words, respondents in self-administered conditions provided higher estimates of offending behavior than respondents in face-to-face interviews. However, since there is no 'gold standard' criterion in order to judge the veracity of reports of offending behavior (Thornberry & Krohn, 2000), we have applied the 'more is better' assumption to determine that the condition resulting in higher estimates of offending corresponds to the most honest responding (Tourangeau & Yan, 2007). Future methodological experiments using illegal drug use may include the collection of physiological data that serves as an external criterion for drug consumption. As demonstrated in Chapter IV, questions about illegal drug use (i.e., Cannabis/marijuana/hash, Ecstasy/LSD/amphetamines, and Heroin/cocaine/crack) are also perceived as highly sensitive. In our experiments described in Chapter V (Gomes et al., 2021c), we have collected self-reports of drug use, as well as hair samples, in order to test for biomarkers of recent drug use which will provide an external criterion to establish the honest responding in the different conditions. We hope to publish these results in the near future after the completion of the chemical analysis of the hair samples (Gomes et al., 2021d).

Finally, since SRO are the most widely used method in the study of the causes of offending, the lack of methodological research can seriously compromise the quality of research conclusions. Therefore,

it is of great importance that future studies consider other aspects of collecting offending data using self-reports, such as different modes of administration, the supervision of the survey completion, the design of the questionnaire, among many others. In our research team, we have developed a series of methodological experiments in order to inform researchers about the best practices in using SRO. For example, we have compared the incidence of offending provided by open-ended and closed-ended response formats (Gomes et al., 2021e; Korotchenko et al., 2020). We have experimentally tested the impact of the number of follow-up questions on participants' willingness to report offending by comparing delinquency questionnaires with zero, one, and five follow-up questions after each offending item (Gomes et al., 2020, 2021f). Further, we have looked at the impact of the interviewer's sex on respondents' willingness to disclose offending information (Gomes et al., 2021g).

Further, more research is needed to study the potential impact of testing effects on SRO in longitudinal studies. Testing effects refer to changes in respondents' answers caused by previous completion of the same questionnaires (Thornberry, 1989). Considering that our knowledge about offending behavior is increasingly reliant on longitudinal studies, exploring the extent to which SRO are affected by testing effects, as well as testing ways to mitigate these impacts, should be a priority of survey researchers in the field of offending behavior. Krohn et al. (2012) suggested that testing effects can be studied by the development of a longitudinal study where participants are randomly selected to enter the study at different waves of data collection. The first group would receive the survey at the first wave, the next group would start at the second wave, and so on. If SRO were to be subject to testing effects, then we would find systematic differences in the estimates of offending behavior between these groups. Adding to the innovative design proposed by Krohn et al. (2012), we believe that, by including methodological manipulations in this study, it would be possible to test ways to mitigate the biasing effects of testing effects and improve the accuracy of SRO in longitudinal studies. For example, in our research team, we have developed a small-scale longitudinal experiment with three waves of data collection separated by six-month intervals. In this longitudinal experiment, we have manipulated some aspects of the questionnaire, such as the number of follow-up questions (zero, one, or five follow-up questions) and the position of the offending questions within the questionnaire (at the beginning or at the end of the questionnaire). These results will provide evidence on the extent to which longitudinal studies are impacted by testing effects, as well as providing valuable information on how to mitigate these biasing factors and improve the validity of longitudinal assessments of offending behavior (Gomes et al., 2021h).

Conclusions

The present dissertation took on the crucial task of improving the quality of offending measurements. Criminal behavior is naturally secretive and presents serious incriminating threats to the offender. For these reasons, crime assessments will never provide perfect evaluations of offending behavior (Krohn et al., 2012). However, “it is not sufficient to attach warning labels to reports of self-reported delinquency, pointing to the possibility that differences in methods may result in different estimates of the amount of crime” (Enzmann, 2013, p. 149). It is crucial that researchers keep exploring the psychometric qualities of offending measures in order to improve the accuracy of criminal assessments and, as a consequence, the validity of research conclusions.

The studies presented in the first three chapters of this dissertation provide comprehensive reviews of the key issues and main assessment techniques for offending behavior. Mainly, we have provided a review of the main issues in measuring crime, discussing the advantages and limitations of the main methods of assessing offending behavior (Gomes et al., 2018). We have demonstrated that observation methods allow us to assess offending behavior as it happens in real-life conditions with high validity, and we have described the advantages of field experiments as a powerful tool to achieve solid and replicable results in the study of the causes of offending (Gomes et al., 2021a). Also, we have provided a systematic review of methodological experiments on SRO, which summarized the available evidence on the best practices in using self-reports to assess offending behavior (Gomes et al., 2019). Most of the current knowledge about the causes of offending is reliant on SRO, which makes the development of the self-report methodology one of the most important tasks in advancing our understanding of offending behavior (Auty & Farrington, 2015; Krohn et al., 2012).

In view of the scarcity of research on the best practices to improve the accuracy of SRO, we have developed a series of methodological experiments to test potential biasing factors in reporting offending behavior. In these experiments, we have explored the sensitivity of offending questions, as well as the impact of recall periods on participants’ perceptions of the sensitivity of offending questions (Gomes et al., 2021b). Further, we have tested the impact of modes of administration on participants’ willingness to disclose information about their own offending behavior (Gomes et al., 2021c). Overall, the methodological evidence developed within this dissertation provides relevant contributions to the improvement of data quality by adding to the guidelines for best practices in collecting offending behavior using the self-report methodology.

References

- Auty, K. M., Farrington, D. P., & Coid, J. W. (2015). The validity of self-reported convictions in a community sample: Findings from the Cambridge Study in Delinquent Development. *European Journal of Criminology*, *12*(5), 562-580. <https://doi.org/10.1177/1477370815578198>
- Beebe, T. J., Harrison, P. A., Park, E., McRae, J. A., Jr., & Evans, J. (2006). The effects of data collection mode and disclosure on adolescent reporting of health behavior. *Social Science Computer Review*, *24*(4), 476-488. <https://doi.org/10.1177/0894439306288690>
- Bradburn, N. M., Sudman, S., Blair, E., Locander, W., Miles, C., Singer, E., & Stocking, C. (1979). *Improving interview method and questionnaire design: Response effects to threatening questions in survey research*. Jossey-Bass.
- Brener, N. D., Eaton, D. K., Kann, L., Grunbaum, J. A., Gross, L. A., Kyle, T. M., & Ross, J. G. (2006). The association of survey setting and mode with self-reported health risk behaviors among high school students. *Public Opinion Quarterly*, *70*(3), 354-374. <https://doi.org/10.1093/poq/nfl003>
- Buckle, A., & Farrington, D. P. (1984). An observational study of shoplifting. *British Journal of Criminology*, *24*(1), 63-73. <https://doi.org/10.1093/oxfordjournals.bjc.a047425>
- Buckle, A., & Farrington, D. P. (1994). Measuring shoplifting by systematic observation: A replication study. *Psychology, Crime and Law*, *1*(2), 133-141. <https://doi.org/10.1080/10683169408411946>
- Clark, J. P., & Tiffit, L. L. (1966). Polygraph and interview validation of self-reported deviant behavior. *American Sociological Review*, *31*(4), 516-523. <https://doi.org/10.2307/2090775>
- Defoe, I. N., Dubas, J. S., & Romer, D. (2019). Heightened adolescent risk-taking? Insights from lab studies on age differences in decision-making. *Policy Insights from the Behavioral and Brain Sciences*, *6*(1), 56-63. <https://doi.org/10.1177/2372732218801037>
- Dur, R., & Vollaard, B. (2019). Salience of law enforcement: A field experiment. *Journal of Environmental Economics and Management*, *93*, 208-220. <https://doi.org/10.1016/j.jeem.2018.11.011>
- Eaton, D. K., Brener, N. D., Kann, L., Denniston, M. M., McManus, T., Kyle, T. M., Roberts, A. M., Flint, K. H., & Ross, J. G. (2010). Comparison of paper-and-pencil versus Web administration of the Youth Risk Behavior Survey (YRBS): Risk behavior prevalence estimates. *Evaluation Review*, *34*(2), 137-153. <https://doi.org/10.1177/0193841X10362491>
- Enzmann, D. (2013). The impact of questionnaire design on prevalence and incidence rates of self-reported delinquency: Results of an experiment modifying the ISRD-2 questionnaire. *Journal of*

- Contemporary Criminal Justice*, 29(1), 147-177.
<https://doi.org/10.1177/1043986212470890>
- Farrington, D. P. (1979). Experiments on deviance with special reference to dishonesty. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 12, pp. 207-252). Academic Press.
[https://doi.org/10.1016/S0065-2601\(08\)60263-4](https://doi.org/10.1016/S0065-2601(08)60263-4)
- Farrington, D. P. (2001). *What has been learned from self-reports about criminal careers and the causes of offending?* Home Office Online Report.
<https://www.crim.cam.ac.uk/sites/www.crim.cam.ac.uk/files/srdrep.pdf>
- Farrington, D. P., & Jolliffe, D. (2004). England and Wales. In D. P. Farrington, P. A. Langan, & M. Tonry (Eds.), *Cross-national studies in crime and justice* (pp. 1–38). Bureau of Justice Statistics.
<http://www.ojp.usdoj.gov/bjs>
- Farrington, D. P., Jolliffe, D., Hawkins, J. D., Catalano, R. F., Hill, K. G., & Kosterman, R. (2003). Comparing delinquency careers in court records and self-reports. *Criminology*, 41(3), 933-958.
<https://doi.org/10.1111/j.1745-9125.2003.tb01009.x>
- Farrington, D. P., Jolliffe, D., Loeber, R., & Homish, D. L. (2007). How many offenses are really committed per juvenile court offender?. *Victims and offenders*, 2(3), 227-249.
<https://doi.org/10.1080/15564880701403934>
- Farrington, D. P., & Knight, B. J. (1980). Stealing from a “lost” letter: Effects of victim characteristics. *Criminal Justice and Behavior*, 7(4), 423-436.
<https://doi.org/10.1177/009385488000700406>
- Farrington, D. P., Lösel, F., Braga, A. A., Mazerolle, L., Raine, A., Sherman, L. W., & Weisburd, D. (2020). Experimental criminology: Looking back and forward on the 20th anniversary of the Academy of Experimental Criminology. *Journal of Experimental Criminology*, 16, 649–673.
<https://doi.org/10.1007/s11292-019-09384-z>
- Ferraro, K. F., & LaGrange, R. (1987). The measurement of fear of crime. *Sociological Inquiry*, 57(1), 70-97. <https://doi.org/10.1111/j.1475-682X.1987.tb01181.x>
- Franzen, A., & Pointner, S. (2013). The external validity of giving in the dictator game. *Experimental Economics*, 16(2), 155-169. <https://doi.org/10.1007/s10683-012-9337-5>
- Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, 47(4), 1237-1259.
<https://doi.org/10.3758/s13428-014-0533-4>

- Gomes, H. S., Farrington, D. P., Defoe, I. N., & Maia, Â. (2021a). Field experiments on dishonesty and stealing: What have we learned in the last 40 years?. *Journal of Experimental Criminology*. Advance online publication. <https://doi.org/10.1007/s11292-021-09459-w>
- Gomes, H. S., Farrington, D. P., Krohn, M. D., Cunha, A., Jurdi, J., Sousa, B., Morgado, D., Hoft, J., Hartsell, E., Kassem, L., & Maia, Â. (2021c). *The impact of modes of administration on self-reports of offending: A two methodological experiment replication* [Manuscript submitted for publication]. School of Psychology, University of Minho.
- Gomes, H. S., Farrington, D. P., Krohn, M. D., & Maia, Â. (2021b). *How sensitive are self-reports of offending?: The impact of recall periods on question sensitivity* [Manuscript submitted for publication]. School of Psychology, University of Minho.
- Gomes, H. S., Farrington, D. P., Krohn, M. D., & Maia, Â. (2021d). *Testing the impact of modes of administration on self-reports of drug use: Using hair analysis of marijuana use* [Manuscript in preparation]. School of Psychology, University of Minho.
- Gomes, H. S., Farrington, D. P., Krohn, M. D., & Maia, Â. (2021h). *Are self-reports of offending in longitudinal studies affected by testing effects? Evidence from a longitudinal experiment with Portuguese adolescents* [Manuscript in preparation]. School of Psychology, University of Minho.
- Gomes, H. S., Farrington, D. P., Maia, Â., & Krohn, M. D. (2019). Measurement bias in self-reports of offending: A systematic review of experiments. *Journal of Experimental Criminology*, *15*(3), 313-339. <https://doi.org/10.1007/s11292-019-09379-w>
- Gomes, H. S., Farrington, D. P., Maia, Â., & Krohn, M. D. (2020, November 18-21). *The impact of question format on self-reports of offending: An experimental manipulation of follow-up questions* [Paper presentation]. Accepted for presentation at the 75th American Society of Criminology Annual Meeting (ASC), Washington, D.C., U.S.
- Gomes, H. S., Farrington, D. P., Maia, Â., & Krohn, M. D. (2021f). *The impact of question format on self-reports of offending: An experimental manipulation of follow-up questions* [Manuscript in preparation]. School of Psychology, University of Minho.
- Gomes, H. S., Farrington, D. P., Maia, Â., & Krohn, M. D. (2021g). *Interviewer effects on self-reports of offending: A quasi-experiment on the impact of interviewer's gender* [Manuscript in preparation]. School of Psychology, University of Minho.
- Gomes, H. S., Korotchenko, S., Farrington, D. P., Maia, Â., & Krohn, M. D. (2021e). *Response formats in measuring self-reports of sensitive behavior among college students* [Manuscript in preparation]. School of Psychology, University of Minho.

- Gomes, H. S., Maia, Â., & Farrington, D. P. (2018). Measuring offending: Self-reports, official records, systematic observation and experimentation. *Crime Psychology Review, 4*(1), 26-44. <https://doi.org/10.1080/23744006.2018.1475455>
- Gribble, J. N., Miller, H. G., Cooley, P. C., Catania, J. A., Pollack, L., & Turner, C. F. (2000). The impact of T-ACASI interviewing on reported drug use among men who have sex with men. *Substance use & misuse, 35*(6-8), 869-890. <https://doi.org/10.3109/10826080009148425>
- Hamby, S., Sugarman, D. B., & Boney-McCoy, S. (2006). Does questionnaire format impact reported partner violence rates?: An experimental study. *Violence and Victims, 21*(4), 507-518. <https://doi.org/10.1891/0886-6708.21.4.507>
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature, 42*(4), 1009-1055. <https://doi.org/10.1257/0022051043004577>
- Horney, J., & Marshall, I. H. (1992). An experimental comparison of two self-report methods for measuring lambda. *Journal of Research in Crime and Delinquency, 29*(1), 102-121. <https://doi.org/10.1177/0022427892029001006>
- Hindelang, M. J., Hirschi, T., & Weis, J. G. (1981). *Measuring delinquency*. Sage.
- Huizinga, D., & Elliott, D. S. (1986). Reassessing the reliability and validity of self-report delinquency measures. *Journal of Quantitative Criminology, 2*(4), 293-327. <https://doi.org/10.1007/BF01064258>
- Jolliffe, D., & Farrington, D. P. (2014). Self-reported offending: Reliability and validity. In G. Bruinsma, & D. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice* (pp. 4716-4723). Springer. https://doi.org/10.1007/978-1-4614-5690-2_648
- Kazemian, L., & Farrington, D. P. (2005). Comparing the validity of prospective, retrospective, and official onset for different offending categories. *Journal of Quantitative Criminology, 21*(2), 127-147. <https://doi.org/10.1007/s10940-005-2489-0>
- King, C. A., Hill, R. M., Wynne, H. A., & Cunningham, R. M. (2012). Adolescent suicide risk screening: the effect of communication about type of follow-up on adolescents' screening responses. *Journal of Clinical Child & Adolescent Psychology, 41*(4), 508-515. <https://doi.org/10.1080/15374416.2012.680188>
- Kivivuori, J., Salmi, V., & Walser, S. (2013). Supervision mode effects in computerized delinquency surveys at school: Finnish replication of a Swiss experiment. *Journal of Experimental Criminology, 9*(1), 91-107. <https://doi.org/10.1007/s11292-012-9162-z>

- Klein, M. (1987). Watch out for that last variable. In S. Mednick, T. Moffitt, & S. Stack (Eds.), *The causes of crime: New biological approaches* (pp. 25-41). Cambridge University Press. <https://doi.org/10.1017/CBO9780511753282>
- Knapp, H., & Kirk, S. A. (2003). Using pencil and paper, Internet and touch-tone phones for self-administered surveys: Does methodology matter? *Computers in Human Behavior*, *19*(1), 117-134. [https://doi.org/10.1016/S0747-5632\(02\)00008-0](https://doi.org/10.1016/S0747-5632(02)00008-0)
- Korotchenko, S., Gomes, H. S., Krohn, M. D., Maia, Â., & Farrington, D. P. (2021, November 17-20). *How response formats impact the disclosure of offending behavior: Open-ended vs. closed-ended follow-up questions* [Paper presentation]. Accepted for presentation at the 76th American Society of Criminology Annual Meeting (ASC), Chicago, IL, U.S.
- Krohn, M., Thornberry, T., Bell, K., Lizotte, A., & Phillips, M. (2012). Self-report surveys within longitudinal panel designs. In D. Gadd, S. Karstedt, & S. Messner (Eds.), *The Sage handbook of criminological research* (pp. 23-35). Sage. <https://dx.doi.org/10.4135/9781446268285.n2>
- Krohn, M. D., Waldo, G. P., & Chiricos, T. G. (1974). Self-reported delinquency: A comparison of structured interviews and self-administered checklists. *Journal of Criminal Law and Criminology*, *65*(4), 545-553. <https://doi.org/10.2307/1142528>
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*(4), 2025-2047. <https://doi.org/10.1007/s11135-011-9640-9>
- Lucia, S., Herrmann, L., & Killias, M. (2007). How important are interview methods and questionnaire designs in 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*(1), 39-64. <https://doi.org/10.1007/s11292-007-9025-1>
- Maxfield, M. G., & Babbie, E. R. (2009). *Basics of research methods for criminal justice and criminology* (2nd ed.). Cengage Learning.
- McCall, G. J. (1984). Systematic field observation. *Annual review of sociology*, *10*, 263-282. <https://doi.org/10.1146/annurev.so.10.080184.001403>
- Osgood, D. W., McMorris, B. J., & Potenza, M. T. (2002). Analyzing multiple-item measures of crime and deviance I: Item response theory scaling. *Journal of Quantitative Criminology*, *18*, 267-296. <https://doi.org/10.1023/A:1016008004010>
- Pepper, J. V., & Petrie, C. V. (Eds.). (2003). *Measurement problems in criminal justice research: Workshop summary*. National Academy Press. <https://doi.org/10.17226/10581>

- Piquero, A. R., Schubert, C. A., & Brame, R. (2014). Comparing official and self-report records of offending across gender and race/ethnicity in a longitudinal study of serious youthful offenders. *Journal of Research in Crime and Delinquency*, 51(4), 526–556. <https://doi.org/10.1177/0022427813520445>
- Potdar, R., & Koenig, M. A. (2005). Does audio-CASI improve reports of risky behavior? Evidence from a randomized field trial among young urban men in India. *Studies in Family Planning*, 36(2), 107-116. <https://doi.org/10.1111/j.1728-4465.2005.00048.x>
- Ramos, J., & Torgler, B. (2012). Are academics messy? Testing the broken windows theory with a field experiment in the work environment. *Review of Law & Economics*, 8(3), 563-577. <https://doi.org/10.1515/1555-5879.1617>
- Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, 84(5), 754-775. <https://doi.org/10.1037/0021-9010.84.5.754>
- Strang, E., & Peterson, Z. D. (2020). Use of a bogus pipeline to detect men's underreporting of sexually aggressive behavior. *Journal of Interpersonal Violence*, 35(1-2), 208-232. <https://doi.org/10.1177/0886260516681157>
- Sudman, S., & Bradburn, N. M. (1974). *Response effects in surveys: A review and synthesis*. Aldine Publishing Company.
- Theobald, D., Farrington, D. P., Loeber, R., Pardini, D. A., & Piquero, A. R. (2014). Scaling up from convictions to self-reported offending. *Criminal Behaviour and Mental Health*, 24(4), 265-276. <https://doi.org/10.1002/cbm.1928>
- Thornberry, T. P. (1989). Panel effects and the use of self-reported measures of delinquency in longitudinal studies. In M. W. Klein (Ed.), *Cross-national research in self-reported crime and delinquency* (pp. 347-369). Springer. https://doi.org/10.1007/978-94-009-1001-0_16
- Thornberry, T. P., & Krohn, M. D. (2000). The self-report method for measuring delinquency and crime. In D. Duffee (Ed.), *Measurement and analysis of crime and justice* (pp. 33–84). U.S. National Institute of Justice.
- Tobol, Y., Siniver, E., & Yaniv, G. (2020). Dishonesty and mandatory mask wearing in the COVID-19 pandemic. *Economics Letters*, 197, 109617. <https://doi.org/10.1016/j.econlet.2020.109617>
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511819322>

- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133(5), 859-883. <https://doi.org/10.1037/0033-2909.133.5.859>
- Tourangeau, R., & Yan, T. (in press). Reporting issues in surveys of drug use. *Substance Use and Misuse*.
- Trapl, E. S., Taylor, H. G., Colabianchi, N., Litaker, D., & Borawski, E. A. (2013). Value of audio-enhanced handheld computers over paper surveys with adolescents. *American Journal of Health Behavior*, 37(1), 62-69. <https://doi.org/10.5993/AJHB.37.1.7>
- van de Looij-Jansen, P. M., & de Wilde, E. J. (2008). Comparison of web-based versus paper-and-pencil self-administered questionnaire: Effects on health indicators in Dutch adolescents. *Health Services Research*, 43(5p1), 1708-1721. <https://doi.org/10.1111/j.1475-6773.2008.00860.x>
- van de Looij-Jansen, P. M., Goldschmeding, J. E., & de Wilde, E. J. (2006). Comparison of anonymous versus confidential survey procedures: Effects on health indicators in Dutch adolescents. *Journal of Youth and Adolescence*, 35(4), 652-658. <https://doi.org/10.1007/s10964-005-9027-0>
- Walser, S., & Killias, M. (2012). Who should supervise students during self-report interviews? A controlled experiment on response behavior in online questionnaires. *Journal of Experimental Criminology*, 8(1), 17-28. <https://doi.org/10.1007/s11292-011-9129-5>

APPENDIX



Universidade do Minho
SECSH

Subcomissão de Ética para as Ciências Sociais e Humanas

Identificação do documento: SECSH 052/2017

Título do projeto: *Self-report bias in measuring delinquent behaviour: Modes of administration, questionnaire design, and testing effects in longitudinal studies*

Investigador(a) Responsável: Angela Maia, Unidade de Investigação de Justiça e Violência do Centro de Investigação em Psicologia (CIPSI), Escola de Psicologia, Universidade do Minho (Orientadora)

Outros Investigadores: Hugo S. Gomes, Estudante de doutoramento em Psicologia Aplicada, Universidade do Minho; David P. Farrington (Instituto de Criminologia, Universidade de Cambridge)

PARECER


A Subcomissão de Ética para as Ciências Sociais e Humanas (SECSH) analisou o processo relativo ao projeto intitulado *Self-report bias in measuring delinquent behaviour: Modes of administration, questionnaire design, and testing effects in longitudinal studies*.

Os documentos apresentados revelam que o projeto obedece aos requisitos exigidos para as boas práticas na investigação com humanos, em conformidade com as normas nacionais e internacionais que regulam a investigação em Ciências Sociais e Humanas.

Face ao exposto, a SECSH nada tem a opor à realização do projeto na UMinho.

Braga, 04 de dezembro de 2017.

O Presidente


Digitally signed by
PAULO MANUEL
PINTO PEREIRA
ALMEIDA MACHADO
- Date: 2017.12.05
16:22:00 Z

Paulo Manuel Pinto Pereira Almeida Machado