Journal of Experimental Criminology
https://doi.org/10.1007/s11292-019-09379-w

# Measurement bias in self-reports of offending: a systematic review of experiments

Hugo S. Gomes[1] · David P. Farrington[2] · Ângela Maia[1] · Marvin D. Krohn[3]

Published online: 29 August 2019


## 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

✉ Hugo S. Gomes
hugo.santos.gomes@gmail.com

Extended author information available on the last page of the article

## 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 (for reviews, see Thornberry and Krohn 2000; Gomes et al. 2018a). 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).

Self-reports of offending were first introduced in an attempt to overcome the limitations of official records of crime (Porterfield 1943; Nye and Short 1957). 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 and Krohn 2000: p. 34). However, the great number of studies on the validity of self-reports of offending seen in the 1960s and 1970s (e.g., Clark and Tifft 1966; Kulik et al. 1968; Farrington 1973; Hardt and Peterson-Hardt 1977; 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 self-reports once and for all (Jolliffe and Farrington 2014).

Despite the scarcity of recent studies on the validity of self-reported offending, 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 and 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 and 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 self-reports of offending (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 self-reports of offending (e.g., Maxfield et al. 2000; Auty et al. 2015). 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 Tifft (1966) interviewed students with and without a polygraph in order to study the validity of self-reported deviant behavior. 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 and 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; Gribble et al. 2000), others found no significant differences in reports of alcohol use (e.g., Sobell and 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 and 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 and Miller 1997) and the recency of the behavior (Tourangeau and McNeeley 2003).

Unfortunately, research on sensitive questions commonly includes questions about income, voting, sexual behaviors, and drug use (Tourangeau and 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 and 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 self-reported offending. 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 self-reports of offending.

## 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 over-report*) 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., Chan et al. 2005; Baly and Cornell 2011; Huang and 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 self-report offending making it impossible to investigate their characteristics or predictors, such as the Item Count Technique (Wolter and Laier 2014), Unmatched Count Technique (Dalton et al. 1994),[1] and Randomized Response Technique (Wolter and 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 Fig. 1, 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 self-reported data on offending 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 (Grysman and 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

[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 and 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 and Preisendörfer 2013).

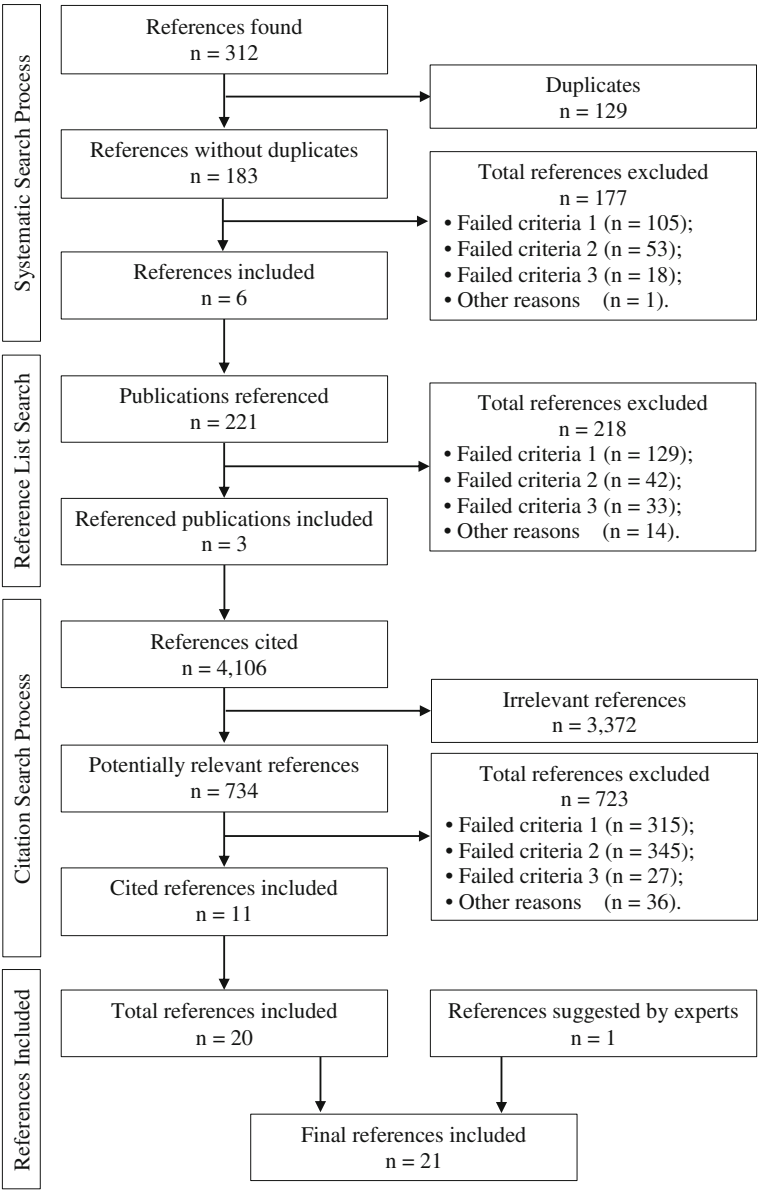


**Fig. 1** Flowchart of the process of systematic search

studies, but 218 studies were excluded for failing to meet the eligibility criteria or not reporting in the included languages.

The 9 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 4106 new references. An initial title search resulted in the exclusion of 3372 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.

### 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 self-reports of offending in different measurement techniques (i.e., experimental manipulations). In order to provide comparable information regarding the magnitude of different measurements, we estimated odds ratio effect sizes for each manipulation by evaluating the difference in the odds of self-reported offending behavior 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 (Hindelang et al. 1981; Baier 2017). 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 odds ratio 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 1 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 self-reports of offending caused by measurement manipulations. However, both outcome and experimental manipulation

**Table 1** Descriptive information on studies in the systematic review

| Study | Country | Sample | Topic | Recall period | Outcome measure |
|---|---|---|---|---|---|
| Baier (2017) | Germany | 2643 adolescent students | Mode of administration | Lifetime prevalence<br>Year prevalence | Delinquent behavior (violence, damage to property,<br>shoplifting, spraying graffiti, selling pirate copies) |
| Beebe et al. (1998) | USA | 368 adolescent students | Mode of administration | Lifetime prevalence<br>Year prevalence | Delinquent behavior (involved in gang, ran away<br>from home, damaged property, beat person up,<br>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 | 4506 adolescent students | Mode of administration and Setting | Year prevalence<br>30-day prevalence | Delinquent risk behavior (drove after drinking alcohol,<br>carried a weapon, carried a gun, physical fight,<br>dating violence) |
| Eaton et al. (2010) | USA | 5227 adolescent students | Mode of administration | Year prevalence<br>30-day prevalence | Delinquent risk behavior (drove after drinking alcohol,<br>carried a weapon, carried a gun, carried a weapon on<br>school property, physical fight, physical fight on<br>school property, dating violence) |
| Enzmann (2013) | Germany | 1629 adolescent students | Questionnaire design | Lifetime prevalence<br>Year prevalence | Delinquent behavior (vandalism, shoplifting, burglary,<br>bicycle theft, car theft, car break, snatching, carrying<br>weapons, extortion, group fight, assault, drug dealing) |

**Table 1** (continued)

| Study | Country | Sample | Topic | Recall period | Outcome measure |
|---|---|---|---|---|---|
| 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<br>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 and 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<br>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 |

**Table 1** (continued)

| Study | Country | Sample | Topic | Recall period | Outcome measure |
|---|---|---|---|---|---|
| | | | | | 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 | 1203 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 male residents in slums | Mode of administration | Lifetime prevalence | Risk behavior (carrying a weapon/gun and engaged in abusive, violent behavior after drinking) |
| Strang and Peterson (2016) | USA | 93 young, community men | Bogus pipeline | Lifetime prevalence | Sexual aggression (verbal coercion tactics, drugs and alcohol tactics, and force tactics of sexual assault) |

**Table 1** (continued)

| Study | Country | Sample | Topic | Recall period | Outcome measure |
|---|---|---|---|---|---|
| Trapl et al. (2013) | USA | 275 adolescent students | Mode of administration | Lifetime prevalence | Sensitive behaviors (shoplifting) |
| Turner et al. (1998) | USA | 1672 adolescent males | Mode of administration | Year prevalence<br>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 | 1197 adolescent students | Supervision | Lifetime prevalence<br>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) |

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 and Peterson 2016). 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 2 summarizes the results of these experiments, organized by measurement manipulations. For each manipulation, we provided information regarding the odds ratio (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 2 provides information on the experimental manipulation under analysis (experimental condition A vs. experimental condition B). Considering the calculation of odds ratio 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 2, we present the comparison of personal interview (i.e., condition A) vs. self-administered questionnaire (i.e., condition B) (Krohn et al. 1974); an $OR = .701$ 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).

## 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

**Table 2** Main findings of experiments in the systematic review

| Study | Comparison ($p < .05$) | OR | 95% CI | $z$ | $p$ |
|---|---|---|---|---|---|
| Modes of administration | | | | | |
| Personal interview (PI) vs. self-administered questionnaire (SAQ) ($k = 3$) | | | | | |
| Krohn et al. (1974) | 0 of 6 | .701 | .340, 1.445 | − .962 | .336 |
| Hindelang et al. (1981) | – | .974 | .917, 1.035 | − .895 | .398 |
| Potdar and Koenig (2005) | 0 of 2 | .827 | .359, 1.907 | − .445 | .656 |
| Random model | | .971 | .915, 1.031 | − .953 | .341 |
| Personal interview (PI) vs. audio computer-assisted self-interview (ACASI) ($k = 1$) | | | | | |
| Potdar and Koenig (2005) | > PI (1 of 4) | 1.229 | .837, 1.804 | 1.052 | .293 |
| Self-administered questionnaire (SAQ) vs. computer-assisted self-interview (CASI) ($k = 10$) | | | | | |
| Beebe et al. (1998) | > SAQ (2 of 5) | 1.416 | .911, 2.202 | 1.545 | .122 |
| Knapp and Kirk (2003) | 0 of 3 | 1.112 | .591, 2.093 | .329 | .742 |
| Beebe et al. (2006) | 0 of 2 | 1.056 | .654, 1.705 | .224 | .823 |
| Brener et al. (2006) | > CASI (2 of 5) | .836 | .704, .991 | − 2.059 | .040 |
| Hamby et al. (2006) | > CASI (1 of 4)<br>> SAQ (1 of 4) | .934 | .493, 1.769 | − .208 | .835 |
| Lucia et al. (2007) | > CASI (2 of 40)<br>> SAQ (5 of 40) | 1.119 | .802, 1.561 | .663 | .507 |
| van de Looij-Jansen and de Wilde (2008) | > CASI (1 of 3) | .807 | .593, 1.099 | − 1.361 | .174 |
| Eaton et al. (2010) | > CASI (5 of 7) | .896 | .770, 1.043 | − 1.417 | .157 |
| Trapl et al. (2013) | 0 of 1 | 1.104 | .580, 2.102 | .301 | .764 |
| Baier (2017) | > SAQ (1 of 10) | .935 | .707, 1.236 | − .474 | .635 |
| Random model | | .919 | .841, 1.005 | − 1.852 | .064 |
| Self-administered questionnaire (SAQ) vs. audio computer-assisted self-interview (ACASI) ($k = 3$) | | | | | |
| Turner et al. (1998) | > ACASI (4 of 5) | .692 | .513, .932 | − 2.421 | .015 |
| Potdar and Koenig (2005) | – | 1.052 | .469, 2.359 | .124 | .902 |
| Trapl et al. (2013) | 0 of 1 | 1.144 | .598, 2.188 | .406 | .685 |
| Random model | | .819 | .591, 1.136 | − 1.196 | .232 |
| Computer-assisted self-interview (CASI) vs. audio computer-assisted self-interview (ACASI) ($k = 1$) | | | | | |
| Trapl et al. (2013) | 0 of 1 | 1.036 | .541, 1.986 | .107 | .914 |
| Self-administered questionnaire (SAQ) vs. automated touch-tone telephone (TACASI) ($k = 1$) | | | | | |
| Knapp and Kirk (2003) | 0 of 3 | 1.180 | .721, 1.930 | .659 | .510 |
| Computer-assisted self-interview (CASI) vs. telephone audio computer-assisted self-interview (TACASI) ($k = 1$) | | | | | |
| Knapp and Kirk (2003) | 0 of 3 | 1.061 | .543, 2.074 | .173 | .863 |
| Procedures of data collection | | | | | |
| Supervision by teachers vs. Supervision by researchers ($k = 2$) | | | | | |
| Walser and Killias (2012) | > teacher (2 of 22) | 1.042 | .828, 1.311 | .350 | .726 |
| Kivivuori et al. (2013) | > research (2 of 26) | .873 | .584, 1.305 | − .662 | .508 |
| Random model | | .998 | .817, 1..218 | − .024 | .981 |
| Non-anonymous vs. Anonymous ($k = 2$) | | | | | |
| Hindelang et al. (1981) | – | .978 | .921, 1.039 | − .705 | .481 |

**Table 2** (continued)

| Study | Comparison ($p < .05$) | OR | 95% CI | $z$ | $p$ |
|---|---|---|---|---|---|
| van de Looij-Jansen et al. (2006) | > Anonym. (3 of 4) | .673 | .514, .880 | − 2.888 | .004 |
| Random model | | .831 | .578, 1.196 | − .997 | .319 |
| No disclosure vs. Disclosure ($k = 1$) | | | | | |
| Beebe et al. (2006) | > No discl. (1 of 2) | 1.690 | .992, 2.877 | 1.931 | .053 |
| Home setting vs. School setting ($k = 1$) | | | | | |
| Brener et al. (2006) | > school (5 of 5) | .752 | .634, .892 | − 3.271 | .001 |
| "Conservative" interviewer vs. "Hip" interviewer ($k = 1$) | | | | | |
| Krohn et al. (1974) | > "Hip" (2 of 6) | .542 | .273, 1.076 | − 1.750 | .080 |
| No in-person follow-up vs. In-person follow-up ($k = 1$) | | | | | |
| King et al. (2012) | 0 of 1 | .619 | .362, 1.057 | − 1.758 | .079 |
| Bogus pipeline (BPL) vs. Control group ($k = 1$) | | | | | |
| Strang and Peterson (2016) | > BPL (2 of 8) | 2.176 | .815, 5.808 | 1.552 | .121 |
| Questionnaire design | | | | | |
| Response format: 2 options vs. 7 options ($k = 1$) | | | | | |
| Hamby et al. (2006) | > 7 options (2 of 4) | 1.185 | .625, 2.246 | .521 | .602 |
| Long vs. Short questionnaire ($k = 1$) | | | | | |
| Enzmann (2013) | > Short (5 of 24) | .888 | .744, 1.061 | − 1.305 | .192 |
| Standard vs. Month-by-month reporting ($k = 1$) | | | | | |
| Horney and Marshall (1992) | 0 of 8 | .978 | .688, 1.390 | − .126 | .900 |
| Reference period: "12 months" vs. "Since October 2003" ($k = 1$) | | | | | |
| Lucia et al. (2007) | 0 of 20 | 1.041 | .623, 1.740 | .154 | .878 |

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

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 self-reported offending collected under PI and SAQ (Krohn et al. 1974; Hindelang et al. 1981; Potdar and 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 = .971, 95% CI [.915, 1.031], $z = -.953$, $p = .341$).

### PI vs. ACASI

Only one study compared PI and ACASI (Potdar and Koenig 2005). Results mainly favored PI (OR = 1.229, 95% CI [.837, 1.804], $z = 1.052$, $p = .293$), 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 (Beebe et al. 1998, 2006; Knapp and Kirk 2003; Brener et al. 2006; Hamby et al. 2006; Lucia et al. 2007; van de Looij-Jansen and de Wilde 2008; Eaton et al. 2010; Trapl et al. 2013; Baier 2017). An analysis of the individual effect sizes showed that 5 comparisons favored CASI, though only one reached statistical significance with an OR of .836 (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 = .919, 95% CI [.841, 1.005], $z = -1.852$, $p = .064$), though with only marginal significance ($p < .10$).

### SAQ vs. ACASI

Three studies provided comparisons of offending behavior collected with SAQ or ACASI (Turner et al. 1998; Potdar and Koenig 2005; Trapl et al. 2013). One out of the three ORs presented statistically significant results in favor of the ACASI mode (OR = .692, $p < .05$). Considering random effects, the average effect size showed an OR = .819 favoring ACASI but with no statistical significance (OR = .819, 95% CI [.591, .136], $z = -1.196$, $p = .232$).

### CASI vs. ACASI

Trapl et al. (2013) conducted the sole experiment comparing self-reports of offending 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 = 1.036, 95% CI [.541, 1.986], $z = .107$, $p = .914$).

### 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 = 1.180, 95% CI [.721, 1.930], $z = .659$, $p = .510$).

### CASI vs. TACASI

Similar to the previous results, the experimental comparison between CASI and TACASI (Knapp and Kirk 2003) showed a nonsignificant effect size (OR = 1.061, 95% CI [.543, 2.074], $z = .173$, $p = .863$).

## Procedures of data collection

The second category of manipulations takes into account different procedures applied in the data collection that might influence the participants' self-reports of offending. 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 (Walser and Killias 2012; Kivivuori et al. 2013). 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 = .998, 95% CI [.817, 1.218], $z = -.024$, $p = .981$).

### Anonymity

From the pooled experiments, two studies focused on the issue of anonymity in self-reports of offending. 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 .978 ($p > .05$). 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 self-reports of offending in the anonymous condition (OR = .673, $p > .01$). 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 = .831, 95% CI [.578, 1.196], $z = -.997$, $p = .319$).

### 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 = 1.690, 95% CI [.992, 2.877], $z = 1.931$, $p = .053$).

### 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 = .752, 95% CI [.634, .892], $z = -3.271$, $p < .01$).

### 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 = .542, 95% CI [.273, 1.076], $z = -1.750$, $p = .080$).

### 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 = .619, 95% CI [.362, 1.057], $z = -1.758$, $p = .079$).

### Bogus pipeline

Finally, Strang and Peterson (2016) 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 = 2.176, 95% CI [.815, 5.808], $z =$

1.552, $p = .121$). Moreover, individual item comparisons revealed that men in the bogus pipeline condition showed 6.5 times greater odds of reporting illegal sexual assault (OR = 6.486, 95% CI [1.776, 23.688], $z = 2.829$, $p < .01$).

### 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 = 1.185, 95% CI [.625, 2.246], $z = .521$, $p = .602$). However, results varied considerably according to the types of crimes. Self-reports of psychological aggression (OR = .795, 95% CI [.309, 2.040]) and physical assault (OR = .771, 95% CI [.408, 1.456]) were slightly higher in the dichotomous condition, but not statistically significant ($p > .05$). For self-reports of sexual coercion (OR = 3.581, 95% CI [1.339, 9.575]) and injury (OR = 3.353, 95% CI [1.032, 10.889]), 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 = .888, 95% CI [.744, 1.061], $z = -1.305$, $p = .192$). 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 and 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 = .978, 95% CI [.688, 1.390], $z = -.126$, $p = .900$).

### 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 = 1.041, 95% CI [.623, 1.740], $z = .154$, $p = .878$).

## 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' self-reports of offending. 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 self-reports of offending in order to summarize the available information about measurement bias in criminology, providing evidence to improve data collection of self-reported offending.

We found that, contrary to other fields of sensitive questions (e.g., Richman et al. 1999; Tourangeau and Yan 2007), experimental research on self-reports of offending 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 self-reported offending, 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 self-reports of offending collected with PI was carried out in India (Potdar and 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 self-reported offending.

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., Tourangeau and Smith 1996; Thornberry and Krohn 2000), present findings on self-reports of offending 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 self-reports of offending 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 .836 (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 self-reports of offending, 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 self-reports of offending. 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 self-reported offending 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 statistically 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 self-reports of offending vary as a function of Standard vs. Month-by-month reporting and Reference period. In other words, we found no evidence that self-reports of offending are subject to telescoping, i.e., a memory distortion in which participants report events that occurred prior to the recall period (e.g., Loftus and 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 and 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 self-reports of offending, 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 and Peterson 2016), to delinquent behavior (e.g., Walser and Killias 2012), ever being in jail (Knapp and Kirk 2003), etc. We have also considered results referring to different recall periods, such as lifetime prevalence (e.g., Knapp and 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 self-reports of offending 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 self-reported offending 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 self-reports of offending 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 self-reports of offending? 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 self-reports of offending? 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 self-reports of offending.

**Funding** This study was conducted at the Psychology Research Centre (PSI/01662), School of Psychology, University of Minho, and supported by the Portuguese Foundation for Science and Technology and the Portuguese Ministry of Science, Technology and Higher Education (UID/PSI/01662/2019), through the national funds (PIDDAC).

The first author was supported by a doctoral grant from the Portuguese Foundation for Science and Technology (FCT - SFRH/BD/122919/2016).

## 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*, 562–580.

Baier, D. (2017). Computer-assisted versus paper-and-pencil self-report delinquency surveys: results of an experimental study. *European Journal of Criminology, 15*, 385–402.

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*, 221–238.

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.

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*, 623–632.

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*, 476–488.

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*, 382–392.

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*, 427–437.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Hoboken: John Wiley & Sons.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2014). *Comprehensive meta-analysis, version 3*. Englewood: 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*, 354–374.

Chaiken, J. M., & Chaiken, M. R. (1982). *Varieties of criminal behavior*. Santa Monica: Rand Corporation.

Chan, J. H., Myron, R., & Crawshaw, M. (2005). The efficacy of non-anonymous measures of bullying. *School Psychology International, 26*, 443–458.

Clark, J. P., & Tifft, L. L. (1966). Polygraph and interview validation of self-reported deviant behavior. *American Sociological Review, 31*, 516–523.

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*, 92–110.

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*, 441–450.

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*, 817–829.

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*. 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*, 439–467.

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*, 137–153.

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*, 147–177.

Farrington, D. P. (1973). Self-reports of deviant behavior: predictive and stable? *Journal of Criminal Law and Criminology, 64*, 99–110.

Gomes, H. S., Maia, A., & Farrington, D. P. (2018a). Measuring offending: self-reports, official records, systematic observation and experimentation. *Crime Psychology Review, 4*, 26–44.

Gomes, H. S., Maia, A. & Farrington, D. P. (2018b). *Method effects in measuring self-reported offending: an experimental comparison of personal, paper-and-pencil, and computer assisted interviews*. (Paper presented at the 74th Annual Meeting of the American Society of criminology, Atlanta, GA).

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*, 798–821.

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*, 869–890.

Grysman, B. & Johnson, C. (2010). *Effects of value affirmation on drug use disclosure in patients entering a community mental health center*. Doctoral dissertation, Hofstra University, Long Island, NY.

Hamby, S., Sugarman, D. B., & Boney-McCoy, S. (2006). Does questionnaire format impact reported partner violence rates?: an experimental study. *Violence and Victims, 21*, 507–518.

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*, 247–259.

Hindelang, M. J., Hirschi, T., & Weis, J. G. (1981). *Measuring delinquency*. London: 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*, 102–121.

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*, 1484–1493.

Jolliffe, D., & Farrington, D. P. (2014). Self-reported offending: reliability and validity. In G. J. N. Bruinsma & D. L. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice* (pp. 4716–4723). New York: Springer-Verlag.

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 and Adolescent Psychology, 41*, 508–515.

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*, 91–107.

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). New York: Springer.

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*, 117–134.

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*, 545–553.

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*, 501–528.

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). Bingley: Emerald.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality and Quantity, 47*, 2025–2047.

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*, 506–509.

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*, 150–168.

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 landmarkevents. *Memory and Cognition, 11*, 114–120.

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*, 39–64.

Maxfield, M. G., Weiler, B. L., & Widom, C. S. (2000). Comparing self-reports and official records of arrests. *Journal of Quantitative Criminology, 16*, 87–110.

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*, 565–587.

Nye, F. I., & Short, J. F. (1957). Scaling delinquent behavior. *American Sociological Review, 22*, 326–331.
Ong, A. D., & Weiss, D. J. (2000). The impact of anonymity on responses to sensitive questions. *Journal of Applied Social Psychology, 30*, 1691–1708.
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*, 526–556.
Porterfield, A. L. (1943). Delinquency and its outcome in court and college. *American Journal of Sociology, 49*, 199–208.
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*, 107–116.
Rehm, J., & Spuhler, T. (1993). Measurement error in alcohol consumption: the Swiss health survey. *European Journal of Clinical Nutrition, 47*, S25–S30.
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.
Schore, J., Maynard, R., & Piliavin, I. (1979). *The accuracy of self-reported arrest data*. Princeton: Mathematica Policy Research.
Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist, 54*, 93–105.
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*, 225–237.
Strang, E., & Peterson, Z. D. (2016). Use of a bogus pipeline to detect men's underreporting of sexually aggressive behavior. *Journal of Interpersonal Violence* (in press). 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.), *Criminal justice* (pp. 33–84). Washington, DC: 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). Washington, DC: National Academies Press.
Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: the impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60*, 275–304.
Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859–883.
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. New York: Cambridge University Press.
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*, 62–69.
Turner, C. F., & Miller, H. G. (1997). Monitoring trends in drug use: strategies for the 21st century. *Substance Use and Misuse, 32*, 2093–2103.
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*, 867–873.
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*, 1708–1721.
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*, 652–658.
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*, 17–28.
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.
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*, 153–168.
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*, 321–353.

**Hugo S. Gomes** is a Fulbright Scholar and a PhD Candidate, funded by the Portuguese Foundation for Science and Technology (FCT – grant SFRH/BD/122919/2016), in the School of Psychology, University of Minho, and a PhD visiting student at the Institute of Criminology, University of Cambridge. He is a Portuguese psychologist with a master's degree in Criminal Psychology (with an Exceptional Academic Achievement award). His research interests are in juvenile delinquency, validity of self-reports of offending, developmental and life-course criminology, and experimental criminology.

**David P. Farrington** is Emeritus Professor of Psychological Criminology at Cambridge University. He has received the Stockholm Prize in Criminology and he has been President of the American Society of Criminology. His major research interest is in developmental criminology, and he is the Director of the Cambridge Study in Delinquent Development, which is a prospective longitudinal survey of over 400 London males from age 8 to age 61. In addition to 788 published journal articles and book chapters on criminological and psychological topics, he has published 111 books, monographs and government publications, and 156 shorter publications (total = 1055).

**Ângela Maia** holds a PhD in Clinical Psychology from the University of Minho and is a Professor in the Department of Applied Psychology at the School of Psychology, University of Minho. She is the Vice President of the School of Psychology and President of the Pedagogical Council. She has collaborated and is the Principal Investigator of multiple research projects in the area of health, trauma, justice, and violence. She has published dozens of national and international publications on these topics.

**Marvin D. Krohn** is a Professor in the Department of Sociology and Criminology & Law at the University of Florida. His primary area of interest is criminological theory with a specific focus in developmental and life-course criminology. He was a Co-Principle Investigator on the Rochester Youth Development Study and continues to be affiliated with the project. He has published several research articles and co-authored or edited a number of books including Gangs and Delinquency in Development Perspective, the 2003 American Society of Criminology Michael Hindelang Award for Outstanding Scholarship. He received an Outstanding Mentor Award from the Academy of Criminal Justice Sciences and was named a Fellow in the American Society of Criminology.

## Affiliations

**Hugo S. Gomes**[1] · **David P. Farrington**[2] · **Ângela Maia**[1] · **Marvin D. Krohn**[3]

[1] Psychology Research Centre (CIPsi), University of Minho, School of Psychology - Campus de Gualtar, 4710-057 Braga, Portugal

[2] Institute of Criminology, Cambridge University, Cambridge, UK

[3] Sociology and Criminology & Law, University of Florida, Gainesville, FL, USA