Application of Randomized Response Models in Criminological Research

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Benedikt Iberl
aus Eberbach

Tübingen
2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	08.11.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Rolf Ulrich
2. Berichterstatter/-in:	Prof. Dr. Jochen Musch

Contents

Acknowledgements

Writing this dissertation as well as the underlying journal papers would not have been possible without the support and guidance of many people.

First and foremost, I would like to thank Rolf Ulrich, who accompanied my journey in the field of research throughout the last seven years as the supervisor of my Bachelor and Master theses, and now this dissertation. Since our very first meeting in the summer of 2017, I was met with nothing but kindness, openness, and support. I appreciate the opportunity to learn from and work with a truly admirable researcher, who never seems to run out of patience, scientific ideas, or interest in many fields!

I would also like to thank Jörg Kinzig, my second supervisor and superior at the Institute of Criminology. Many great differences between the fields of psychology and legal studies, from language to publishing practices to the dress code at scientific conferences, did not stand in the way of a cooperation that I perceived as respectful and fruitful from day one, and that I deem quite successful. I am grateful for the opportunity to work in the interesting and multifaceted field that is criminology, and to do so in a work environment I greatly appreciate.

I am especially thankful to Fabiola Reiber, for being a mentor and colleague who I could always turn to for advice. I very much appreciate your support which I hope to repay someday, not only because of how helpful it was, but also because it was solely motivated by collegiality and a willingness to help.

Furthermore, many thanks to Anesa Aljovic, for being such a hard-working and reliable co-worker and -author during your master thesis and in the writing of our shared paper.

Thanks to Sarah Schreier, for being such a dependable colleague since our time at the Institut of Criminology began, and especially for being willing to help in some language- and many non-language related matters, even if you do not have to.

For supporting me unconditionally in and out of work- and research- related matters, I am eternally grateful to my wife Kirstin, my parents Gabriele and Bernhard, my brother Max, and my friends, on whom I can always rely when things get difficult.

Last but not least, I would probably not have written this dissertation if it was not for the teaching and mentorship of Florian Wickelmaier, who unexpectedly managed to awaken my interest in statistics and quantitative data analysis during my studies.

Thank you!

Abstract

This dissertation addresses the questions of whether or not Randomized Response Models (RRMs) are suitable for use in criminological research and how RRM applications can be improved in general. RRM is the umbrella term for certain survey methods which were developed to minimize socially desirable response behavior. As socially desirable responses may especially occur when asking questions on sensitive topics, RRMs are — at least in theory — especially well-suited for such questions. This method could be particularly promising in the field of quantitative criminology, as topics such as crime, criminal prosecution, and victimization provide for a large number of sensitive topics.

To answer these research questions, a form of case study was conducted initially, and evidence for the existence of social desirability bias in a criminological survey was collected. Additionally, the Poisson model, a new method that enables particularly efficient measurement of a behavior's prevalence and rate of occurrence, was introduced to improve RRM applications. In an online survey on the prevalence of drinking and driving, the Poisson model was combined with one specific RRM, the Unrelated Question Model (UQM). This survey study revealed problems regarding the functionality of the UQM, a problem not uncommon for RRM applications. Furthermore, the extent to which comprehension aids during participant instruction can contribute to improving the validity of RRM applications was investigated. However, it remained unclear how important the comprehension of instructions actually is for the functionality of RRMs and whether comprehension aids do, in fact, improve the validity of such applications.

Overall, this dissertation illustrated that while RRMs provide for a promising methodological approach, the conditions for their optimal use have yet to be identified completely. Consequently, RRMs cannot yet be fully recommended for use in criminological research.

Zusammenfassung

Diese Dissertation befasst sich mit der Frage, ob Randomized Response Modelle (RRMs) für die Anwendung in der kriminologischen Forschung geeignet sind und wie RRM-Anwendungen allgemein verbessert werden können. RRM ist der Überbegriff für bestimmte Umfragemethoden, die entwickelt wurden, um sozial erwünschtes Antwortverhalten zu minimieren. Da sozial erwünschte Antworten insbesondere bei Fragen nach sensiblen Themen auftreten können, sind RRMs — zumindest theoretisch — insbesondere für solche Fragen geeignet. Im Bereich der quantitativen Kriminologie könnte diese Methode besonders vielversprechend sein, denn die Themenbereiche Kriminalität, Strafverfolgung und Viktimisierung bieten eine Vielzahl sensibler Fragestellungen.

Um die Forschungsfrage zu beantworten, wurde zunächst eine Art Fallstudie durchgeführt, in der Hinweise für das Vorliegen des Effekts der sozialen Erwünschtheit in einer kriminologischen Befragung nachgewiesen wurden. Für die Verbesserung von RRM-Anwendungen wurde zudem das Poisson-Modell vorgestellt, eine neue Methode, die eine besonders effiziente Messung von Verhaltensprävalenzen und -frequenzen ermöglicht. In einer Online-Befragung zur Häufigkeit von Alkohol am Steuer wurde das Poisson-Modell mit dem Unrelated Question Model (UQM), einem RRM, kombiniert. Dabei zeigten sich Probleme bei der Funktionsweise des UQM, die bei RRM-Anwendungen häufiger auftreten. Zuletzt wurde untersucht, inwiefern Verständnishilfen bei der Instruktion der Teilnehmenden dazu beitragen können, die Validität von RRM-Anwendungen zu verbessern. Dabei blieb unklar, wie wichtig das Instruktionsverständnis für die Funktionsweise von RRMs ist, und ob Verständnishilfen die Validität solcher Anwendungen verbessern.

Im Ergebnis zeigte sich, dass RRMs zwar einen vielversprechenden methodischen Ansatz verfolgen, die Rahmenbedingungen für deren optimalen Einsatz aber noch immer nicht vollumfänglich identifiziert wurden. Somit können RRMs noch nicht uneingeschränkt für den Einsatz in der kriminologischen Forschung empfohlen werden.

1 Introduction

One of the cornerstones in survey research is the assumption that participants respond honestly to the survey questions — at least most of the time. Only then can a survey lead to reliable and thus valuable results. Hence, it is a major problem for survey research if participants systematically respond in a way that is unreliable or dishonest. One systematic effect that causes such problems is the *social desirability bias*, which is defined as respondents' tendency towards answers that are in line with societal norms (Crowne & Marlowe, 1960; Edwards, 1953). This well-researched effect is especially prevalent in surveys on topics that participants perceive as sensitive or in cases where participants have to fear negative consequences when they respond honestly (e.g., Krumpal, 2013; Lee, 1993; Nederhof, 1985; Rasinski, Willis, Baldwin, Yeh, & Lee, 1999; Tourangeau & Yan, 2007). For instance, honest answers to survey questions about criminal behavior could include self-incriminating information. Participants who exhibited such behavior could fear legal, professional, or personal repercussions when responding honestly. Therefore, they may choose to give dishonest answers instead.

The overarching topic of this dissertation is the application of *Randomized Response Models* (RRMs) in criminological research. RRMs are specifically designed to lower social desirability bias in surveys on sensitive topics by guaranteeing complete and objective anonymity for participants. One recent example of an RRM application perfectly underlines why such methods can be very valuable for survey research: Shortly after Russia's invasion of Ukraine in 2022, researchers asked Russian citizens whether they supported the war (Chapkovski & Schaub, 2022). Since the participants might have had to expect repercussions for criticizing the authoritarian Russian government (e.g., Oliker, 2017), the researchers used an RRM to both protect the participants and to enhance the validity of their responses. This study is a prime example for the high value that RRMs could offer survey research in many fields, as many surveys on socially and politically relevant topics might benefit from an increased level of anonymity.

In the following sections, the origins and general functionality of RRMs will be explained specifically emphasizing the RRM which has been used in the research papers presented in this dissertation. This is followed by a short overview on the discipline of criminology and on possible areas of RRM applications in criminological research. Lastly, the objectives of this dissertation will be presented, transitioning to the second chapter, in which each of the four underlying papers will be summarized and discussed in detail.

1.1 Randomized Response Models and the Unrelated Question Model

The original RRM, the *Randomized Response Technique* (RRT), was introduced by Warner (1965). The main idea behind this method is to use statistical noise to mask the true answer of a respondent. For this purpose, the participant is asked to perform a random experiment using a certain randomization device (e.g., draw a card or roll a die), and to keep the outcome to themselves. Hence, this random experiment guarantees complete anonymity to the participants.

The RRT is depicted in the probability tree in Figure 1. First, each participant conducts the random experiment. The outcome, known only to them, determines whether the participant is supposed to respond to one of two yes/no-questions, A or B, that are both openly presented to them. With a probability of p, as represented by the upper branch in the tree, the randomization device will lead the participant to question A and with a probability of $1 - p$ — as shown in the lower branch — participants will be directed to question B. For instance, the participants could be told to roll a regular dice and to answer question A if they roll a 1 or 2, or question B if they roll a 3, 4, 5, or 6. In this example, $p = \frac{2}{6}$.

In the RRM, question A and question B are designed to be complementary to each other. Assume that a researcher wants to conduct a study on the lifetime prevalence of criminal behavior. In this example, question A could be "Have you ever committed a crime?", and question B "Have you never committed a crime?". As shown in the probability tree, the probability of a yes-answer to question A would be equal to the probability of a no-answer to question B, represented by the parameter π. The participant is meant to answer the question openly so that the researcher can observe whether they responded with "yes" or "no" . However, since the participant kept the outcome of the random experiment secret, it remains unknown to the researcher and any observers whether their yes- or no-response refers to question A or B. In our example, it will remain unclear to anyone but the participant whether they have ever committed a crime or not; thus, the participant's anonymity is guaranteed completely.

When numerous participants respond to an RRT survey in this way, it becomes possible to draw conclusions about the prevalence of the behavior (or opinion, etc., depending on the question) of interest. With a probability of p, set by design, and the observed proportion of yes-answers in the sample, γ,[1] it is possible to estimate π, which is the

[1] In most RRMs, the parameter representing the observed proportion of yes-answers is called λ. To avoid confusion with another parameter introduced later in a different model, it will be called γ here instead.

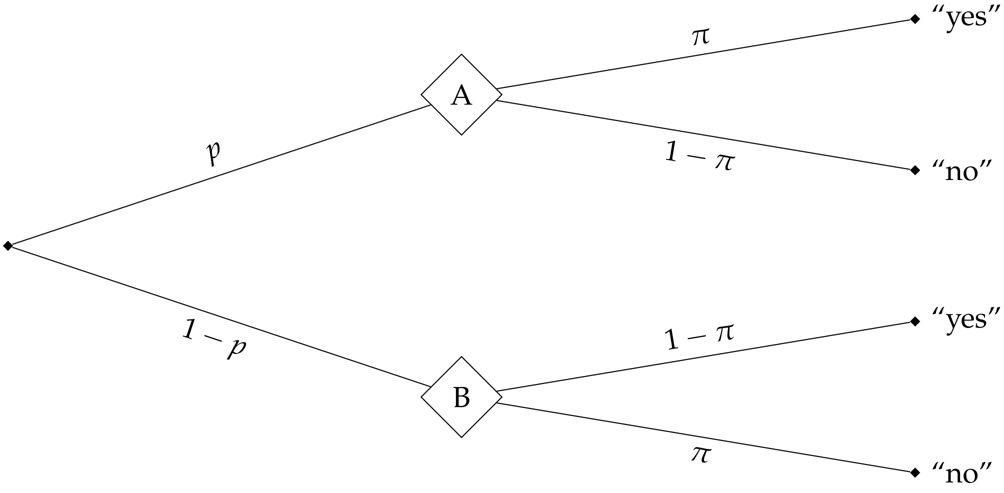
probability of responding "yes" to question A or "no" to question B. So, in the example at hand, the parameter π represents the proportion of people in the sample who have committed a crime before. Then, π can be estimated by

$$\hat{\pi} = \frac{(\hat{\gamma} + p - 1)}{2p - 1} \tag{1.1}$$

for $p \neq .5$, and with

$$\hat{\sigma}^2(\hat{\pi}) = \frac{\hat{\pi}(1 - \hat{\pi})}{n} + \frac{p(1-p)}{n(2p-1)^2}. \tag{1.2}$$

Figure 1: Probability tree of the RRT



Note. The sample is divided into two groups. First, respondents drawing question A, with a probability of p, and second, respondents drawing question B, which is formulated complementary to question A, with the counter probability of $1 - p$. Participants will respond "yes" to question A or "no" to question B, each with a probability of π. Respectively, participants will respond "no" to question A or "yes" to question B with the counter probability $1 - \pi$.

Since the introduction of the RRT by Warner (1965), multiple similar models have been proposed and tested. Most of these models build on Warner's basic idea of employing random experiments to mask the responses of single participants, while simultaneously allowing for prevalence estimation when numerous responses are collected. Some examples of rather well established RRMs are the *Unrelated Question Model* (UQM; Greenberg, Abul-Ela, Simmons, & Horvitz, 1969), the *Item Count Technique* (ICT; J. D. Miller, 1984) as used by Chapkovski and Schaub (2022), the *Forced Choice Model* (FCM; Boruch, 1971), or the *Cheater Detection Model* (CDM; Clark & Desharnais, 1998). Some other similar models do not rely on a randomization device and therefore, strictly speaking, are *non-Randomized Response Models*. The probably most widely used non-RRM is the *Cross-*

wise Model (CM) by>Yu2008. Instead of a randomization device, the participants are presented with two distinct questions in the CM, and are asked whether their answer is the same for both. One of these questions is chosen in a way that the probabilities of a yes- or no-answer can be estimated a priori. Even though non-RRMs do not utilize a random experiment, for the sake of simplification and readability, these models are subsequently included in the abbreviation RRM.

A common advantage of RRMs compared to traditional *Direct Questioning* methods (DQ) is the complete anonymity for participants that they provide. Said anonymity should theoretically lead to more valid responses in surveys on sensitive topics where social desirability bias might influence participants' responses in DQ surveys. In fact, some studies not only suggest that this supposed advantage of RRMs does exist in surveys on sensitive topics, but also that this advantage gets bigger the more delicate the respective questions are (Lensvelt-Mulders, Hox, Van der Heijden, & Maas, 2005). However, due to the additional statistical noise added by the randomization device — or the second question in the CM (Yu, Tian, & Tang, 2008) — RRMs are less efficient than DQ. As a result, bigger samples are needed for RRMs in order to achieve the statistical power of DQ applications (Ulrich, Schröter, Striegel, & Simon, 2012).

Ulrich et al. (2012) showed that the UQM has some desirable statistical characteristics, and that it is more efficient than other RRMs. Additionally, the UQM's instructions are rather simple and do not require participants to lie (other than, e.g., the FCM by Boruch, 1971), which they might be reluctant to do. Therefore, the UQM is assumed to have a decent level of psychological acceptability for participants (Höglinger, Jann, & Diekmann, 2016; Reiber, Pope, & Ulrich, 2020; Reiber, Bryce, & Ulrich, 2022; Ulrich et al., 2012). For these reasons, the UQM was used as the "RRM of choice" in this dissertation and the underlying research papers.

Figure 2 shows the probability tree of the UQM as proposed by Greenberg et al. (1969). The similarity to the RRT becomes apparent when comparing both trees (see Figure 1). As in the RRT, the random experiment is conducted first and its outcome kept secret by the participants. Participants draw the question A with a probability of p, as illustrated in the upper branch of the probability tree, and question B with a probability of $1 - p$, as shown in the lower branch. In contrast to the RRT, question B is not formulated complementary to question A in the UQM. Instead, question A is the only *sensitive question*, as represented by the S-node in Figure 2, while question B is a *neutral question*, as represented by the N-node.

Similar to the procedure in the RRT, participants who drew the sensitive question are assumed to answer "yes" with a probability of π, and "no" with the counter probability $1 - \pi$. Along with the neutral question, a new parameter, q, is introduced, which describes the probability of a participant to answer "yes" to this question. The sensitive question is referring to the behavior (or opinion) of interest, e.g., "Have you ever committed a crime?". Meanwhile, the neutral question is chosen in a way that the probability q can be estimated a priori. As the name suggests, the neutral question is not sensitive in nature so that it should not be influenced by social desirability bias. For instance, this question could refer to the date of birth of the participants, e.g., "Is your birthday in the first half of the year (before July 1st)?". Then, it can be assumed that $q \approx .5$. Even though dates of births are not exactly uniformly distributed, this approximation is precise enough for the purpose of the UQM (Ulrich et al., 2012). Similar questions are also frequently used as the randomization device in UQM applications (e.g., Dietz et al., 2018; Meisters, Hoffmann, & Musch, 2020; Reiber et al., 2022; Ulrich et al., 2018).

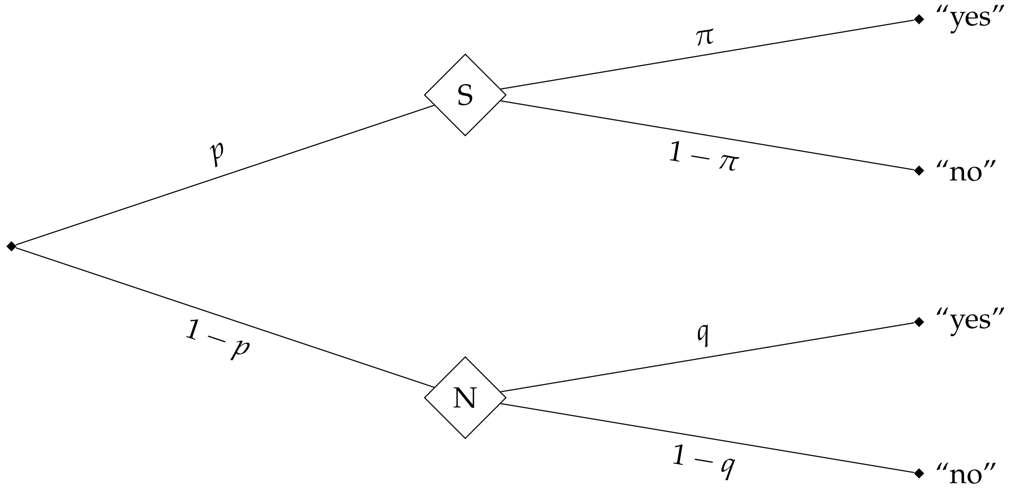
As in the RRT, observation of the proportion of yes-answers in a sample of numerous participants allows for estimation of the parameter π, with

$$\hat{\pi} = \frac{\hat{\gamma} - (1-p) \cdot q}{p} \tag{1.3}$$

and

$$\hat{\sigma}^2(\hat{\pi}) = \frac{\hat{\gamma} \cdot (1-\hat{\gamma})}{n \cdot p^2}. \tag{1.4}$$

Figure 2: Probability tree of the UQM



Note. The sample is divided into two groups. First, respondents drawing the sensitive question (S), and second, respondents drawing the neutral question (N), with the probabilities of p and $1 - p$, respectively. With the probability π, participants who drew the sensitive question will respond "yes" . The probability of a no-response for these participants is the counter probability $1 - \pi$. The probabilities of respondents who drew the neutral question to answer "yes" or "no" are q and $1 - q$, respectively.

RRMs have been used to research many different sensitive topics. Examples — in no particular order — are surveys on the prevalence of (non-)voting behavior (Moshagen, Musch, & Erdfelder, 2012), (non-)compliance with medication (Ostapczuk, Musch, & Moshagen, 2011), intimate partner violence (Reiber et al., 2022; Moshagen et al., 2012), doping in elite athletics competitions (Ulrich et al., 2018), pharmacological neuroenhancement in students (Dietz et al., 2018), drug users (Goodstadt & Gruson, 1975), tax evasion (Houston & Tran, 2001; Musch, Bröder, & Klauer, 2001), sexual violence (Soeken & Damrosch, 1986), illegal resource use, e.g., poaching, in a Ugandan national park (Solomon, Jacobson, Wald, & Gavin, 2007), or — as previously mentioned — surveys on the prevalence of Russians with a negative opinion towards the invasion of Ukraine (Chapkovski & Schaub, 2022). From these examples alone, it becomes apparent that RRMs are often applied in the context of researching criminal behavior or victimization. This might be obvious, since crime is historically connected to the violation of societal norms. Talking openly about committed crimes is socially undesirable and can even have legal consequences. On the other hand, sharing information about personal victimization might be very intimate or even be perceived as shameful, and is often underreported in surveys (Krumpal, 2013). So, surveys on crime and victimization can be assumed to be susceptible to social desirability bias, and should therefore be well suited for RRM applications. Such applications inevitably overlap with the field of criminology.

1.2 Criminology: A (very) short overview

While a universally accepted definition of criminology does not exist, crime and criminal behavior are the central foci of this distinct discipline, as its name already suggests (Hagan, 1986). The famous sociologist and criminologist Edwin Sutherland defined criminology as the "body of knowledge regarding crime and delinquency as social phenomena", including "the processes of making laws, breaking laws, and reacting to the breaking of laws" (Sutherland, Cressey, & Luckenbill, 1992, p. 3). Sutherland refers to crime as "certain acts regarded as undesirable" that are thus "defined by the political society as crimes" (Sutherland et al., 1992, p. 3). When such acts are committed, the "political society reacts by punishment, intervention, and prevention" (Sutherland et al., 1992, p. 3). According to Hagan (1986), the main topics of interest in criminology are criminal behavior, juvenile delinquency, and victimization, as well as the etiology and the sociology of crime. Furthermore, criminology overlaps significantly with the field of criminal justice (Hagan, 1986).

Criminology is an empirical and interdisciplinary field of research that is influenced by sociology, psychology, legal studies, as well as educational and political science (Neubacher, 2020). In the Anglo-American context, criminological research is traditionally conducted by sociologists and psychologists (Neubacher, 2020). In Germany, however, criminological research is more heavily influenced by legal studies, with most criminological research facilities being located at faculties of law (Neubacher, 2020). Overall, the distinction between criminology and criminal justice as separate disciplines is less common in Germany, and topics such as policing and crime control are central elements within German criminological research (Neubacher, 2020).

Historically, criminology has used a wide variety of empirical research methods. For instance, Cesare Lombroso, one of the first empirical scientists who systematically researched the etiology of crime in the 19th century, used the "methods" of craniometrics to identify "born criminals" (Becker, 2002; Hagan, 1986; Neubacher, 2020). To name other examples, the famous *Broken Windows Theory* builds on a field experiment by Philip Zimbardo (Wilson & Kelling, 2017), and some of the very influential *Chicago School*'s criminological studies rely on the quantitative analysis of spatial crime data (Snodgrass, 1976). In modern times, most standard qualitative or quantitative research methods have been established as tools of criminological research (Neubacher, 2020).

As Sutherland's definition indicates, social desirability plays a key factor in criminology (Sutherland et al., 1992). It is therefore reasonable to assume that RRMs, designed to circumvent social desirability bias in surveys, could be relevant tools for criminological research. As described above, RRMs have been used in criminologically relevant research as well. However, none of the cited RRM studies on criminologically relevant topics have been conducted by criminologists or in a criminological research facility. It seems that despite their apparent suitability, RRMs are far from being established as a standard method in criminological research.

1.3 Objective: Randomized Response Models in criminological research

As has been established in the section before, criminology relies on a variety of empirical research methods. In terms of quantitative methods, this primarily includes survey and prevalence research (Neubacher, 2020). With sensitive topics such as criminal behavior and victimization being of central interest in this field of research, minimizing social desirability bias should be an important concern for quantitative criminologists. RRMs could thus offer an interesting alternative to traditional methods in criminological survey research. Still, so far RRMs are barely employed in criminology.

Considering this, the guiding research question of this dissertation was whether RRMs are suitable for quantitative criminological survey research. The dissertation is based on four papers that have been published in peer-reviewed journals and that revolve around topics that contribute to an answer to this question. The papers refer to the

UQM as the RRM of choice, under the assumption that the results would be somewhat transferable and indicative of RRMs in general.

Paper 1 focused on a criminological research project on plea bargaining in Germany and targeted the question of whether the UQM, in principle, is suited for criminological research. Since the limited ways to formulate questions in RRMs can be suspected to contribute to their rare use in and outside of criminological research, Papers 2 and 3 propose a novel method of prevalence measurement and combine this method with the UQM. As insufficient comprehension of instructions might be another possible issue of RRMs that prevents more practical applications, Paper 4 was based on a survey study evaluating a possible effect of comprehension checks in the UQM. In the following, each paper will be summarized and discussed. Afterwards, in order to answer the central research question, the results of all four papers will be integrated in a general conclusion.

2 The UQM as an alternative for traditional methods

[2]The question of whether RRMs could theoretically be valuable for criminological research might seem trivial at first. Nonetheless, this theoretical question appears to be thoroughly under-researched (one rare example is a short theoretical article on methods of (online-)survey research methods in criminological research by Treibel & Funke, 2004). It thus seemed reasonable to approach this question systematically. In essence, two major conditions have to be met in order for RRMs to have the potential to be a promising alternative to traditional survey research methods. The first condition is that the researched topic should be sensitive in nature — or for the surveyed population —, so that participants might be susceptible to responding in a socially desirable way. If there is no risk of social desirability bias, the costs of RRMs in terms of additional sample size and statistical noise outweigh the benefits. This directly implies the second condition: RRMs are a valuable option only if it can be expected that a certain sample size is available.

Considering the role of these two conditions, Paper 1 analyzed whether both conditions were met in a specific criminological research project that was used as a case study. This project was previously conducted at the Institute of Criminology in Tübingen. The original survey data which used traditional direct questioning was re-analyzed in terms of the research question at hand. The goal of this form of case study was to determine a) whether a social desirability bias might have occurred in the original study, and b) whether the original sample was theoretically suitable for UQM application.

2.1 Need for anonymity in criminological survey research

The original project the article was based on was initiated by the German Federal Ministry of Justice ("Bundesministerium der Justiz", BMJ). This research project evaluated a German law from 2009 that regulates plea bargaining in criminal cases ("Verständigung im Strafverfahren"). The project's main research question was to determine

[2] This chapter is based on Paper 1 of the four underlying publications (Iberl & Kinzig, 2022), see Appendix A.

whether the legal regulations were sufficiently adhered to in practice.[3] Starting point of the project was a decision by the German Constitutional Court ("Bundesverfassungsgericht", BVerfG) in 2013. It ruled that the plea bargaining law is in principle conforming to constitutional law — however, it suspected that violations of the respective regulations might be widespread and obliged the Ministry of Justice to continuously evaluate how closely the plea bargaining law is complied with in practice. Therefore, the Ministry of Justice commissioned a research alliance of three teams from the University of Frankfurt/Main, University of Düsseldorf and the Institute of Criminology at the University of Tübingen to carry out this evaluation. Using different methodological approaches, the research alliance published its findings in 2020 — with the main result being that illegal plea bargaining was still occurring in Germany on a regular basis (Altenhain, Jahn, & Kinzig, 2020).

In this context, the Institute of Criminology conducted a quantitative online survey, interviewing 1,567 judges, defense attorneys and prosecutors (Kinzig, Iberl, & Koch, 2020). A surprising result of this survey was that the answers of the participants varied significantly depending on their occupation: According to defense attorneys, illegal plea bargaining was much more prevalent than according to judges and prosecutors. While the assessments of judges and prosecutors were quite similar overall, judges reported the least violations. For example, when asked about how often they engaged in illegal plea bargaining, 45.5% of defense attorneys selected the option "often" or "very often". For prosecutors and judges, this percentage was only 15.5% and 11.1%, respectively. Similar response patterns were found throughout the whole 46 items of the survey: Defense attorneys selected the socially undesirable options considerably more often, while prosecutors and judges apparently tended more towards socially desirable answers.

Hence, social desirability bias might be a possible explanation for the response pattern at hand. After all, illegal procedures in legal practice, including illegal plea bargaining, are most probably viewed negatively by society. Besides, the respective questions should be perceived as sensitive by legal practitioners, since violations of plea bargaining regulations could have serious professional and/or legal consequences for them. As previously mentioned, expecting such repercussions might lead to socially desirable responding (e.g., Krumpal, 2013; Rasinski et al., 1999; Lee, 1993; Tourangeau & Yan, 2007). The strength of this effect can apparently vary between different populations (Tourangeau & Yan, 2007); so, judges and prosecutors might be more susceptible to social desirability bias than defense attorneys.

Naturally, the tendency of judges — and, to a lesser extent, prosecutors — towards

[3] It is questionable how fitting the term of *plea bargaining* is in the German context, since the history and importance of this concept is significantly different in Germany compared to Anglo-American countries (Hodgson, 2015; Langbein, 2022). While this term describes a consensual alternative to criminal proceedings in both contexts, the possibilities in the German "Verständigung" are much more restricted compared to plea bargaining in other countries, e.g., the United States.

more socially desirable response options is not in itself proof of the existence of social desirability bias. Within this study, however, there were some additional indicators that point towards social desirability bias playing an important part in the response behavior of the participants.

First, the varying responses between professions regarding the prevalence of illegal plea bargains hint towards the existence of some kind of bias. This is due to the fact that all three professional groups are required in order for a plea bargain to come into effect. Plea bargains, legal or illegal, are only effective when all parties agree to them. Since the judges are responsible for the verdicts in German law, excluding the judge from a plea bargain would be rather pointless as it would for example not guarantee the lower sentence the defense might aim for. On the other hand, excluding either the prosecution or the defense from any bargain would carry the risk of the excluded party to appeal the judgement, moving the trial to a higher court. As a consequence, one would expect all professions to participate more or less equally frequently in illegal and legal plea bargaining. That the resulting response pattern differs greatly between professions is therefore illogical in itself — the responses have to be influenced by some kind of third factor. This might be a first hint towards the existence of social desirability bias. Still, other explanations are possible as well: Perhaps there is a disconnect between defense attorneys and the other groups regarding their understanding of the relevant regulations, causing defense attorneys to wrongfully interpret some bargains as illegal. Such alternative explanations cannot be ruled out entirely. However, more results point towards a social desirability bias influencing the results.

Second, some substantial differences were found between the judges' and prosecutors' self-perception and the perception of those professional groups by the respective others. In one question of the survey, participants were asked how often illegal plea bargains were initiated by the prosecution, the defense, the defendant, or the court. The results showed that judges, prosecutors, and defense attorneys all assumed a very similar prevalence of plea bargaining initiatons by the defense. However, both judges and prosecutors rated initiations of illegal plea bargains by their own professional group as less prevalent than members of the other professions. This pattern was most clear with respect to illegal plea bargains initiated by the courts: Less than 30% of judges but around 50% of prosecutors and 60% of defense attorneys stated that such initiations were "frequent" or even "very frequent". Such differences between self-perception and the assessments of third parties could be a hint towards socially desirable responding.

Third, the participants' responses differed significantly between professional groups when queried about the perceived risk of professional and/or criminal consequences due to illegal plea bargaining. The surveyed judges rated the risk of both consequences as greater than defense attorneys and prosecutors. Once again, defense attorneys responded in a socially undesirable way most often, rating both risks lower than members of the other professions. As previously explained, the fear of negative consequences to honest answers seems to play an important role in the occurrence of social desirability bias (e.g., Krumpal, 2013; Rasinski et al., 1999; Lee, 1993; Tourangeau & Yan, 2007). Judges and prosecutors apparently fear negative consequences more than defense attorneys, which could help explain the suspected social desirability bias.

Fourth, judges and prosecutors chose evasive response options, such as "don't know", considerably more often across all questions (14.3 times and 13.6 times, respectively) than defense attorneys (6.8 times). This could be explained by judges and prosecutors feeling more anxious, or less anonymous, or because they perceive the questions as more delicate compared to defense attorneys. While other explanations come to mind as well — maybe the defense attorneys, who are not professionally bound to certain court districts or regions, are able to answer more questions on average because their professional day-to-day varies more than the one of a judge or prosecutor —, this can be interpreted as another hint towards the existence of a social desirability bias, especially in combination with the aforementioned findings.

Why prosecutors and especially judges might have perceived the survey as more sensitive than defense attorneys can be explained quite easily. For one, some respondents might have considered the possible indirect consequences that this survey could cause — after all, the survey was part of the evaluation of the plea bargaining law commissioned by the Federal Ministry of Justice. Most of these possible consequences, e.g., stricter regulations or control mechanisms, would probably have the biggest impact on judges and, to a lesser extent, on prosecutors. After all, judges in Germany speak the sentence in criminal law and are thus mainly responsible for guaranteeing the lawfulness of criminal proceedings. Prosecutors on the other hand were named as "guardians" of the plea bargaining regulations by the German Constitutional Court. A high prevalence of illegal plea bargaining could therefore be interpreted as a failure of the prosecution to fulfill this role. Additionally, both of these professional groups have been famously understaffed in recent years (e.g., Koerth, 2019). The mentioned possible indirect consequences would probably lead to even more work for these groups, not necessarily accompanied by additional staffing. Lastly, contrary to defense attorneys, judges and prosecutors are servants of the state, and are very well respected in society (forsa, 2023) — which is why the public expectations towards judges and prosecutors are (understandably) high. Meanwhile, defense attorneys can be assumed to be motivated by economic factors more often. Other than judges and prosecutors, defense attorneys do not represent the state, but their clients' interests in the court of law — usually in opposition to the authorities. That said, it seems natural that public expectations and social and professional norms differ substantially between defense attorneys on the one, and judges and prosecutors on the other hand. This could have led to judges and prosecutors perceiving the survey on plea bargaining as more sensitive than defense attorneys, leading to the suspected social desirability bias.

In conclusion, social desirability bias seems to be a very likely explanation for the varying results between the professional groups, with judges being the most susceptible to socially desirable responding. Hence, the first condition for the usage of RRMs, which is the possibility of response bias due to questions being perceived as sensitive, was apparently fulfilled in this survey study.

2.2 Fulfillment of the mathematical requirements of RRMs

Besides the sensitivity of the survey questions, some mathematical conditions need to be met in order for RRMs to be an improvement to traditional methods. As explained above, a major drawback of RRMs is the greater statistical noise, which means that additional efforts have to be made to reach the same statistical power as traditional methods. Generally, this is achieved through bigger sample sizes.

Besides, it is necessary to mention that most RRMs are based on the assumption that the randomization device and higher level of anonymity lead to 100% honest responses. However, some studies showed that even with the level of anonymity offered by RRMs, not all respondents answer truthfully. Certain RRMs specialized on this issue, introducing additional model parameters which enable the estimation of the prevalence of *cheaters*, i.e., respondents who do not answer honestly, but opt for self-protective answers instead (e.g., Clark & Desharnais, 1998; Reiber et al., 2020).

So, while big samples are a necessary condition for successful RRM applications, instruction adherence might be important as well. In the paper at hand, both factors were taken into account to answer the question of whether the survey on the practice of plea bargaining would have been well-suited for UQM application from a mathematical standpoint.

To start off, we carried out multiple simulations of UQM applications based on the judges' answers and sub-sample size. The central variable of the simulations was the estimate of lifetime prevalence of illegal plea bargaining in the judges' personal practice (29.4%, or $\pi_0 = .294$). The main focus was to test under which circumstances there would be a significant difference between this estimate and the estimate of a UQM application using the parameters $q = .5$ and $p = .67$. To do so, we calculated the power of the underlying statistical tests (one-sided 95%-confidence intervals) and manipulated the values for the sample size N, the true prevalence π_s, and the proportion of self-protective answers (0 to 60%). Additionally, we compared the performance of the UQM to that of the UQMC, the *cheating extension* to the UQM (as proposed by Reiber et al., 2020).

As a result, we found that a UQM application would have been feasible given the sample size of the study. Even under the assumption of a considerable amount of cheaters, the UQM performed quite well in many scenarios. For instance, with 20% cheaters and a true value of $\pi_s = .587$ (lifetime prevalence of the whole original sample for "illegal plea bargaining by hearsay"), a significant difference to the estimate of $\pi_0 = .294$ could have been detected with a sample of $N = 200$ and with a statistical power of $1 - \beta = .8$. As another example, it would have been possible to detect a difference to the original estimate in the original sample of judges ($N = 558$) even if there were 10% cheaters and if the true value for illegal plea bargaining in the judges' practice were $\pi_s \approx .47$, with a power of $1 - \beta = .9$. Unsurprisingly, the UQMC was superior to the UQM only in simulated cases where the proportion of cheaters was high. In such cases, the UQMC needed bigger sample sizes than those available in the original study to reach satisfactory power levels. However, this is more of an indicator for the UQMCs complexity compared to traditional methods and simpler RRMs than it is an indicator for the superiority of traditional methods — such high proportions of protective answers would render any DQ result invalid regardless of the available sample size.

Overall, the case study showed that a UQM application would be suited as an alternative to the traditional approach used in the underlying study for measuring the prevalence of illegal plea bargaining from judges' survey responses.

2.3 Discussion

As shown in the presented Paper 1, two main conditions for successful UQM application, namely the risk of a social desirability bias occurring and a sufficient sample size, were met in the underlying survey study on legal practitioners. Thus, it would have been promising to apply the UQM in the original survey, or to do so in similar future surveys. While this case study cannot claim to be representative for the entirety of criminological survey research, it allows for the conclusion that RRMs possess the potential to be a genuine alternative to traditional methods in some criminological studies. First, in the context of criminological research there undoubtedly exist countless other sensitive topics and questions besides illegal plea bargaining. Therefore, it seems plausible that socially desirable response behavior might influence many criminological survey studies. Second, many criminological research projects have access to rather big samples that should be sufficient for RRM applications (e.g., in the studies of Birkel, Church, Erdmann, Hager, & Leitgöb-Guzy, 2020; Dreißigacker & Riesner, 2018; Ellrich & Baier, 2015; Lutz, Stelly, Bartsch, Thomas, & Bergmann, 2021; Treibel, Dölling, & Hermann, 2017; Kerner, Stroezel, & Wegel, 2011).

Despite their theoretical suitability for criminological research, RRMs are rarely used in this field. As already mentioned, there exist some studies in which RRMs were used to research criminologically relevant topics (e.g., Dietz et al., 2018; Reiber et al., 2022; Goodstadt & Gruson, 1975; Houston & Tran, 2001; Musch et al., 2001; Soeken & Damrosch, 1986; Solomon et al., 2007; Wolter, 2012; Ulrich et al., 2018). However, none of the cited studies was conducted by criminological research facilities. Thus, the question remains why RRMs are not used more often in criminological survey studies when they should in many cases be both mathematically and theoretically suitable for application.

One possible answer to this question is that RRMs might simply not be well known in criminology. This would come to no surprise, as RRMs are generally not widely used in comparison to more traditional methods. Especially in Germany, where criminology is strongly rooted in criminal law and legal studies (Neubacher, 2020), the rare use of non-standard quantitative methods is to be expected.

Another possible explanation could lie in the complexity of RRMs which goes hand in hand with some impractical properties. First, the large sample sizes needed for sufficient statistical power come to mind. Even though lots of criminological research projects have access to big samples, the additional statistical noise in RRM applications is considerable. Second, using RRMs instead of DQ in surveys is more complex not only for the researchers but also for the respondents. This leads to increased efforts and amounts of time spent for both sides. Third, while they enable completely anonymized questioning, most RRMs reduce this possibility to simple yes/no-questions. This naturally limits the information that can be gained from such RRM applications.

Both RRMs not being well known and their inherent impracticabilities are probably relevant factors for their rare use in criminology. That many RRM applications are designed to be limited to simple yes/no-questions might be especially important in this context, as it significantly reduces the information one can gain from asking questions in RRM designs.[4] An example for this drawback of RRMs can be found in the discussed Paper 1: In the underlying study, the respondents were asked about the frequency of certain behaviors (e.g., illegal plea bargaining) and could answer by selecting options such as "very frequent", "frequent", "rarely", or "never". However, for the post-hoc simulations regarding UQM applications, a simplification was necessary which is why the lifetime prevalence (including the options "very frequent", "frequent" and "rarely") was derived from the original results. Clearly, some information is lost here.

Therefore, a researcher applying classical RRMs with yes/no-questions generally has to consider very carefully how they phrase those questions to gain a maximum of information. It gets particularly difficult when one wants to measure the prevalence of

[4] Some RRMs tackle this problem, enabling anonymized question formats with multiple response options; however, they are generally even more inefficient than classical RRMs (e.g., De Jong, Pieters, & Fox, 2010; Greenberg, Kuebler Jr, Abernathy, & Horvitz, 1971).

certain events rather than, e.g., opinions, since a question inferring about the occurrence of an event always refers to a certain time frame. For example, one could ask whether the participant shoplifted within a specific time frame, e.g., in the last week, month, year, etc. Alternatively, one could only ask about the lifetime prevalence of the behavior of interest, e.g., whether the participant has ever shoplifted before. However, the concept of lifetime prevalence is rather vague and does not yield reliable information about the current situation (see e.g., Fiedler & Schwarz, 2016). Although this is not a problem exclusive to RRMs, but to prevalence research in general, a solution to this problem might especially benefit RRMs, as their reliance on yes/no-questions cannot be circumvented easily. In the following, such a potential solution is presented and then applied to the UQM.

3 An overarching problem: Prevalence estimation using time-constrained yes/no-questions

[5]Most RRMs rely on yes/no-questions. When measuring prevalence of events or behavior with such a question, one has to select a time frame which the question refers to, for example: "Have you committed shoplifting during the last week?". In the following, such questions will be called *time-constrained yes/no-questions*.

Not only RRMs rely on this type of question; time-constrained yes/no-questions are generally widely used in survey research (e.g., Andrie et al., 2019; Beck et al., 2021; Birkel et al., 2020; Burr, Butland, King, & Vaughan-Williams, 1989; Han, Compton, Gfroerer, & McKeon, 2015; Isolauri & Laippala, 1995; Linton, Hellsing, & Halldén, 1998; McCabe, Cranford, & Boyd, 2006; McKetin, McLaren, Lubman, & Hides, 2006; Sawyer, Smith, & Benotsch, 2018; Virudachalam, Long, Harhay, Polsky, & Feudtner, 2014; Wittenberg, Reinecke, & Boers, 2009). The popularity of such questions might be rooted in their simplicity: Yes/no-questions are both very easily understood and quickly answered by respondents. Plus, they result in simple, dichotomous data which is easy to analyze and interpret.

However, time-constrained yes/no-questions come with one clear disadvantage: the limited information one can gain from them. In general, when conducting prevalence research, two kinds of information are central. First, one wants to find out *how many people* engage in a certain behavior — for example, how many people are shoplifters. Second, one wants to know *how often* those people exhibit the behavior of interest — so, in our example, how frequently a shoplifter shoplifts. Hence, one is interested in both the *proportion* of shoplifters and in the *rate* of shoplifting offenses committed by them.

The limited information of time-constrained yes/no-questions is relevant to the measurement of both the proportion of *trait carriers*, and the rate of a given behavior. For instance, Wittenberg et al. (2009) conducted a longitudinal study on youth delinquency in Germany. Among other questions, the authors used a time-constrained yes/no-question to ask respondents whether they shoplifted within the last year. Their 2006 sample from

[5] This chapter is based on Paper 2 of the four underlying publications (Iberl & Ulrich, 2023), see Appendix B.

the city of Duisburg reported a past-year prevalence of 6% for shoplifting. Obviously, this percentage in itself does not yield any information about the rate of shoplifting. The information on the proportion of shoplifters is ambiguous as well: There might be shoplifters in the sample who did not shoplift in the past year.

Naturally, there exist different methods to solve both problems, i.e., to enable the measurement of the proportion of trait carriers and of the underlying rate. For example, to measure the proportion of trait carriers, one could simply ask participants whether they would describe themselves as shoplifters, without referring to a certain time frame. This could be problematic — not only due to the risk of social desirability bias but also because participants' individual definitions of "being a shoplifter" might differ and be quite subjective.

Regarding the rate, one could simply ask participants how often they shoplifted during a certain time frame, e.g., "How often have you committed shoplifting during the last week?" (this method is widely used in prevalence research as well, see e.g., Cullen et al., 2018; C. Miller et al., 2020; Molinaro et al., 2018; Seitz, Rauschert, Atzendorf, & Kraus, 2020; Soga, Evans, Tsuchiya, & Fukano, 2021). While this method is certainly suited for the measurement of rates, it still faces the problem of how to measure the proportion of trait carriers. Additionally, compared to simple yes/no-questions, such questions are probably harder and more time-consuming for participants to answer, because they have to recall multiple events instead of only one.

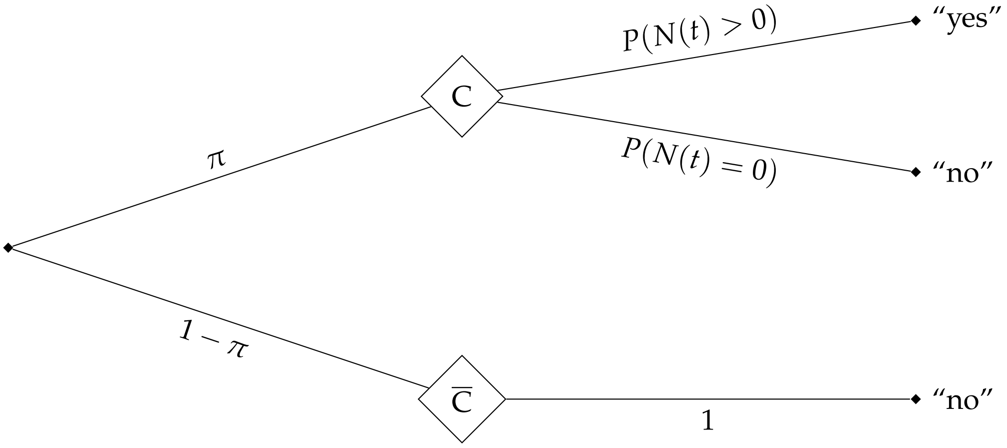
Considering these issues, Paper 2 introduced a novel method that allows for both the measurement of the proportion of trait carriers and the rate of the underlying behavior while still employing straightforward time-constrained yes/no-questions.

3.1 The Poisson model as a possible solution

The in Paper 2 proposed solution to the problem of ambiguity inherent to time-constrained yes/no-questions is a statistical model based on a Poisson process. The core assumptions of the model are twofold: The behavior in question occurs approximately independent of time and is equally distributed across all trait carriers. Considering the aforementioned example of shoplifting, this would mean that the probability of shoplifting within a certain time frame is equal for all shoplifters. When these assumptions are met, the behavior of shoplifting can be modeled via the Poisson distribution.

The precondition for an application of this *Poisson model* is to divide the participants into at least two groups. Each group gets a yes/no-question referring to the same behavior of interest. However, the time frame t the questions refer to is varied between groups. For instance, with three groups, the question for group 1 could be: "Did you shoplift during the *past week*?", the question for group 2, "Did you shoplift during the *past month*?", and the question for group 3, "Did you shoplift during the *past three months*?". Or, mathematically speaking, with the unit of t being weeks, one would define $t_1 = 1$, $t_2 = 4$, and $t_3 = 12$. In this way, simple yes/no count data is collected for each of the groups.

Figure 3: Probability tree of the proposed Poisson model



Note. The sample is divided into carriers C and non-carriers $\overline{C}$ by the parameter π, describing the probability of a random participant being a carrier of the researched attribute. Non-carriers answer "no" with a probability of 1. Carriers answer "yes" with a probability of $P(N(t) > 0)$ or "no" with a probability of $P(N(t) = 0)$.

The proposed Poisson model is depicted in Figure 3. The upper branch of the probability tree represents the already established carriers, so people who in principle engage in the behavior in question — for instance, shoplifters. Their proportion of the underlying population is described by the parameter π.[6] On the other hand, the lower branch of the probability tree represents *non-carriers*, i.e., people who do not engage in the behavior in question. Or, in the example at hand, participants of the survey who do not shoplift. The probability of a random participant being a non-carrier is the counter probability to π, so, $1 - \pi$. It is assumed that non-carriers will answer "no" regardless of the time frame t_i referred to in the question.

As the answer of the carriers will depend on the time frame t_i, the upper branch is divided into two further branches: There are carriers who have exhibited the behavior at least once in the time frame included in the question ($N(t) > 0$) and therefore would answer "yes" . Carriers who have not shown the behavior in the time frame in question ($N(t) = 0$) would answer "no" , respectively. With the assumptions presented above, $N(t)$ is a Poisson-distributed random variable with the intensity parameter λ. In the

[6] The parameter π is called p in the original research paper. However, since p denotes the probability of drawing the sensitive question in the UQM, p will be called π in the following, since the parameter π in the UQM also describes the proportion of trait carriers.

Poisson model, λ corresponds to the rate of the behavior in question. The probability $P(N(t) = k)$ that a carrier showed the behavior in question, e.g., shoplifting, k times during the time interval t can thus be described as

$$P(N(t) = k) = \frac{(\lambda \cdot t)^k \cdot e^{-\lambda \cdot t}}{k!}. \tag{3.1}$$

For $k = 0$, this probability is

$$P(N(t) = 0) = e^{-\lambda \cdot t}. \tag{3.2}$$

The probability of a random participant responding with "yes" to the time-constrained yes/no-question can be denoted as

$$P(\text{“yes”}\,|t) = \pi \cdot P(N(t) > 0), \tag{3.3}$$

or

$$P(\text{“yes”}\,|t) = \pi \cdot [1 - P(N(t) = 0)]. \tag{3.4}$$

Inserting Equation 3.2 yields

$$P(\text{“yes”}\,|t) = \pi \cdot (1 - e^{-\lambda \cdot t}). \tag{3.5}$$

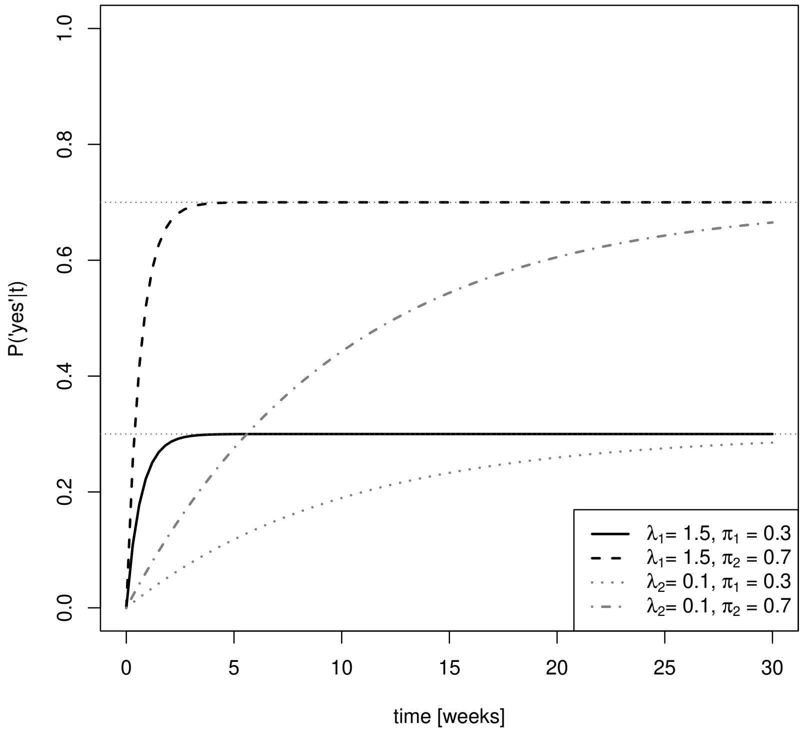
This *prevalence curve* describes the progression of the prevalence of the behavior in question relative to the time t.

As evident from the prevalence curves shown in Figure 4, the curves' asymptote is determined by the parameter π. Meanwhile, the slope of the curves is defined by the intensity parameter λ. The parameter π is to be interpreted directly as the proportion of trait carriers. The parameter λ is more difficult to interpret. It has the dimension $[\text{time unit}]^{-1}$ and refers to the time unit used as the baseline for the design parameter t_i, e.g., weeks. For example, if $\lambda = 2$, the behavior in question occurs twice per week on average. Thus, the inverse of λ yields the time intervals in which the behavior of interest occurs. In the given example, the behavior in question occurs every $\frac{1}{\lambda=2} = 0.5$ weeks.

When at least two different groups of participants are surveyed on their behavior during distinct time frames t_i, it becomes possible to estimate both parameters π and λ, e.g., through the maximum likelihood procedure. Adding a third group enables the possibility of testing model fit.

To summarize, via estimation of π and λ, the Poisson model allows one to derive information on both the proportion of trait carriers and the rate of the behavior in question, while still utilizing standard, easy-to-answer time-constrained yes/no-questions. Additionally, it becomes possible to predict the prevalence of a behavior within an ar-

Figure 4: Example prevalence curves with $\lambda_1 = 1.5$, $\lambda_2 = 0.1$, $\pi_1 = 0.3$, and $\pi_2 = 0.7$.



bitrary time frame, even if no data was collected for this particular time frame.

The Poisson model was first applied in a survey study presented in Paper 2. A sample of 872 students of Tübingen University — 839 after exclusions — were divided into three groups and queried about six types of everyday behavior. They were asked whether...

1. ...they watched the weekly sports program *Sportschau*,
2. ...they watched the weekly crime thriller *Tatort*,
3. ...they ate pizza,
4. ...they drank coffee,
5. ...they congratulated a relative on their birthday, and
6. ...they participated in another survey

...within a certain time frame. The respective time frame per question varied between groups: The questions for group 1 referred to the past week, those for group 2 to the past month and those for group 3 to the past three months.

Table 1: Sample data collected for the six questions regarding everyday behavior.

		time frame		
Question		last week	last month	last three months
sports program	"yes"	27	48	65
	"no"	256	236	207
crime thriller	"yes"	22	59	69
	"no"	261	225	203
pizza	"yes"	131	250	257
	"no"	152	34	15
coffee	"yes"	171	190	188
	"no"	112	94	84
birthday wishes	"yes"	77	191	239
	"no"	206	93	33
other surveys	"yes"	53	140	194
	"no"	230	144	78

Table 1 shows the observed sample data, i.e., the amount of "yes" and "no" answers per group and question. The predicted prevalence curves in Figure 5, together with the estimated parameter values and G-tests depicted in Table 2 as well as comparisons with data from third parties, support the claim that the proposed procedure is well suited to model the prevalence of everyday behavior.

Some additional Monte Carlo simulations (presented in Appendix D of the original paper) revealed that the accuracy of the method is not only dependent on sample size and the true values of π and λ but also on the time frames t_i chosen for the application. Generally, the Poisson model performed well in simulations on parameter retrieval. However, estimation of π is better when the points of measurement lie on the asymptote of the assumed true prevalence curve. On the other hand, estimation of λ improves when shorter time frames are used, so that the slope of the true prevalence curve is captured by the points of measurement. Before applying the Poisson model, it might therefore be useful to conduct pilot studies or to rely on third party data on the researched behavior. When the behavior is expected to have a high rate λ, at least one group should be queried about the behavior within a rather short time interval. In turn, when it is to be expected that a high proportion of participants engage in the target behavior, at least one of the employed time frames should refer to a rather long time interval.

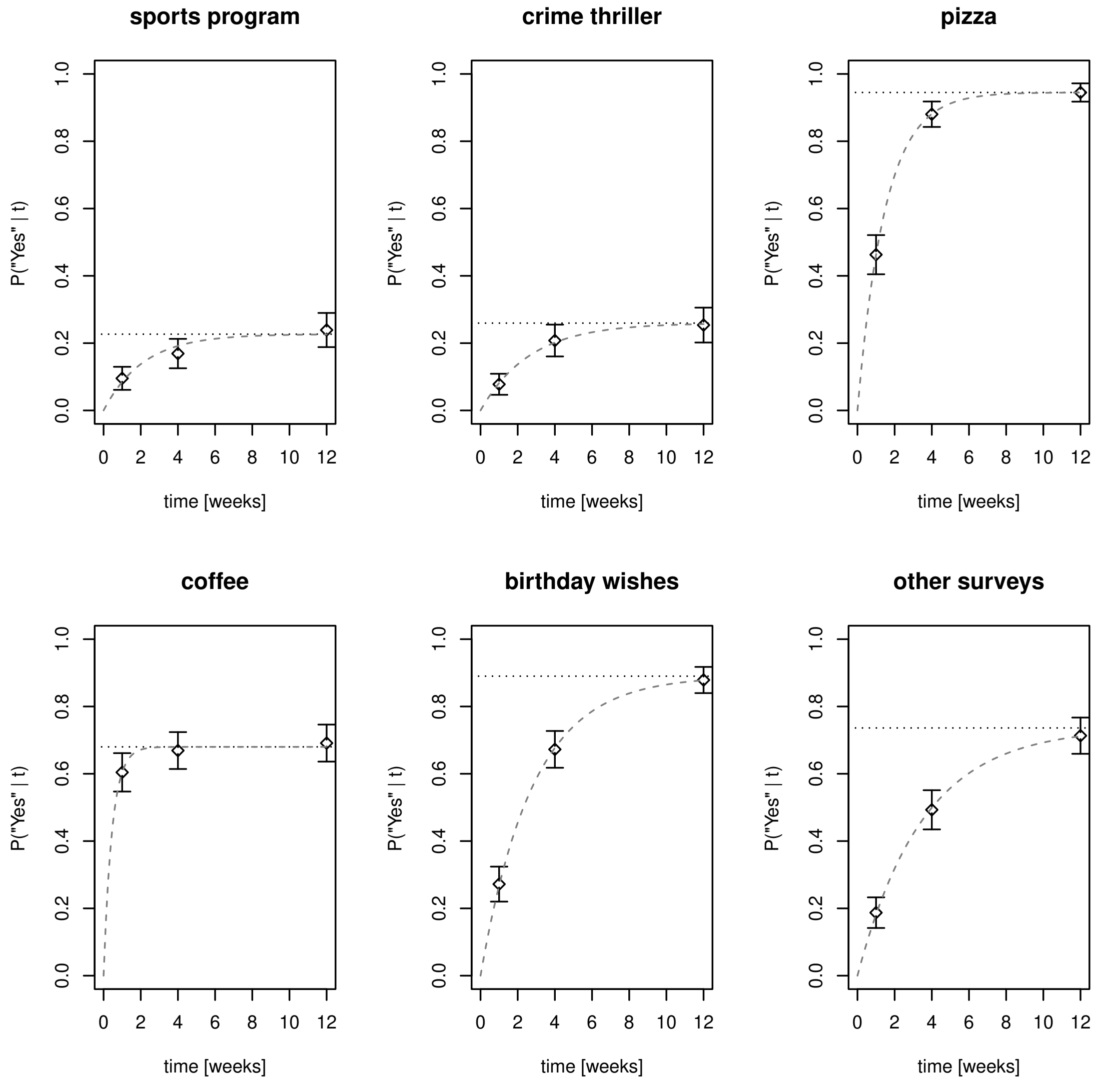
In conclusion, the Poisson model seems to be a promising alternative to traditional methods for researchers interested in measuring the prevalence and the rate of a behavior. As Paper 2 showed, the model seems suitable for surveys on everyday — so,

Table 2: Point estimates, standard errors, and 95% confidence intervals for the parameters π and λ, and results of G-tests (G statistics and p values) to evaluate goodness of fit.

	π			λ				
Question	Estimate	SE	95%-CI	Estimate	SE	95%-CI	G	p value
sports program	.230	.028	[.181; .291]	0.502	0.193	[0.241; 0.979]	1.649	.199
crime thriller	.263	.028	[.211; .323]	0.382	0.104	[0.220; 0.625]	0.123	.726
pizza	.945	.013	[.919; .968]	0.675	0.055	[0.574; 0.789]	0.001	.976
coffee	.679	.019	[.642; .718]	2.394	0.862	[1.567; 4.956]	0.311	.577
birthday congrat.	.891	.024	[.847; .937]	0.360	0.033	[0.300; 0.425]	0.058	.810
other surveys	.739	.037	[.673; .817]	0.285	0.039	[0.211; 0.369]	0.107	.743

Note. λ has the dimension $[\text{week}]^{-1}$. The point estimates, standard errors, and confidence intervals were calculated using a parametric bootstrap algorithm with 1,000 bootstrap samples, and the maximum likelihood procedure. G-tests were calculated with one degree of freedom ($df = 1$).

Figure 5: Observed proportion of "yes" answers and prevalence curves for the sample data in Table 1.



Note. The error bars represent 95% confidence intervals.

non-sensitive — behavior. A particular strength of the model rests in the simplicity of the employed time-constrained yes/no-questions. Thus, it is possible to present participants with a multitude of such simple questions, while simultaneously extracting information on a higher level via the parameters π and λ.

Theoretically, the approach of the Poisson model could benefit any field of prevalence research, including quantitative criminology. Obviously, there is still a need for more research to examine how valid and valuable applications of the Poisson model truly are. Its simplicity which could be seen as a weakness due to the strict core assumptions might actually be a strength of the model, as it allows for flexible adaptations or extensions. For instance, one could use alternatives to the Poisson distribution in cases where different probability distributions are better suited (e.g., to model relapse behavior, the exponential distribution might be promising). Additionally, in order to improve the Poisson model for measurement of sensitive behavior, it seems promising to pair it with RRMs. In Paper 3, a "Poisson extension" to the UQM was proposed and tested.

3.2 The Poisson extension to the UQM

[7]In Paper 3, we proposed a combination of the UQM with the Poisson model described previously. We called this model combination the *Poisson extension to the UQM*, or *UQMP*. The goal of the UQMP is to disambiguate the UQM's estimations of a behavior's prevalence when using time-constrained yes/no-questions. When applying the original UQM to measure a behavior's prevalence, the parameter π simply represents a time-constrained prevalence of the behavior in question. However, in the UQMP, π is equivalent to the parameter π in the Poisson model, representing the proportion of trait carriers independent of time. Additionally, it enables the measurement of the behavior's rate λ. With this extension, the UQM gains the benefits of the Poisson model without losing its essential strength of guaranteeing complete anonymity for participants.

The combination of both models is shown in the probability tree in Figure 7. As in the original UQM, the tree first splits into two branches: One branch represents the participants drawing the sensitive question (S) with the probability p, while the other branch represents the participants drawing the neutral question (N) with the counter probability $1 - p$. Like in the UQM, the bottom branch then splits into two further branches which represent the outcomes of a yes-response (with the probability q) and a no-response (with the counter probability $1 - q$). After the first node, the upper branch is basically equivalent to the original Poisson model.

As in the Poisson model, the UQMP requires the sample to be split into at least two groups, and more than two to test model fit. In each group, the sensitive question of the UQM refers to a different time frame t_i. The parameters λ and π can be estimated via the maximum likelihood procedure. Like in the original UQM, p and q are constants set by design.

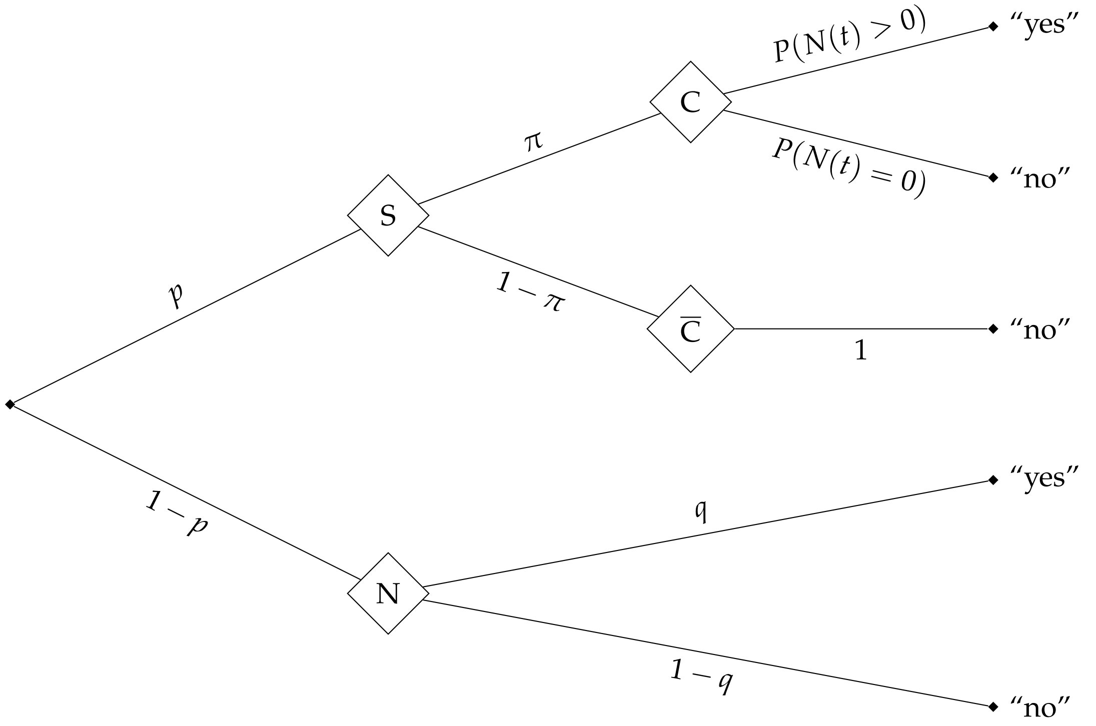
In the UQMP, the probability of a random participant responding with "yes" is given by

$$P(\text{“yes”}\,|t) = p \cdot \pi \cdot (1 - e^{-\lambda \cdot t}) + (1 - p) \cdot q. \tag{3.6}$$

Note that this probability does not describe a prevalence curve because the yes-

[7] This section is based on Paper 3 of the four underlying publications (Iberl, Aljovic, Ulrich, & Reiber, 2024), see Appendix C.

Figure 6: Probability tree of the UQMP



Note. The sample is divided into two groups. First, respondents drawing the sensitive question (S), and second, respondents drawing the neutral question (N), with the probabilities of p and $1 - p$, respectively. A respondent who draws the sensitive question is a trait carrier (C) with a probability of π, or a non-carrier ($\overline{\text{C}}$) with the counter probability $1 - \pi$. Non-carriers will respond with "no" with a probability of 1. With a probability of $P(N(t) > 0)$, a carrier will give a yes-answer to the sensitive question. $P(N(t) = 0)$ describes the counter probability of a carrier giving a no-answer to the sensitive question. As in the standard UQM, the probabilities for respondents who drew the neutral question to answer with "yes" or "no" are q and $1 - q$, respectively.

responses of participants drawing the neutral question are also represented in this equation. Instead, the conditional probability of a yes-response, given one draws the sensitive question, is the equivalent of the prevalence curve in the UQMP, that is

$$P(\text{"yes"}\,|t, \text{sensitive question}) = \frac{P(\text{"yes"}\,|t) - (1 - p) \cdot q}{p}, \tag{3.7}$$

which is equal to the prevalence curve of the standard Poisson model (see Equation 3.5) after inserting Equation 3.6.

In the presented Paper 3, the UQMP was applied in a survey on the prevalence of drinking and driving in Germany, and compared to its DQ equivalent, the original Poisson model.

Generally, and maybe surprisingly so, the prevalence of drinking and driving in Germany is fairly under-researched. The most elaborate research on this behavior was conducted by Krüger and Vollrath (1998), who, with help of the police, randomly pulled over drivers from traffic and measured their blood alcohol level (BAC). In this *roadside survey*, they found that 1.2% of the tested drivers had a BAC higher than 0.05% and therefore violated German traffic law. It is not quite clear how indicative these study results are for the prevalence of drinking and driving in the modern day. For one, the study was conducted more than 25 years ago, so the findings might not be transferable to the current situation. Furthermore, as the roadside survey was conducted only on specific days, times, and places, it is possible that the researchers "missed" some drunk drivers who deliberately avoid driving in certain areas or at certain times in order to minimize the risk of getting caught. More recently, Goldenbeld, Torfs, Vlakveld, and Houwing (2020) researched the prevalence of drunk drivers in multiple countries, including Germany. They conducted an online survey, employing direct questions to ask participants whether they might have violated the traffic law in regards to drinking and driving. They found a past month prevalence of 9%. In another study, presented later in Paper 4, the UQM was used to measure the lifetime prevalence of "driving under the influence while accepting the possibility of a rule violation" in students, resulting in an estimate of 44% (Iberl, 2021).

So, while some studies measured the prevalence of drinking and driving in Germany using different methods, the respective results vary greatly. Additionally, none of the studies measured the proportion of trait carriers independently of time or the rate of the behavior. So, applying the approach of the Poisson model to this field of research and comparing DQ to the UQM(P) seemed very promising. We expected that the UQMP would be suitable for measuring these variables. Also, assuming that drinking and driving is a sensitive topic for participants, we hypothesized that the UQMP would yield a higher estimate for the proportion of carriers than the standard Poisson model (using DQ).

Relying on the service of the market research company *bilendi*, we surveyed a sample of $N = 3,529$ (after exclusions) that was drawn to be representative for German motorized traffic participants in regards to age and gender. The sample was split into two groups, one smaller group ($n = 878$) for application of the standard Poisson model, and one larger group ($n = 2,651$) for the UQMP application. Obviously, since the UQM is less efficient than DQ, the UQMP is less efficient than the standard Poisson model as well, thus a bigger sample is needed. A priori simulations led to the conclusion that four time frames would most probably yield the most valid UQMP estimations. Hence, each of the two groups was again divided into four subgroups in which the time frame t_i the sensitive "drinking and driving" question referred to was varied. Overall, the study

Table 3: Sample data for the question on drinking and driving for each group.

group	time frame	n	"yes"	"no"
UQM	one week	672	283	389
	one month	670	272	398
	six months	654	276	378
	one year	655	277	378
DQ	one week	210	20	190
	one month	221	23	198
	six months	215	17	198
	one year	232	22	210

employed a 2 x 4 between-subjects design (DQ or UQM x drinking and driving in the past week, past month, past six months or past year).

Participants in the four UQM subgroups were asked to think about a person whose date of birth they knew. Afterwards, the following UQM design was presented to them:

Is the birthday of the person you thought about between the 1st and 10th day of the respective month? Then please answer question A honestly.
Is the birthday of the person you thought about between the 11th and 31st day of the respective month? Then please answer question B honestly.

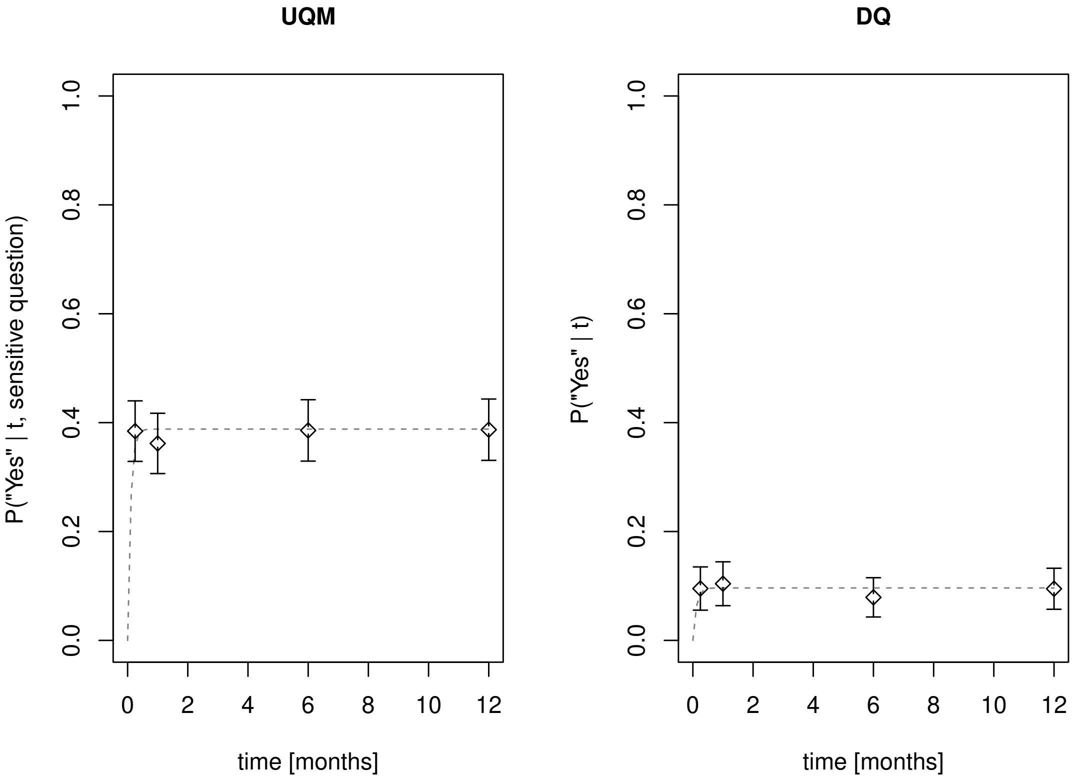
Question A: Is the birthday of the person you thought about in the first half of the year, so before the 1st July of a year?
Question B: Did you drive a motorized vehicle (a car, motorcycle, scooter, etc.) in the last week/month/six months/year while being drunk or knowing that you had too much to drink?

In the DQ group, participants were directly presented with the sensitive question.

To ensure that the UQM functioned as intended, a control question was presented following the drinking and driving question. Depending on the group, participants were asked whether their eye color was blue either via DQ or the UQM. In theory, as this question was assumed to be neutral in nature, the resulting prevalence estimates should not differ between groups.

Table 3 and Figure 7 present the main results of the study. Clearly, the proportion of yes-answers to the drinking and driving question is much higher in the UQM-group, with a π estimate for trait carriers of .388 versus .096 in the DQ-group. At first glance, this seems to reflect the superiority of RRMs when asking sensitive questions, leading to more valid results in the UQM-group compared to those in the DQ-group which could

Figure 7: Prevalence curves for the parameter estimates in the UQMP and the standard Poisson model.



Note. The points represent the proportion of yes-responses (to the sensitive question) per time frame. Error bars represent 95% confidence intervals.

be biased due to socially desirably responding. However, this interpretation could be misleading in this case, as two results point towards methodological problems within the study.

First, the responses shown in Table 3 do not vary based on the time frame used in the questions — neither for the UQM-group, nor for the DQ-group. This causes the prevalence curves in Figure 7 to rise very steeply and to already reach the asymptote π on the first point of measurement ($t_1 = \frac{1}{4}$ months). In turn, this leads to the upper boundary of the λ estimates for both groups to reach the limit of $\lambda = 10$ set in the parameter estimation algorithm. Considering this, a line would be more fitting than a curve to describe the data. Theoretically, when the true data is indeed a horizontal line, λ could be infinite. Hence, while such high λ values suggest a very high rate of this behavior on paper, the estimates should be interpreted carefully. It is unclear why this result occurred. Maybe the data is indeed valid, indicating that almost all drunk drivers in Germany show the behavior in question very frequently. Another explanation is that many participants did not read the questions carefully enough, causing them to overlook the time frame referred to in the question on drinking and driving. Random or careless responding might explain this result as well; however, this explanation could not be supported by post-hoc analyses. For one, we found no effects of response time on the responses, testing the assumption that careless or random responding lead to faster responses on average. Plus, the amount of careless or random responses that would be necessary to lead to the presented results seems unrealistically high.

Second, the question inquiring about the respondents' eye color revealed surprising results. Similar to the difference between the π estimates, the eye color prevalence varied considerably between groups: The UQM estimate (.517) was significantly higher than the DQ (.355) estimate, even though the eye color question should be neutral and thus yield similar results regardless of questioning technique. The reasons for this, however, remain unclear as well. Neither response time nor order effects were found in post-hoc analyses and an additional survey study. The most plausible explanation for the big difference between UQM and DQ estimates in the eye color question is that the UQM did not work as intended or that the question was not as neutral as assumed. Besides the possibility of careless or random responding, it might be that many participants had trouble understanding the instructions or chose to not comply with the UQM procedure for unknown reasons. Alternatively, the very "survey experienced" participants might not care too much about their level of anonymity in surveys that they participate in for monetary compensation. Participants who actually do not perceive the sensitive question as sensitive could contribute to the models not working as intended. In such cases, the additional anonymity enabled by RRMs might be useless. Regardless of the reasons for these surprising results, this clearly indicates that the UQM estimate of almost 40% for the proportion of drunk drivers in Germany is too high.

Evidently, more research is needed to test under which circumstances the UQMP functions best, or, at least, as intended. Regardless of the unexpected results, the study showed that, in principle, a combination of the UQM and the Poisson model is feasible. Interestingly, the Poisson model itself does not seem to be responsible for the unexpected results — rather, the UQM seems to have been the problem. It might be that the online sample had something to do with the surprising results as well. Perhaps samples consisting of recruited, extrinsically motivated respondents are less reliable than samples with intrinsic motivation to participate. Generally, surveying samples not recruited in online panels might lead to more valid results. Overall, it still does not seem clear how the data of survey research utilizing (commercial) online panels compares to data generated by other sampling methods (for an overview on relevant research, see Callegaro, Villar, Yeager, & Krosnick, 2014). One study by Litman, Robinson, and Rosenzweig

(2015) suggests that the level of compensation for panel participants can be influential, with higher payment increasing data quality. Another study found that commercial panels performed worse than *Mechanical Turk*, a crowd-sourcing platform by *Amazon*, despite the latter being cheaper (B. Zhang & Gearhart, 2020). In future applications of the UQMP, it might be advisable to refrain from using commercially recruited samples.

3.3 Discussion

In conclusion, the newly proposed Poisson model seems to be a promising and efficient alternative to measure the prevalence of trait carriers as well as the rate of the underlying behavior of interest. While the Poisson model seems to work well in itself and for non-sensitive questions, it showed some problems in combination with the UQM. However, the unexpected — and, most likely, in part invalid — results of the UQMP study seem to originate from the UQM and/or the sample used in the study rather than from the Poisson model.

One possible explanation for the UQM not working as intended is random or careless responding. It is disputed how much of a problem random responding is in RRM applications. While Walzenbach and Hinz (2019) found evidence for random responding having a major impact in an application of the CM, a validation study by Meisters, Hoffmann, and Musch (2022a) showed that only a small proportion of their respondents answered randomly. Wolter and Diekmann (2021) found that both random responding and unexpected results regarding the neutral question in the CM might be responsible for high rates of false positive answers. However, our results indicate that random or careless responding could have been one reason for the UQM not working as intended, but that it can not fully explain the high difference between the results of the UQM and DQ groups.

Furthermore, the fact that the sample in this study was recruited via an online panel might have contributed to the unexpected results. Paradoxically, settings where participants feel that their anonymity is not well protected might be especially well-suited for RRM applications, since RRMs are assumed to work best when participants perceive the questions at hand as sensitive (Lensvelt-Mulders et al., 2005). Participants might value the additional anonymity more in such settings, which could increase motivation and instruction adherence. In settings where participants already feel that their anonymity is well protected, RRMs might lose their advantage compared to DQ. This could especially be the case when querying participants with high experience in online surveys. Some research suggests that participants generally feel most anonymous in online surveys, and/or that the social desirability bias is weakest in online surveys (e.g., Chang & Krosnick, 2009; Holbrook, Green, & Krosnick, 2003; Kreuter, Presser, &

Tourangeau, 2008; Robertson, Tran, Lewark, & Epstein, 2018). Even though other studies have found no differences between survey modes in terms of social desirability bias (e.g., Dodou & de Winter, 2014; X. Zhang, Kuchinke, Woud, Velten, & Margraf, 2017), both participants' intrinsic motivation and instruction adherence might be higher in personal settings compared to online surveys (e.g., Alfonsson, Johansson, Uddling, & Hursti, 2017; Clancy & Taylor, 2016; Heerwegh, 2009). Therefore, RRMs might work better in face-to-face settings and with samples that perceive the questions as particularly sensitive.

Additionally, it has to be considered that the instructions of an RRM application are undoubtedly more difficult to understand than traditional direct questions. Multiple studies revealed that comprehension is an issue in RRM applications (e.g., De Schrijver, 2012; Hoffmann & Musch, 2016; Landsheer, Van Der Heijden, & Van Gils, 1999; Lensvelt-Mulders & Boeije, 2007; Meisters et al., 2020; Wolter, 2012). Hence, it cannot be ruled out that comprehension problems contributed to the unexpected results as well.

To summarize, multiple possible explanations for the UQM's lack of validity in the presented study, but also for RRMs' lack of validity in general come to mind. Most probably, a mixture of these explanations, and possibly other factors as well, lead to the validity problems in RRM applications.

Still, since RRMs have been used successfully many times before (e.g., Dietz et al., 2018; Lensvelt-Mulders et al., 2005; Ulrich et al., 2018), and due to their promising core principle (see the example of its usage to survey Russians on the Russian-Ukraine war, Chapkovski & Schaub, 2022), the UQMP should not be given up on just yet. Rather, future research should identify the circumstances under which the UQMP leads to valid results. If these circumstances can be found, the combination of truly anonymous responding and the advantages of the Poisson model might in many instances be worth the large sample sizes needed for parameter estimation.

The problems with the UQM in the second "Poisson study" indicate that, after more than 50 years since its invention (Greenberg et al., 1969), it still is not completely clear when or why the UQM does (not) work. A review of the available literature indicates that these problems are not exclusive to the UQM, but appear in other RRMs as well. The following chapter focuses on comprehension as a possible key factor in the functionality of RRMs and the UQM.

4 Instruction (non-)adherence and comprehension in RRM applications

[8]As previously established, the goal of RRMs is to offer a more valid alternative to traditional methods when researching sensitive topics. However, in cases where their validity is not superior, the additional statistical noise and required samples sizes are simply not worthwhile. Historically, when comparing RRMs to DQ, a ″more-is-better″-logic was often applied (see e.g., Buchman & Tracy, 1982; Lensvelt-Mulders et al., 2005; Umesh & Peterson, 1991): When researching a sensitive topic, one assumes the existence of social desirability bias and thus expects that the prevalence of interest is underestimated in a DQ design. Hence, when judging the performance of RRMs on the same sensitive question, it seems logical to expect a higher prevalence estimate to be more valid than one close to the DQ estimate.

In more recent publications, this ″more-is-better″-assumption has been put into question due to the possibility of both false positive and false negative responses occurring in DQ as well as in RRM applications (Höglinger & Jann, 2018). In some cases, higher prevalence estimates in RRM studies could simply occur due to false positive responding. To evaluate the validity of RRMs, *strong validation studies* have been carried out as an alternative to the ″more-is-better″- paradigm (Hoffmann & Musch, 2016; Höglinger & Jann, 2018). These studies enable the estimation of false positive and false negative responses by design. Generally, this is accomplished by choosing the sensitive question in a way that the prevalence of yes and no answers is known or can be estimated a priori. Multiple strong validation studies have shown that the higher prevalence estimates in RRM studies can be caused by a higher proportion of false positive responses (Höglinger et al., 2016; Höglinger & Jann, 2018; Meisters et al., 2020). This is precarious for RRMs, since it questions their raison d′être of superior validity in research on sensitive topics. Both this chapter and Paper 4 focus on the role that participants′ comprehension of instructions play for the validity of RRM applications.

[8] This chapter is based on Paper 4 of the four underlying publications (Iberl, 2021), see Appendix D.

4.1 The theoretical importance of comprehension

While the circumstances under which false positive answers can occur in RRMs are not yet fully understood, the most natural explanation is that some participants do not adhere to the provided instructions. Such instruction non-adherence could either be intentional or unintentional.

When it comes to intentional non-adherence, some participants might mistrust RRMs for various reasons, leading to intentional self-protective responses. According to a study by Bullek, Garboski, Mir, and Peck (2017), some participants prefer to answer DQs over participating in an RRM design even though their anonymity is better protected in the latter. This could also lead to intentional instruction non-adherence: In the UQM, some participants might choose to ignore the instructions and respond to the sensitive question even when the experiment's outcome points them to the neutral question, basically pretending to be presented with a DQ instead of a UQM. Furthermore, some participants might even choose to respond randomly, be it due to feeling patronized by the RRM design or due to losing interest in the comparatively complex instructions.

In regards to unintentional non-adherence, some participants might have problems comprehending the instructions. This could lead to unintentional false-positive or false-negative responses, for instance, when participants do not comprehend which of two questions they are supposed answer.

The importance of comprehension within RRMs transcends the issue of instruction adherence, as it is key for minimizing social desirability bias. Certainly, it is not necessary for a participant to understand how a researcher calculates a prevalence despite not knowing the outcome of their random experiment. It might not even be necessary for participants to comprehend *why* RRMs offer a higher level of anonymity. However, comprehending *that* RRMs offer a higher level of anonymity should most definitely be important: If a participant does not feel more anonymous when presented with an RRM application to a sensitive topic, there is no obvious reason for them to answer more honestly than in a DQ design. Hence, the participants' comprehension of the increased anonymity offered by RRMs should in theory be vital for their functionality.

While it is to be expected that many respondents are more confused by RRM instructions compared to traditional methods, some studies have shown that respondents' lack of comprehension might be problematic for the validity of RRM applications (e.g., De Schrijver, 2012; Hoffmann & Musch, 2016; Landsheer et al., 1999; Lensvelt-Mulders & Boeije, 2007; Meisters et al., 2020; Wolter, 2012). Apparently, the educational level of the participants can constitute an important factor for instruction comprehension and thus, compliance: Higher educated participants comprehend the instructions better on average, leading to more valid results (e.g., Böckenholt & Van der Heijden, 2007; Hoffmann & Musch, 2016; Landsheer et al., 1999; Meisters et al., 2020; Wolter, 2012).

4.2 Comprehension checks as a guarantee of validity?

One possibility to ensure that participants understand the RRM instructions is adding comprehension aids to the survey design. Such a comprehension aid could, for instance, include "test runs" of the employed RRM design in which the participants are asked how a fictional participant — who is a carrier or non-carrier of the sensitive attribute — would have to respond given a certain outcome of the randomization device.

Such a "comprehension check" could look as followed: Assume that a participant is presented with the UQM design from Section 3.2, where birthdays are employed as both the randomization device and the neutral question, with the sensitive question referring to drinking and driving — except, in this case, they are supposed to think about their own birthday. Now, one could tell the participant that a fictional person with a given date of birth drove under the influence during the past year. Then, the participant would be asked how the fictional person would have to respond, given they would answer honestly. For example, one could ask them: "Last month, Tom drove under the influence. His birth date is on September 6th. How would Tom have to answer?". In this case, because Tom's date of birth is between the 1st and 10th day of the respective month, he would be supposed to respond to the neutral question ("Is the birthday of the person you thought about in the first half of the year, so before the 1st July of a year?"), to which he would be supposed to answer "no" . If the participant follows the instructions correctly, they should be able to respond correctly ("Tom would be supposed to answer 'no'").

Presenting participants with such comprehension checks should familiarize them with the design. Additionally, they should have a higher chance of understanding that their answers are completely anonymous in RRM designs. That said, *incongruent* examples in comprehension checks, like the given example above, where a trait carrier has to answer "no"— or where a non-carrier has to answer "yes"— could be especially effective.

Theoretically speaking, such comprehension checks could have two major benefits. First, they enable researchers to filter out participants who gave too many wrong answers in the comprehension checks, which could point towards a lack of instruction comprehension. Second, comprehension checks should enhance the instruction comprehension across the sample and thus increase the validity of the study.

Meisters et al. (2020) tested the effect of such comprehension checks in a strong validation study using a CM design. Additionally, they examined possible effects of educational level. They found that comprehension checks increased validity by decreasing the amount of false positive answers. However, this effect only occurred within the subgroup of participants with a higher educational level. Another study by the same research team confirmed positive effects of comprehension checks (Meisters, Hoffmann, & Musch, 2022b).

Inspired by Meisters et al. (2020), the study in Paper 4 employed comprehension checks in a UQM application, however without using a strong validation study design. Apparently, at the time of the study in Paper 4, the effect of comprehension checks in UQM applications had not yet been researched. Instead of employing a strong validation study design, the sensitive question inquired about a behavior of which the prevalence can not be objectively verified. While this is a possible limitation of the study, it seemed reasonable to use a very simple and inexpensive study design to examine whether comprehension checks would have an obvious influence on the prevalence estimates.

For the study, a student sample was invited to participate in a survey on drinking and driving. Considering that a student sample is inherently comprised of participants with a high educational level, this can be counted as another limitation of the study. Thus, it was not possible to control for a possible effect of educational level. However, as Meisters et al. (2020) found an effect of comprehension checks only in highly educated participants, the sample still seemed suitable to examine a possible effect of comprehension checks.

The sample was divided into three groups: One group was presented with a DQ design in which they were asked directly about their behavior (DQ group). The second group was presented with a UQM design which employed dates of birth as both the randomization device and the basis for the neutral question (UQM group). The instructions for this group were kept simple and short. The third group was presented with a UQM design as well — however, the instructions were more elaborate for the third group and explained in detail how birth dates can be used as a randomization device (UQMC group). Furthermore, comprehension checks were used in the UQMC group. The comprehension checks consisted of three test runs of the UQM design which were based on fictional participants with various combinations of birth dates and drinking and driving behavior. Every participant in the UQMC group was presented with at least one incongruent comprehension check.

Surprisingly, the prevalence estimates for drinking and driving did not differ between the three groups. Apparently, comprehension checks alone do not have a strong effect on participants responses — at least with respect to the sensitive question used in the survey. While in Paper 3 (see Section 3.2), drinking and driving was defined as "driving while drunk", a broader definition was used in Paper 4, where the participants were asked whether they ever "accepted the possibility that they were driving under the influence". The reasoning behind this was to not only include people who drive drunk but also people who drive even though they are unsure whether they might have had too much to drink. It is possible that the sensitive question was not perceived as very delicate due to this wording. This could explain that no differences were found between the prevalence estimates in the three groups. With respect to comprehension of instructions, it is possible that the comprehension checks might have had no effect on the prevalence estimates because most participants complied with the UQM design regardless of how elaborate the instructions were. In other words, comprehension checks did not seem to be necessary in order for the student sample to follow the instructions. This hypothesis is supported by the finding that most participants in the UQMC group had no problems with answering the test questions correctly.

In conclusion, the participants might have been both motivated and smart enough to follow the instructions even with minimal instructions. Assuming that this is the case, this would render comprehension checks useless for highly motivated and educated samples. However, this somewhat contradicts the findings of Meisters et al. (2020), who — employing a CM design — found an effect of comprehension checks in highly educated participants. Multiple explanations come to mind. It may be possible that the effect of comprehension checks functions differently for CM and UQM designs. Another explanation could be that the design of the study at hand was simply not suited to answer the research questions; perhaps the comprehension checks did actually have an effect that was cancelled out by a higher amount of both false-positive and false-negative responses in the other UQM group. Furthermore, it also seems possible that the surveyed students had high trust in researchers from their university to comply with the usual ethical guidelines of survey research and therefore felt highly anonymous regardless of the question design. Lastly, comprehension might not be as important a factor for RRM validity as suspected — even though, in theory, realizing that one's anonymity is granted completely and objectively in RRM designs should be essential for RRMs working as intended. Perhaps *intrinsic motivation* of participants is actually a more important factor in this regard than comprehension.

4.3 Discussion

Theoretically, basic comprehension of instructions should be a vital factor for RRMs validity. If a participant does not comprehend that they are more anonymous in an RRM design compared to a DQ design, they should have no logical reason to answer more honestly. Thus, the RRMs goal of minimizing social desirability bias should at least in some capacity depend on participants' level of comprehension.

However, both a recent study by Meisters et al. (2020) and the study in Paper 4, where test questions were employed to ensure better comprehension in certain subgroups, showed that this suspected comprehension effect is not as simple as assumed. While Meisters et al. (2020) found a comprehension effect in the CM only for participants with higher education, no such comprehension effect was found in the UQM study.

Even though multiple explanations come to mind for this result, it is also possible that comprehension of the higher level of anonymity offered by RRMs might not be the most important reason for participants to answer more honestly. Rather, motivation to participate could be the more important factor; in turn, participants' motivation could also be connected with their level of comprehension, since intrinsically motivated participants can be assumed to read more carefully and try harder to comprehend the instructions than participants with lower intrinsic motivation. Additionally, highly motivated participants should be more reluctant to resort to careless or random responding. There could also exist an important interaction between highly motivated participants and participants who appreciate the additional anonymity provided by RRMs: Inquiring about a topic that one finds sensitive might automatically lead to a higher intrinsic motivation and more careful responding. Assuming that participants recruited via online panels are less motivated than student samples, this could also explain the surprising results in Paper 3 (see Section 3.2), where an online panel was used to draw the sample. Unfortunately, the role of motivation in RRM applications has apparently not been researched as of yet.

Without a doubt, the study at hand came with a number of limitations in terms of examining the effects of comprehension checks in UQM designs. Still, the findings suggest that a simple comprehension effect — in the sense of: "More comprehension equals lower prevalence estimates" — is unlikely and that simply including comprehension checks and test questions in RRM designs is most likely not the answer to the validity problems of RRMs. Additionally, it raised further questions regarding the role of motivation. Future studies, at best strong validation studies, should examine how motivational factors, comprehension, and instruction adherence interact with each other, and how they influence the validity of RRMs.

5 Summary and Conclusion

This chapter includes a short summary of the four research articles presented in this dissertation. Following that, some general conclusions will be drawn based on the main findings from Papers 1-4 and regarding the central question of whether RRMs have a purpose in criminological research.

5.1 Summary

Paper 1 revolved around the question of whether RRMs, specifically the UQM, are in principle suited for use in criminological survey research. The article focused on an already concluded empirical research project on plea bargaining in Germany. The results illustrated that social desirability bias can be a problem in criminological survey research. Additionally, the study suggested that the necessary methodological requirements for UQM applications were met both in the example project on plea bargaining as well as in many other cited criminological studies. Therefore, it was concluded that the requirements for the UQM, and most probably for other RRMs as well, can generally be met in such survey studies. However, the limited ways in which questions can be formulated in RRM applications and their dependence on questions referring to a single time interval may hinder their use in criminology.

Paper 2 proposed an alternative way to utilize such time-constrained questions, the Poisson model. This method offers the possibility to disambiguate prevalence estimates referring to certain time frames. The model proved to be promising for researching everyday activities.

Paper 3 introduced a combination of the Poisson model and the UQM, the UQMP. Said model was tested in an online survey on drinking and driving. However, the UQM did not work as intended in this study, which led to unexpected results. Presumably, these unexpected results had nothing to do with the Poisson model itself, but with instruction non-adherence and/or lacking intrinsic motivation of the participants that were recruited via a commercial online panel.

Paper 4 focused on the importance of comprehension for the validity of RRMs, specifically in the UQM. In theory, some basic level of instruction comprehension is the prerequisite for RRMs to work as intended and thus lead to more valid responses. In a student survey on drinking and driving, traditional DQ and UQM applications were compared to a UQM application with elaborate comprehension aids in the form of test questions. Surprisingly, the comprehension checks did not lead to differing results between groups. A possible explanation for this finding is that highly educated and motivated samples comprehend RRMs' instructions well, even without comprehension checks. In fact, motivation might be a more important factor for successful RRM applications than instruction comprehension.

5.2 Conclusion

This dissertation showed that the UQM — and probably any other RRM — is principally suited for researching criminological topics such as drunk driving or plea bargaining (and probably for criminological research in general). However, it was shown once more that the functionality and problems of the UQM are complex and not fully understood yet. According to the available literature, this finding can be transferred to RRMs in general to a certain degree. Most likely, this is the main reason for the rare use of RRMs in criminology and other sciences. Due to these problems, RRMs cannot be recommended as a standard method for quantitative criminology as of now. First, the circumstances under which RRMs work as intended and can develop their full potential as a more valid way to ask sensitive questions need to be identified. Surprisingly, comprehension of instructions does not seem to be as important to RRMs' — or at least the UQM's — functionality as previously assumed. A more likely reason for failing RRM applications might be lowly motivated participants in online samples.

Based on this, it can be concluded that inquiring about certain behaviors or events that are especially sensitive for specific populations while using a face-to-face design should be the ideal setup for RRM applications. To name an example of this hypothetical best-practice use of RRMs, Ulrich et al. (2018) used the UQM in a face-to-face setting to survey elite athletes on doping behavior. In online surveys, RRMs should only be used if the underlying question can be assumed to be highly sensitive. As an example for such an online application one could name the study of Chapkovski and Schaub (2022) who surveyed Russians on their opinion on the invasion of Ukraine. For further examples, one could think of investigating the prevalence of police brutality with a sample of police officers, querying a sample of judges about the prevalence of illegal plea bargains, or asking any sample whether they have committed severe crimes. As these examples show, there exist plenty of highly relevant research topics that rely on participants responding honestly to sensitive questions. Therefore, the basic idea of RRMs is still too good to be abandoned, despite their problematic validity.

Criminology is an excellent example for a discipline that could benefit from RRMs, as, by nature, it offers a multitude of sensitive topics that are highly relevant on a societal and political level. Future research should focus on establishing conditions under which RRMs can produce valid and reliable results. This includes research on the role of motivation for RRMs' validity and comparisons between online and face-to-face applications. Furthermore, the importance of instruction comprehension and (non-)adherence is still not fully understood. Research on possible interactions between motivation, instruction comprehension and (non-)adherence, as well as careless or random responding could be the key to successful future RRM applications. In spite of some remaining open questions in terms of their functionality, more criminological applications of RRMs are called for. Whenever traditional methods are used in control groups and presumably highly sensitive questions are posed to presumably highly motivated groups, criminological research might immediately benefit from RRMs.

Since it allows for efficient measurement of the proportion of trait carriers that exhibit a certain kind of behavior as well as the behavior's underlying rate, the Poisson model seems to be a promising alternative to traditional DQ prevalence surveys. Naturally, the model should be well suited for applications in quantitative criminology, as prevalence estimates are so widely used in this field of research. In fact, there is hardly any reason not to utilize the model immediately in prevalence research. The only apparent weakness of the Poisson model seems to be its simplicity, i.e., its rather strict assumptions. For instance, dividing a population into carriers (e.g., regular drug users) and non-carriers (e.g., people who never use drugs) is certainly an oversimplification. Additionally, assuming an invariant rate for all carriers might not be realistic in many cases — even though we showed that extending the Poisson model with a variance parameter for the rate does not have much influence on the other parameter estimations (Iberl & Ulrich, 2023). However, said simplicity is also a core strength of the Poisson model, since it allows for simple combinations with different models, such as RRMs. Naturally, more research is needed to flesh out the Poisson model's potential and ideal use cases. Such future research could focus on adapting and/or improving the model for specific purposes. While promising, Poisson extensions of RRMs such as the UQMP obviously need more testing as well — ideally in studies where the employed RRM is functioning as intended.

Bibliography

Alfonsson, S., Johansson, K., Uddling, J., & Hursti, T. (2017). Differences in motivation and adherence to a prescribed assignment after face-to-face and online psychoeducation: An experimental study. *BMC Psychology*, *5*, 1–13. doi: 10.1186/s40359-017-0172-5

Altenhain, K., Jahn, M., & Kinzig, J. (2020). *Die Praxis der Verständigung im Strafprozess [The practice of plea bargaining in criminal proceedings]*. Baden-Baden, Germany: Nomos. doi: 10.5771/9783748922094

Andrie, E. K., Tzavara, C. K., Tzavela, E., Richardson, C., Greydanus, D., Tsolia, M., & Tsitsika, A. K. (2019). Gambling involvement and problem gambling correlates among European adolescents: Results from the European Network for Addictive Behavior study. *Social Psychiatry and Psychiatric Epidemiology*, *54*(11), 1429–1441. doi: 10.1007/s00127-019-01706-w

Beck, F., Léger, D., Fressard, L., Peretti-Watel, P., Verger, P., & Group, C. (2021). Covid-19 health crisis and lockdown associated with high level of sleep complaints and hypnotic uptake at the population level. *Journal of Sleep Research*, *30*(1), Article e13119. doi: 10.1111/jsr.13119

Becker, P. (2002). *Verderbnis und Entartung — Eine Geschichte der Kriminologie des 19. Jahrhunderts als Diskurs und Praxis [Depravity and degeneration — A history of 19th century criminology as discourse and practice]*. Göttingen, Germany: Vandenhoeck & Ruprecht.

Birkel, C., Church, D., Erdmann, A., Hager, A., & Leitgöb-Guzy, N. (2020). Sicherheit und Kriminalität in Deutschland–SKiD 2020: Bundesweite Kernbefunde des Viktimisierungssurvey des Bundeskriminalamts und der Polizeien der Länder [Safety and crime in Germany–SKiD 2020: Nationwide findings of the Victimization Survey by the German Federal Office of Criminal Investigation and by the Police of the federal states]. Bundeskriminalamt. Retrieved from https://www.bka.de/DE/UnsereAufgaben/Forschung/ForschungsprojekteUndErgebnisse/Dunkelfeldforschung/SKiD/Ergebnisse/Ergebnisse_node.html

Böckenholt, U., & Van der Heijden, P. G. (2007). Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. *Psychometrika*, *72*(2), 245–262. doi: 10.1007/s11336-005-1495

-y

Boruch, R. F. (1971). Assuring confidentiality of responses in social research: A note on strategies. *The American Sociologist*, *6*(4), 308–311.

Buchman, T. A., & Tracy, J. A. (1982). Obtaining responses to sensitive questions: Conventional questionnaire versus randomized response technique. *Journal of Accounting Research*, *20*(1), 263–271. doi: 10.2307/2490775

Bullek, B., Garboski, S., Mir, D. J., & Peck, E. M. (2017). Towards understanding differential privacy: When do people trust Randomized Response Technique? In *Proceedings of the 2017 chi conference on human factors in computing systems* (pp. 3833–3837). doi: 10.1145/3025453.3025698

Burr, M. L., Butland, B., King, S., & Vaughan-Williams, E. (1989). Changes in asthma prevalence: Two surveys 15 years apart. *Archives of Disease in Childhood*, *64*(10), 1452–1456. doi: 10.1136/adc.64.10.1452

Callegaro, M., Villar, A., Yeager, D., & Krosnick, J. A. (2014). A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples. In M. Callegaro, R. P. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online Panel Research: A Data Quality Perspective* (pp. 23–53). Hoboken, NJ, USA: John Wiley & Sons. doi: 10.1002/9781118763520.ch2

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4), 641–678. doi: 10.1093/poq/nfp075

Chapkovski, P., & Schaub, M. (2022). Do Russians tell the truth when they say they support the war in Ukraine? Evidence from a list experiment. LSE European Politics and Policy (EUROPP) blog. Retrieved from `https://eprints.lse.ac.uk/116770/1/europpblog_2022_04_06_do_russians_tell_the_truth_when_they_say_they.pdf`

Clancy, R., & Taylor, A. (2016). Engaging clinicians in motivational interviewing: Comparing online with face-to-face post-training consolidation. *International Journal of Mental Health Nursing*, *25*(1), 51–61. doi: 10.1111/inm.12184

Clark, S. J., & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods*, *3*(2), 160–168. doi: 10.1037/1082-989X.3.2.160

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. doi: 10.1037/h0047358

Cullen, K. A., Ambrose, B. K., Gentzke, A. S., Apelberg, B. J., Jamal, A., & King, B. A. (2018). Notes from the field: Use of electronic cigarettes and any tobacco product among middle and high school students—United States, 2011–2018. *Morbidity and Mortality Weekly Report*, *67*(45), 1276–1277. doi: 10.15585/mmwr.mm6745a5

De Jong, M. G., Pieters, R., & Fox, J.-P. (2010). Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research*, *47*(1), 14–27. doi: 10.1509/jmkr.47.1.14

De Schrijver, A. (2012). Sample survey on sensitive topics: Investigating respondents' understanding and trust in alternative versions of the randomized response technique. *Journal of Research Practice*, *8*(1), M1.

Dietz, P., Iberl, B., Schuett, E., van Poppel, M., Ulrich, R., & Sattler, M. C. (2018). Prevalence estimates for pharmacological neuroenhancement in Austrian university students: Its relation to health-related risk attitude and the framing effect of caffeine tablets. *Front. Pharmacol.*, *9*(494), 1–9. doi: 10.3389/fphar.2018.00494

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. doi: 10.1016/j.chb.2014.04.005

Dreißigacker, A., & Riesner, L. (2018). Private Internetnutzung und Erfahrung mit computerbezogener Kriminalität [Private internet usage and experience with computer-related crime]. Kriminologisches Forschungsinstitut Niedersachsen e.V. Retrieved from `https://tobias-lib.ub.uni-tuebingen.de/xmlui/bitstream/handle/10900/86122/FB_139.pdf?sequence=1&isAllowed=y`

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, *37*(2), 90. doi: 10.1037/h0058073

Ellrich, K., & Baier, D. (2015). Gewaltausübung durch Polizeibeamte — Ausmaß und Einflussfaktoren [Use of force by police officers — extent and influencing factors]. *Rechtspsychologie*, *1*(1), 22–45. doi: 10.5771/2365-1083-2015-1-22

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. doi: 10.1177/1948550615612150

forsa. (2023). dbb Bürgerbefragung "Öffentlicher Dienst" 2023 — Einschätzungen, Erfahrungen und Erwartungen der Bürger [dbb Citizens' survey "Public Service" 2023 — Citizens' assessments, experiences and expectations]. forsa. Retrieved from `https://www.dbb.de/fileadmin/user_upload/globale_elemente/pdfs/2023/230815_dbb_Buergerbefragung_2023_final.pdf`

Goldenbeld, C., Torfs, K., Vlakveld, W., & Houwing, S. (2020). Impaired driving due to alcohol or drugs: International differences and determinants based on E-Survey of Road Users' Attitudes first-wave results in 32 countries. *IATSS Research*, *44*(3), 188–196. doi: 10.1016/j.iatssr.2020.07.005

Goodstadt, M. S., & Gruson, V. (1975). The randomized response technique: A test on drug use. *Journal of the American Statistical Association*, *70*(352), 814–818. doi: 10.1080/01621459.1975.10480307

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, *64*(326), 520–539.

Greenberg, B. G., Kuebler Jr, R. R., Abernathy, J. R., & Horvitz, D. G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, *66*(334), 243–250. doi: 10.2307/2283916

Hagan, F. E. (1986). *Introduction to criminology — theories, methods, and criminal behavior*. Chicago, IL, USA: Nelson-Hall.

Han, B., Compton, W. M., Gfroerer, J., & McKeon, R. (2015). Prevalence and correlates of past 12-month suicide attempt among adults with past-year suicidal ideation in the United States. *The Journal of Clinical Psychiatry*, *76*(3), 295–302. doi: 10.4088/JCP.14m09287

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, *21*(1), 111–121. doi: 10.1093/ijpor/edn054

Hodgson, J. (2015). Plea bargaining: A comparative analysis. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 226–231). Oxford, UK: Elsevier. doi: 10.1016/B978-0-08-097086-8.86091-2

Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques: A Stochastic Lie Detector versus the Crosswise Model. *Behavior Research Methods*, *48*(3), 1032–1046. doi: 10.3758/s13428-015-0628-6

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *PLOS ONE*, *13*(8), e0201770. doi: 10.1371/journal.pone.0201770

Höglinger, M., Jann, B., & Diekmann, A. (2016). Sensitive questions in online surveys: An experimental evaluation of different implementations of the Randomized Response Technique and the crosswise model. *Survey Research Methods*, *10*(3), 171–187. doi: 10.18148/srm/2016.v10i3.6703

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*(1), 79–125. doi: 10.1086/346010

Houston, J., & Tran, A. (2001). A survey of tax evasion using the randomized response technique. In *Advances in Taxation* (pp. 69–94). Leeds, UK: Emerald Group Publishing Limited. doi: 10.1016/S1058-7497(01)13007-3

Iberl, B. (2021). Ein, zwei Bier und ab ans Lenkrad? — Prävalenzschätzung von Alkohol am Steuer durch das Unrelated Question Model [One or two drinks before going for a ride? — Prevalence estimation of driving under the influence via the Unrelated Question Model]. *Kriminologie–Das Online-Journal [Criminology-The Online Journal]*, *3*(3), 270–292. doi: 10.18716/ojs/krimoj/2021.3.5

Iberl, B., Aljovic, A., Ulrich, R., & Reiber, F. (2024). The Poisson Extension of the Unrelated Question Model: Improving surveys with time-constrained questions on sensitive topics. *Survey Research Methods*, *18*(1), 21–38. doi: 10.18148/srm/2024.v18i1.8252

Iberl, B., & Kinzig, J. (2022). Nemo tenetur se ipsum accusare — Systematische Antwortverzerrungen bei der Befragung justizieller Akteure zur Verständigung im Strafprozess [You have the right to remain silent — Systematic response bias of legal practitioners in a survey about plea bargaining]. *RPsych Rechtspsychologie*, *8*(4), 499–517. doi: 10.5771/2365-1083-2022-4-499

Iberl, B., & Ulrich, R. (2023). On estimating the frequency of a target behavior from time-constrained yes/no survey questions: A parametric approach based on the Poisson process. *Psychological Methods*, Advance online publication. doi: 10.1037/met0000588

Isolauri, J., & Laippala, P. (1995). Prevalence of symptoms suggestive of gastroesophageal reflux disease in an adult population. *Annals of Medicine*, *27*(1), 67–70. doi: 10.3109/07853899509031939

Kerner, H.-J., Stroezel, H., & Wegel, M. (2011). Gewaltdelinquenz und Gewaltaffinität bei jungen Menschen in verschiedenen sozialen Milieus [Violent delinquency and affinity for violence among young people in different social milieus]. *Trauma & Gewalt*, *5*(1), 20–35.

Kinzig, J., Iberl, B., & Koch, J. (2020). Online-Befragung justizieller Akteure (Modul 4) [Online survey of judicial actors (Module 4)]. In K. Altenhain, M. Jahn, & J. Kinzig (Eds.), *Die Praxis der Verständigung im Strafverfahren [The practice of plea bargaining in criminal proceedings]* (pp. 191–305). Baden-Baden, Germany: Nomos. doi: 10.5771/9783748922094

Koerth, K. (2019, 05 03). Personalnot: Die Justiz sieht alt aus [Staff shortage: The judiciary looks old]. *Spiegel Online*. Retrieved from `https://www.spiegel.de/karriere/arbeitsueberlastung-im-gericht-warum-die-justiz-alt-aussieht-a-1265194.html`

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. doi: 10.1093/poq/nfn063

Krüger, H.-P., & Vollrath, M. (1998). Fahren unter Alkohol in Deutschland: Die Ergebnisse des Deutschen Roadside Surveys [Driving under the influence in Germany: The results of the German Roadside Survey]. In H.-P. Krüger (Ed.), *Fahren unter Alkohol in Deutschland [Driving under the influence in Germany]* (pp. 33–57). Stuttgart, Germany: Gustav Fischer.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*(4), 2025–2047. doi: 10.1007/s11135-011-9640-9

Landsheer, J. A., Van Der Heijden, P., & Van Gils, G. (1999). Trust and understanding, two psychological aspects of Randomized Response. *Quality and Quantity*, *33*(1), 1–12.

Langbein, J. H. (2022). The turn to confession bargaining in German criminal procedure: Causes and comparisons with American plea bargaining. *The American Journal of Comparative Law*, *70*(1), 139–161. doi: 10.1093/ajcl/avac025

Lee, R. M. (1993). *Doing research on sensitive topics*. Thousand Oaks, CA, USA: Sage.

Lensvelt-Mulders, G. J., & Boeije, H. R. (2007). Evaluating compliance with a computer assisted randomized response technique: A qualitative study into the origins of lying and cheating. *Computers in Human Behavior*, *23*(1), 591–608. doi: 10.1016/j.chb.2004.11.001

Lensvelt-Mulders, G. J., Hox, J. J., Van der Heijden, P. G., & Maas, C. J. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research*, *33*(3), 319–348. doi: 10.1177/0049124104268664

Linton, S. J., Hellsing, A.-L., & Halldén, K. (1998). A population-based study of spinal pain among 35-45-year-old individuals: Prevalence, sick leave, and health care use. *Spine*, *23*(13), 1457–1463.

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*, *47*(2), 519–528. doi: 10.3758/s13428-014-0483-x

Lutz, P., Stelly, W., Bartsch, T., Thomas, J., & Bergmann, B. (2021). Islamische Seelsorge im Jugendstrafvollzug [Islamic pastoral care in juvenile detention centers]. *Kriminologie-Das Online-Journal | Criminology-The Online Journal*, *3*(3), 228–248. doi: 10.18716/ojs/krimoj/2021.3.3

McCabe, S. E., Cranford, J. A., & Boyd, C. J. (2006). The relationship between past-year drinking behaviors and nonmedical use of prescription drugs: Prevalence of co-occurrence in a national sample. *Drug and Alcohol Dependence*, *84*(3), 281–288. doi: 10.1016/j.drugalcdep.2006.03.006

McKetin, R., McLaren, J., Lubman, D. I., & Hides, L. (2006). The prevalence of psychotic symptoms among methamphetamine users. *Addiction*, *101*(10), 1473–1478. doi: 10.1111/j.1360-0443.2006.01496.x

Meisters, J., Hoffmann, A., & Musch, J. (2020). Can detailed instructions and comprehension checks increase the validity of crosswise model estimates? *PLOS ONE*, *15*(6), e0235403. doi: 10.1371/journal.pone.0235403

Meisters, J., Hoffmann, A., & Musch, J. (2022a). More than random responding: Empirical evidence for the validity of the (Extended) Crosswise Model. *Behavior Research Methods*, *55*(2), 716–729. doi: 10.3758/s13428-022-01819-2

Meisters, J., Hoffmann, A., & Musch, J. (2022b). A new approach to detecting cheating in sensitive surveys: The Cheating Detection Triangular Model. *Sociological Methods & Research*, *53*(1), 328–368. doi: 10.1177/00491241211055764

Miller, C., Ettridge, K., Wakefield, M., Pettigrew, S., Coveney, J., Roder, D., … Dono, J. (2020). Consumption of sugar-sweetened beverages, juice, artificially-sweetened soda and bottled water: An Australian population study. *Nutrients*, *12*(3), Article 817. doi: 10.3390/nu12030817

Miller, J. D. (1984). *A new survey technique for studying deviant behavior*. Washington, D.C., USA: The George Washington University.

Molinaro, S., Benedetti, E., Scalese, M., Bastiani, L., Fortunato, L., Cerrai, S., … others (2018). Prevalence of youth gambling and potential influence of substance use and other risk factors throughout 33 European countries: First results from the 2015 ESPAD study. *Addiction*, *113*(10), 1862-1873. doi: 10.1111/add.14275

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods*, *44*(1), 222–231. doi: 10.3758/s13428-011-0144-2

Musch, J., Bröder, A., & Klauer, K. C. (2001). Improving survey research on the World-Wide Web using the randomized response technique. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet Science* (pp. 179–192). Lengerich, Germany: Pabst Science Publishers.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. doi: 10.1002/ejsp.2420150303

Neubacher, F. (2020). *Kriminologie [Criminology]*. Baden-Baden, Germany: Nomos.

Oliker, O. (2017). Putinism, populism and the defence of liberal democracy. *Survival*, *59*(1), 7–24. doi: 10.1080/00396338.2017.1282669

Ostapczuk, M., Musch, J., & Moshagen, M. (2011). Improving self-report measures of medication non-adherence using a cheating detection extension of the randomised-response-technique. *Statistical Methods in Medical Research*, *20*(5), 489–503. doi: 10.1177/0962280210372843

Rasinski, K. A., Willis, G. B., Baldwin, A. K., Yeh, W., & Lee, L. (1999). Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *13*(5), 465–484.

doi: 10.1002/(SICI)1099-0720(199910)13:5<465::AID-ACP609>3.0.CO;2-Y

Reiber, F., Bryce, D., & Ulrich, R. (2022). Self-protecting responses in randomized response designs: A survey on intimate partner violence during the coronavirus disease 2019 pandemic. *Sociological Methods & Resarch*, Advance online publication. doi: 10.1177/00491241211043138

Reiber, F., Pope, H., & Ulrich, R. (2020). Cheater detection using the unrelated question model. *Sociological Methods & Research*, Advance online publication. doi: 10.1177/0049124120914919

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. doi: 10.1007/s10508-017-1044-z

Sawyer, A. N., Smith, E. R., & Benotsch, E. G. (2018). Dating application use and sexual risk behavior among young adults. *Sexuality Research and Social Policy*, *15*(2), 183–191. doi: 10.1007/s13178-017-0297-6

Seitz, N.-N., Rauschert, C., Atzendorf, J., & Kraus, L. (2020). IFT-Berichte Bd. 190: Berlin, Hessen, Nordrhein-Westfalen, Sachsen und Thüringen. Ergebnisse des Epidemiologischen Suchtsurvey 2018 [IFT-Reports Vol. 190: Substance use and substance use disorders in Berlin, Hesse, North Rhine-Westphalia, Saxony and Thuringia. Results of the 2018 Epidemiological Survey of Substance Abuse]. München, Germany: Institut für Therapieforschung München. Retrieved from `https://www.esa-survey.de/fileadmin/user_upload/esa_laenderberichte/Bd_190_ESA_2018_Bundeslaender.pdf`

Snodgrass, J. (1976). Clifford R. Shaw and Henry D. McKay: Chicago Criminologists. *The British Journal of Criminology*, *16*(4), 1–19.

Soeken, K. L., & Damrosch, S. P. (1986). Randomized response technique: Applications to research on rape. *Psychology of Women Quarterly*, *10*(2), 119–126. doi: 10.1111/j.1471-6402.1986.tb00740.x

Soga, M., Evans, M. J., Tsuchiya, K., & Fukano, Y. (2021). A room with a green view: The importance of nearby nature for mental health during the COVID-19 pandemic. *Ecological Applications*, *31*(2), Article e2248. doi: 10.1002/eap.2248

Solomon, J., Jacobson, S. K., Wald, K. D., & Gavin, M. (2007). Estimating illegal resource use at a Ugandan park with the randomized response technique. *Human Dimensions of Wildlife*, *12*(2), 75–88. doi: 10.1080/10871200701195365

Sutherland, E. H., Cressey, D. R., & Luckenbill, D. F. (1992). *Principles of Criminology*. Lanham, MD, USA: Altamira Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. doi: 10.1037/0033-2909.133.5.859

Treibel, A., Dölling, D., & Hermann, D. (2017). Determinants of reporting crimes against sexual self-determination. *Forensische Psychiatrie, Psychologie, Kriminologie*, *11*(4), 355–363. doi: 10.1007/s11757-017-0438-z

Treibel, A., & Funke, J. (2004). Die internetbasierte Opferbefragung als Instrument der Dunkelfeldforschung — Grenzen und Chancen [The internet-based victim survey as an instrument of dark field research — limits and opportunities]. *Monatsschrift für Kriminologie und Strafrechtsreform*, *87*(2), 146–151. doi: 10.1515/mks-2004-00018

Ulrich, R., Pope, H. G., Cléret, L., Petróczi, A., Nepusz, T., Schaffer, J., … Simon, P. (2018). Doping in two elite athletics competitions assessed by randomized-response surveys. *Sports Medicine*, *48*(1), 211–219. doi: 10.1007/s40279-017-0765-4

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking sensitive questions: A statistical power analysis of randomized response models. *Psychological Methods*, *17*(4), 623–641. doi: 10.1037/a0029314

Umesh, U. N., & Peterson, R. A. (1991). A critical evaluation of the randomized response method: Applications, validation, and research agenda. *Sociological Methods & Research*, *20*(1), 104–138. doi: 10.1177/0049124191020001004

Virudachalam, S., Long, J. A., Harhay, M. O., Polsky, D. E., & Feudtner, C. (2014). Prevalence and patterns of cooking dinner at home in the USA: National Health and Nutrition Examination Survey (NHANES) 2007–2008. *Public Health Nutrition*, *17*(5), 1022–1030. doi: 10.1017/S1368980013002589

Walzenbach, S., & Hinz, T. (2019). Pouring water into wine: Revisiting the advantages of the Crosswise Model for asking sensitive questions. *Survey Methods: Insights from the Field*, 1–16. doi: 10.13094/SMIF-2019-00002

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63–69.

Wilson, J. Q., & Kelling, G. L. (2017). The police and neighborhood safety Broken Windows. In J. T. Walker (Ed.), *Social, Ecological and Environmental Theories of Crime* (pp. 169–178). London, UK: Routledge. doi: 10.4324/9781315087863

Wittenberg, J., Reinecke, J., & Boers, K. (2009). Verbreitung, Entwicklung und Erklärung von Delinquenz im Jugendalter. Ergebnisse einer aktuellen Längsschnittstudie [Prevalence, development and explanation of delinquency in adolescence. Results of a current longitudinal study]. *Journal for Educational Research Online*, *1*(1), 106–134. doi: 10.25656/01:4558

Wolter, F. (2012). *Heikle Fragen in Interviews — Eine Validierung der Randomized Response-Technik [Sensitive questions in interviews — a validation of the randomized response technique]*. Wiesbaden, Germany: Springer VS. doi: 10.1007/978-3-531-19371-7

Wolter, F., & Diekmann, A. (2021). False positives and the "more-is-better" assumption in sensitive question research: New evidence on the crosswise model and the item count technique. *Public Opinion Quarterly*, *85*(3), 836–863. doi:

10.1093/poq/nfab043

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, *67*(3), 251–263. doi: 10.1007/s00184-007-0131-x

Zhang, B., & Gearhart, S. (2020). Collecting online survey data: A comparison of data quality among a commercial panel & MTurk. *Survey Practice*, *13*(1). doi: 10.29115/SP-2020-0015

Zhang, X., Kuchinke, L., Woud, M. L., Velten, J., & Margraf, J. (2017). Survey method matters: Online/offline questionnaires and face-to-face or telephone interviews differ. *Computers in Human Behavior*, *71*, 172–180. doi: 10.1016/j.chb.2017.02.006

A Paper 1

Iberl, B., & Kinzig, J. (2022). Nemo tenetur se ipsum accusare - Systematische Antwortverzerrungen bei der Befragung justizieller Akteure zur Verständigung im Strafprozess [You have the right to remain silent - Systematic response bias of legal practitioners in a survey about plea bargaining]. *Rechtspsychologie, 8*(4), 499-517. https://doi.org/10.5771/2365-1083-2022-4-499.

Candidate contributions to the article

Status	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Published	100%	100%	90%	75%

Nemo tenetur se ipsum accusare – Systematische Antwortverzerrungen bei der Befragung justizieller Akteure zur Verständigung im Strafprozess

Benedikt Iberl

Jörg Kinzig

Institut für Kriminologie | Eberhard Karls Universität Tübingen

Nemo tenetur se ipsum accusare – Systematische Antwortverzerrungen bei der Befragung justizieller Akteure zur Verständigung im Strafprozess

Zusammenfassung

Der vorliegende Artikel stützt sich auf die Daten eines bereits abgeschlossenen Forschungsprojekts zur Verständigung im Strafprozess. In den Ergebnissen dieser Studie finden sich im Antwortverhalten deutliche Berufsgruppenunterschiede und mehrere Hinweise für das Vorliegen einer systematischen Verzerrung durch den Effekt der sozialen Erwünschtheit, insbesondere bei Richtern. Aus dem Forschungsstand werden Erklärungsansätze für diese Beobachtung abgeleitet und erläutert. Abschließend erfolgt die Vorstellung von Randomized Response Modellen als Lösungsansatz für den Umgang mit systematischen Antwortverzerrungen durch sensible Fragen in der kriminologischen Forschung.

Schlüsselworte: Soziale Erwünschtheit, Richter, Verständigung im Strafprozess, Absprachen, Randomized Response Technique

You have the right to remain silent - Systematic response bias of legal practitioners in a survey about plea bargaining

Abstract

The paper at hand is based on data from an already completed research project regarding plea bargaining. The results of this study contain clear differences in response behavior across professions and multiple hints towards the existence of systematic social desirability bias, especially within judges. Different explanatory approaches for this observation are derived and explained from the current state of research. Finally, randomized response models are presented as a possible solution for handling systematic response bias caused by sensitive questions in criminological research.

Keywords: social desirability, judges, plea bargaining, randomized response technique

Anmerkung der Autoren

Bei diesem Manuskript handelt es sich um eine von den Autoren formatierte Version des Artikels ***Nemo tenetur se ipsum accusare - Systematische Antwortverzerrungen bei der Befragung justizieller Akteure zur Verständigung im Strafprozess***, der erstmals in der Zeitschrift ***Rechtspsychologie*** (Jahr 2022, Heft 4, Jahrgang 4, doi.org/10.5771/2365-1083-2022-4-499) des Verlags ***Nomos*** (Baden-Baden) veröffentlicht wurde.

This manuscript is a version of the original article that was formated by the authors. The article was published in the journal *Rechtspsychologie [Forensic Psychology]* (year 2022, issue 4, volume 8, doi.org/10.5771/2365-1083-2022-4-499) by the publisher *Nomos* (Baden-Baden, Germany).

1. Einleitung

1.1 „Deals“ in Strafverfahren

Unter „Deals“ in Strafverfahren wird sich jeder etwas vorstellen können, der sich schon einmal die Sendungen „Suits“, „Better Call Saul“ oder andere US-amerikanische Anwaltsserien zu Gemüte geführt hat. Um ein aufwendiges Verfahren zu vermeiden und einen bestmöglichen Prozessausgang zu erreichen, werden dort bei informellen Verhandlungen alle Register gezogen: Von listigen Tricks und Bluffs über Erpressungsversuche bis hin zu Betrügereien – oder gar der Sabotage eines Aufzugs, um mit der vielbeschäftigten Staatsanwältin in Ruhe feilschen zu können („Better Call Saul“, Staffel 5, Folge 2). Verlässt man das Gebiet der Fiktion und blickt auf „echte“ Strafverfahren in der Bundesrepublik Deutschland, erscheinen solche Vorgänge zunächst realitätsfern. Sicherlich – so ist zu vermuten – existieren in Deutschland hierzu strenge Regeln, die penibel eingehalten werden und durch die unseriösen „Deals“ in Richterzimmern[1] (und Aufzügen) ein Riegel vorgeschoben wird – oder etwa nicht?

Tatsächlich trifft auf einer formalen Ebene die erste Annahme zu: Zwar erst seit dem Jahr 2009 (BGBl. I 2009, S. 2353), aber durchaus detailliert, ist die „Verständigung im Strafprozess“ in Deutschland gesetzlich geregelt. Dass sich Verteidigung, Staatsanwaltschaft und Gericht über ein Urteil absprechen oder, wie es das Gesetz formuliert, „verständigen“ und somit den Prozess wesentlich verkürzen können, ist unter Einhaltung der gesetzlichen Vorgaben erlaubt. Dazu gehören u. a. aufwendige Transparenz- und Dokumentationspflichten, etwa darüber, ob, wann und durch wen die Verständigung initiiert und über was genau sich inhaltlich geeinigt wurde. Auch darf keine genaue Strafe abgesprochen werden; lediglich ein Strafrahmen kann Gegenstand der

[1] Aufgrund der in diesem Beitrag sehr häufigen Nennung von Richterinnen und Richtern, Staatsanwältinnen und Staatsanwälten und Strafverteidigerinnen und Strafverteidigern wurde sich zur Förderung des Leseflusses für die Verwendung des generischen Maskulinums entschieden. Dies soll selbstverständlich alle Geschlechtszugehörigkeiten beinhalten.

Verständigung sein. Außerdem ist es Staatsanwaltschaft und Verteidigung verboten, bereits vor dem Urteilsspruch zuzusichern, auf Rechtsmittel verzichten zu wollen - d.h., niemand darf im Voraus garantieren, das auf eine Verständigung folgende Urteil auf keinen Fall anzufechten.

1.2 Das Forschungsprojekt „Die Praxis der Verständigung im Strafprozess“

Soweit zu den gesetzlichen Prämissen und deren Intentionen. Doch wie sieht es mit der Einhaltung dieser Vorgaben aus? Das beschäftigte auch bereits das Bundesverfassungsgericht, das den Gesetzgeber schon im Jahr 2013 in einem aufsehenerregenden Urteil (BVerfGE 133, 168) dazu verpflichtete, die „Wirksamkeit der vorgesehenen Schutzmechanismen“ zur Einhaltung der in der Strafprozessordnung geregelten Vorgaben fortlaufend zu überprüfen. In der Folge beauftragte das Bundesjustizministerium die Universitäten Düsseldorf (Prof. Dr. Karsten Altenhain), Frankfurt (Prof. Dr. Matthias Jahn) und Tübingen (Prof. Dr. Jörg Kinzig) im Jahr 2018, die Einhaltung der gesetzlichen Regelungen zur Verständigung in der juristischen Praxis zu untersuchen (Altenhain, Jahn & Kinzig, 2020).

Die zugrundeliegende Forschungsfrage wurde im Rahmen des Projekts in sechs Teilmodulen aus mehreren Perspektiven und mithilfe verschiedener Methoden (u.a. Interviews, schriftliche und Online-Fragebögen) beantwortet (Altenhain, Jahn & Kinzig, 2020). Der Forschungsverbund kam dabei modulübergreifend zu einem identischen Ergebnis: Informelle Absprachen, bei denen sich die justiziellen Akteure nicht an das Gesetz halten, finden in Deutschland nach wie vor regelmäßig statt. Es besteht damit, ebenso wie vor dem Urteil des Bundesverfassungsgerichts aus dem Jahr 2013 (Altenhain, Dietmeier & May, 2013), unverändert ein eindeutiges Defizit bei der Einhaltung der Vorgaben der Strafprozessordnung. Teil des genannten Forschungsprojekts war unter anderem eine bundesweite quantitative Online-Befragung (Kinzig, Iberl & Koch, 2020), bei der Antworten von 1567 Richtern, Staatsanwälten und Strafverteidigern ausgewertet werden konnten. Diese Online-Befragung liefert die Datengrundlage für die nachfolgenden

Ausführungen.

1.3 Online-Befragung justizieller Akteure

Ein markanter Befund, der bei den Ergebnissen der Online-Befragung (und bei den Interviews; Altenhain, Brandt & Herbst, 2020) konsistent auftritt, ist, dass deutliche Unterschiede im Antwortverhalten zwischen den befragten Berufsgruppen bestehen (Kinzig, Iberl & Koch, 2020). Exemplarisch zeigt sich dies an den Reaktionen auf die Fragen nach der Häufigkeit sogenannter informeller Absprachen – selbige entsprechen qua Definition nicht den gesetzlichen Anforderungen – nach dem Hörensagen und in der eigenen Praxis (s. Abb. 1).

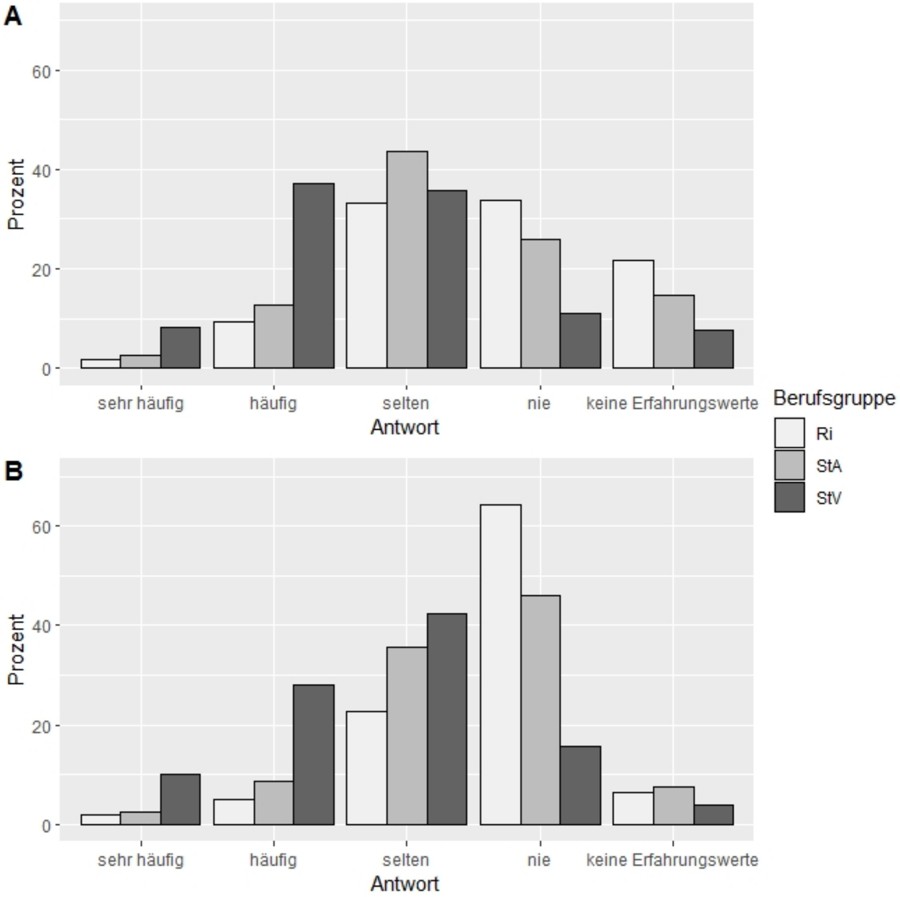
Wie in Abbildung 1) deutlich zu erkennen, berichten die Strafverteidiger sowohl nach dem Hörensagen als auch in der eigenen Praxis von den meisten informellen Absprachen, wohingegen die Richter mit weitem Abstand angeben, am wenigsten von Regelverstößen Kenntnis zu erlangen oder daran gar beteiligt zu sein. Dieses Ergebnis ist nicht plausibel, da für erfolgreiche Absprachen alle Verfahrensbeteiligten konsensual zusammenwirken müssen. Strafverteidiger und Richter scheinen also in verschiedenen Realitäten unterwegs zu sein, zumal Absprachen ohne eine Beteiligung des Gerichts, das am Ende das Urteil fällen muss, schlichtweg nicht möglich sind.

Dieses Antwortmuster – Strafverteidiger und Richter liegen mit ihren Antworten nicht selten extrem weit voneinander entfernt, die Staatsanwälte regelmäßig dazwischen – zeigt sich bei den meisten im Rahmen des Online-Surveys gestellten Fragen. Auffallend ist dabei, dass die Antworten der Strafverteidiger eher gesetzeswidrige Vorgänge nahelegen, während die der Richter eher ordnungsgemäße Abläufe als die Norm darstellen. Als mögliche Erklärung dieses Befundes drängt sich geradezu auf, dass die Angaben der Richter (und in geringerem Maße auch die der Staatsanwälte) durch den Effekt der sozialen Erwünschtheit beeinflusst worden sind.

Natürlich sind auch andere Erklärungsansätze denkbar. Die Vorschriften zur Verständigung könnten etwa einen gewissen Interpretationsspielraum zulassen. Wenn beim

Abbildung 1

Angaben zu den Häufigkeiten informeller Absprachen; Antwortverteilungen nach Berufsgruppen in Prozent. A: Häufigkeit informeller Absprachen nach dem Hörensagen. B: Häufigkeit informeller Absprachen in der eigenen Praxis. „Ri“ = Richter, „StA“ = Staatsanwälte, „StV“ = Strafverteidiger. Unterschiede zwischen Berufsgruppen sind jeweils statistisch signifikant (s. Kinzig, Iberl & Koch, 2020, S. 237).



Verständnis der gesetzlichen Regelungen zwischen den Berufsgruppen systematische Unterschiede vorliegen, dürften auch die Einschätzungen, ab wann es sich bei einem Vorgang um eine informelle Absprache handelt, voneinander abweichen. Dessen ungeachtet lassen sich klare Hinweise auf Antwortverzerrungen durch das Phänomen sozialer Erwünschtheit identifizieren. Diese werden im Folgenden vorgestellt.

2. Soziale Erwünschtheit bei Richtern?

2.1 Der Begriff der sozialen Erwünschtheit

Der Effekt der sozialen Erwünschtheit beschreibt eine Beeinflussung des Antwortverhaltens im Rahmen von Befragungen. Dabei tendieren Personen zu Angaben, die sie als sozial gebilligt bzw. als der gesellschaftlichen Norm entsprechend wahrnehmen (Crowne & Marlowe, 1960; Edwards, 1953; Häcker & Stapf, 2009; Wenninger, 2001). Besonders bei Fragen zu sensiblen Themen können durch diesen Effekt beachtliche Verzerrungen auftreten, wodurch eine Unterschätzung der Prävalenzen sensitiver Attribute entsteht (Krumpal, 2013; Lee, 1993; Nederhof, 1985; Tourangeau & Yan, 2007). Welche Fragen bzw. Attribute als heikel wahrgenommen werden, kann selbstverständlich abhängig von der befragten Bevölkerungsgruppe variieren (Tourangeau & Yan, 2007). In Bezug auf den vorliegenden Untersuchungsgegenstand ist die Annahme naheliegend, dass justizielle Akteure Fragen nach der Beteiligung an informellen, also gesetzwidrigen Absprachen als heikel betrachten. Immerhin handelt es sich dabei um Verstöße, die berufs- oder sogar strafrechtliche Konsequenzen nach sich ziehen können.

Das Vorkommen sozial erwünschter Antworten wird unter anderem dadurch beeinflusst, inwieweit die Befragten befürchten, dass ehrliche Antworten negative Konsequenzen zur Folge haben (Krumpal, 2013; Lee, 1993; Rasinski et al., 1999; Tourangeau & Yan, 2007). Dabei spielen sowohl direkte Konsequenzen, z. B. strafrechtlicher Art nach einer etwaigen Verletzung der Anonymität der Teilnehmenden, eine Rolle, als auch befürchtete indirekte Folgen. Im Fall informeller Absprachen könnten beispielsweise schärfere gesetzliche Vorgaben „de lege ferenda“ oder ein Ansehensverlust des betreffenden Berufsstandes als indirekte Folgen befürchtet werden. Es ist indes wichtig zu erwähnen, dass sozial erwünschte Antworten nicht auf bewussten Falschangaben basieren müssen, sondern auch in Folge einer Selbsttäuschung auftreten können, um im Sinne einer kognitiven Dissonanzreduktion das Selbstbild im Kontext eigener Wertevorstellungen zu wahren (Holtgraves, 2004; Krumpal & Näher, 2012; Krumpal, 2013;

Nederhof, 1985; Paulhus, 1984; Stocké & Hunkler, 2007; Wenninger, 2001).

2.2 Hinweise auf soziale Erwünschtheit

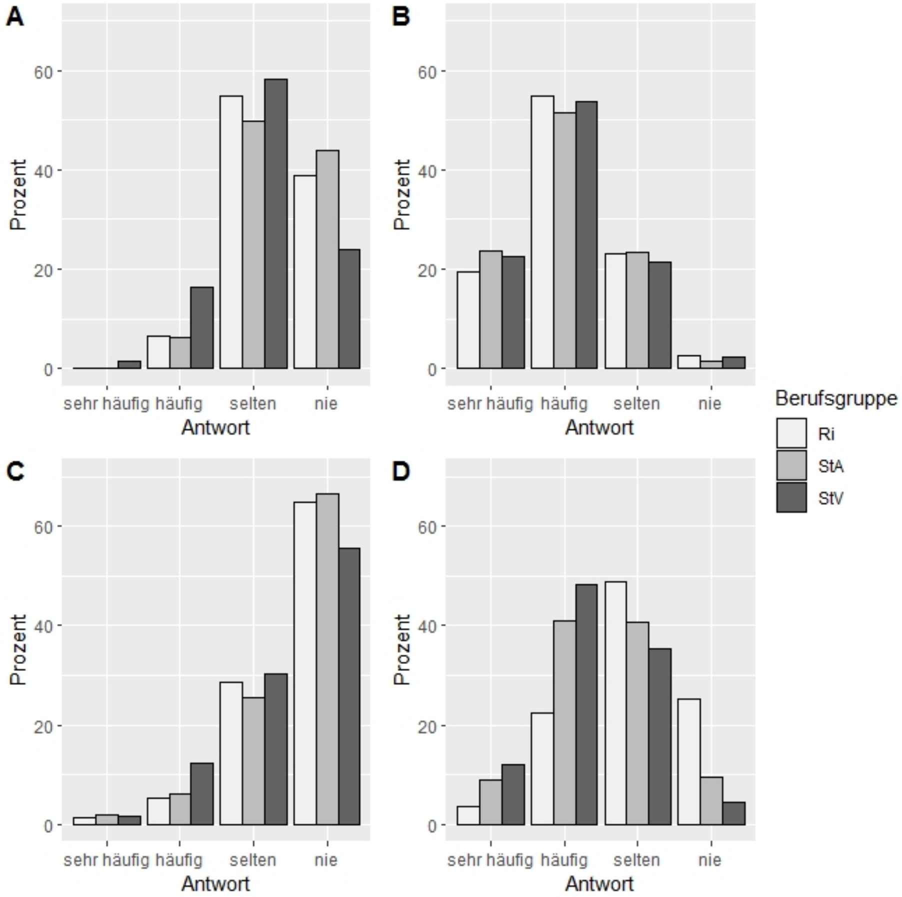
Freilich ist allein die schlichte Beobachtung, dass eine Berufsgruppe die „sozial erwünschtesten“ Antworten gibt, noch kein hinreichender Nachweis, um das Vorliegen einer nennenswerten Antwortverzerrung zu belegen. Jedoch existieren darüber hinaus weitere Indizien dafür, dass der Effekt der sozialen Erwünschtheit bei der Befragung eine Rolle gespielt und vor allem die Antworten der Richter beeinflusst haben könnte.

So wurden die justiziellen Akteure danach gefragt, wie häufig die Initiative zu informellen Absprachen von der Verteidigung, der Staatsanwaltschaft, den Angeklagten oder dem Gericht ausgeht (s. Abb. 2)). Auffällig ist dabei, dass die Selbst- und Fremdeinschätzung bei der Initiierung informeller Absprachen durch Strafverteidiger übereinstimmt. Strafverteidiger, Richter und Staatsanwälte antworten also sehr ähnlich auf die Frage, wie oft informelle Absprachen durch die Verteidigung angeregt werden. Dies spricht für die Validität der Antworten der Strafverteidiger. Gleichzeitig sind jedoch signifikante Unterschiede zwischen Selbst- und Fremdeinschätzungen bei den anderen Berufsgruppen zu beobachten (Kinzig, Iberl & Koch, 2020). Rein deskriptiv sind diese Abweichungen bei Richtern besonders groß. Das bedeutet, dass Richter die Initiierung informeller Absprachen durch ihre Berufsgruppe als deutlich seltener beschreiben als Strafverteidiger und Staatsanwälte. Solche Abweichungen zwischen Selbst- und Fremdeinschätzung können durchaus als Hinweis dafür gewertet werden, dass die Antworten der Richter (und in geringerem Maß auch der Staatsanwälte) in sozial erwünschter Richtung verzerrt sind.

Weiterhin ergibt sich in Bezug auf die zu befürchtenden negativen Konsequenzen ein Indiz dafür, dass insbesondere bei Richtern ein Effekt sozialer Erwünschtheit vorliegt. Zwei Items des Fragebogens zielten darauf ab, eine Risikoeinschätzung bezüglich negativer Konsequenzen im Fall gesetzwidrigen Verhaltens zu erheben (s. Abb. 3)). Eine der Fragen thematisierte das erwartete Risiko, dass eine informelle Absprache zu einer Beanstandung

Abbildung 2

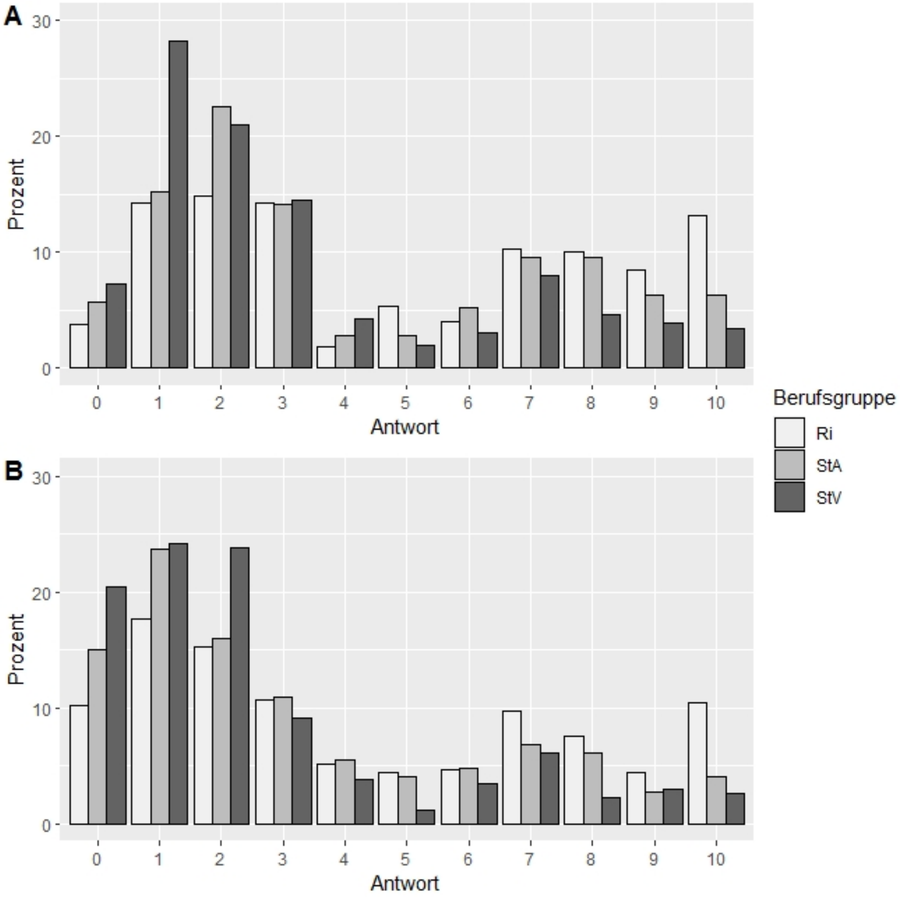
Angaben zur Initiierung informeller Absprachen durch verschiedene Akteure; Antwortverteilungen nach Berufsgruppen in Prozent. A: Häufigkeit der Initiierung durch die Staatsanwaltschaft. B: Häufigkeit der Initiierung durch die Verteidigung. C: Häufigkeit der Initiierung durch Angeklagte. D: Häufigkeit der Initiierung durch das Gericht. „Ri“ = Richter, „StA“ = Staatsanwälte, „StV“ = Strafverteidiger. Unterschiede zwischen Berufsgruppen sind statistisch signifikant bei Initiierung durch Staatsanwaltschaft (A) und Gericht (D) und nicht statistisch signifikant bei Initiierung durch Verteidigung (B) und Angeklagte (C) (s. Kinzig, Iberl & Koch, 2020, S. 244).



im Rechtsmittelverfahren führt – dies könnte beispielsweise bedeuten, dass ein auf einer informellen Absprache basierendes Urteil von einer höheren Instanz aufgehoben wird. Die andere Frage ventilierte die Befürchtung strafrechtlicher Konsequenzen nach einer

Abbildung 3

Angaben zur Befürchtung negativer Konsequenzen nach erfolgter informeller Absprache; Antwortverteilungen nach Berufsgruppen in Prozent. A: Einschätzung des Risikos einer Beanstandung des Urteils im Rechtsmittelverfahren. B: Einschätzung des Risikos strafrechtlicher Konsequenzen. Die Bewertung erfolgte jeweils auf einer elfstufigen Likert-Skala von 0 (niedriges Risiko) bis 10 (hohes Risiko). „Ri“ = Richter, „StA“ = Staatsanwälte, „StV“ = Strafverteidiger. Unterschiede zwischen Berufsgruppen sind statistisch signifikant (s. Kinzig, Iberl & Koch, 2020, S. 257).



informellen Absprache. Auch hier zeigen sich deutliche Unterschiede in der Einschätzung der einzelnen Berufsgruppen. So stufen die Richter die Risiken informeller Absprachen als bedeutend höher ein als die anderen Berufsgruppen. Offenbar halten sie derartige negative Folgen für wahrscheinlicher als die anderen Berufsgruppen, was wiederum die These stützt, dass der Effekt sozialer Erwünschtheit bei ihnen besonders stark ausgeprägt ist.

Bemerkenswert ist auch der unterschiedliche Umgang der Berufsgruppen mit der Auswahl der ausweichenden Antwortoption „keine Erfahrungswerte“, die bei den meisten Fragen zur Verfügung stand. Insgesamt antworteten Richter im Schnitt 14,3 Mal mit „keine Erfahrungswerte“, Staatsanwälte 13,6 Mal, Strafverteidiger aber nur 6,8 Mal. Strafverteidiger wählten also weniger als halb so oft die Kategorie „keine Erfahrungswerte“ aus als Angehörige der beiden anderen Berufsgruppen. Dieser Befund lässt sich in unterschiedliche Richtungen interpretieren. So könnte dieses Antwortverhalten möglicherweise Ausdruck einer angemessenen Zurückhaltung der Richter und Staatsanwälte sein, welche im Vergleich zu Strafverteidigern über einen weniger breiten Überblick über die Praxis an verschiedenen Gerichten und Spruchkörpern verfügen dürften. Es ist jedoch auch möglich, dass durch diese Abstinenz die Beantwortung etwaiger sensibler Fragen vermieden werden sollte. Insgesamt kann dies ebenfalls als Hinweis darauf gewertet werden, dass Richter und Staatsanwälte eher gemäß einer sozial erwünschten Norm antworten als Strafverteidiger.

2.3 Mögliche Erklärungen für die erhöhte soziale Erwünschtheit bei Richtern

Ausgehend von den bisher geschilderten Beobachtungen stellt sich die Frage, weshalb Richter womöglich sozial erwünschter antworten als andere Berufsgruppen, insbesondere Strafverteidiger. Bis dato existieren zu dieser Fragestellung (soweit ersichtlich) keinerlei Befunde. Nachfolgend sollen erste Erklärungsansätze geliefert werden.

Einiges spricht dafür, dass sozial erwünschte Antworten in der Gruppe der Richter deswegen in einem erhöhten Maße auftreten, weil die befürchteten indirekten negativen Konsequenzen für diese Berufsgruppe bei einem der Realität gerecht werdenden Antwortverhalten am größten sind. Denn Gesetzesänderungen oder strengere Kontrollmechanismen, die eine ungeschminkte Wiedergabe der realen Vorgänge in den Gerichtssälen zur Folge haben könnten, dürften vor allem Richter betreffen. Schließlich tragen diese durch ihr Urteil am Ende eines Verfahrens unzweifelhaft die

Letztverantwortung für das erfolgreiche Zustandekommen informeller Absprachen. Dabei werden bereits jetzt die Gesetze zur Verständigung seitens der justiziellen Akteure als undurchsichtig oder praxisuntauglich angesehen (Altenhain, Brandt & Herbst, 2020, S. 368 f.; Kinzig, Iberl & Koch, 2020, S. 251). Auch weil sie aus Sicht der anderen Berufsgruppen besonders von informellen Absprachen profitieren (Altenhain, Brandt & Herbst, 2020, S. 335 f.), würden schärfere Regeln oder zusätzliche Kontrollmechanismen wohl primär zulasten der Richter gehen. So könnte etwa ihre ohnehin hohe Arbeitsbelastung (vgl. etwa Koerth, 2019) dadurch noch weiter ansteigen.

Eine weitere Erklärung leitet sich aus der besonderen rechtlichen und gesellschaftlichen Stellung ab, die mit dem Richterberuf einhergeht, wobei sich diese Erwägung mit Abstrichen auch auf Staatsanwälte übertragen lässt. So sind, um sich für diesen Beruf zu qualifizieren, hohe Hürden zu überwinden. Denn in der Regel braucht man überdurchschnittlich gute Staatsexamina, um Richter oder Staatsanwalt zu werden (Böning & Schultz, 2019; Lippert, 2021). Sind diese Hürden einmal genommen, genießen Richter traditionell ein hohes Ansehen in der Gesellschaft (forsa, 2021). Wie auch Staatsanwälte sind sie im Gegensatz zu Strafverteidigern Diener des Staates und darüber hinaus Repräsentanten einer der Säulen der Gewaltenteilung. Im Gegenzug werden an Richter jedoch auch hohe Anforderungen seitens der Gesellschaft gestellt: Ungeachtet von Personalproblemen und umfangreichen, schwierigen Fällen wird von ihnen erwartet, rechtmäßige und möglichst auch gerechte Urteile zu fällen. Zudem können Fehlentscheidungen von Gerichten weitreichende Folgen unterschiedlicher Art und Schwere haben, die gerade im Strafrecht mit einem langjährigen Freiheitsentzug verbunden sein können. Resümiert man die Anforderungen an diesen Beruf, entsteht das Bild einer Personengruppe, die zwar hohes Ansehen genießt, aber auch große Verantwortung besitzt und möglicherweise einem erhöhten Druck ausgesetzt ist. Die Strafverteidigung, wenngleich ebenfalls von zentraler Bedeutung für Gesellschaft und Rechtsstaat, ist dagegen immer auch mit finanziellen Erwägungen und einer einseitigen Wahrnehmung der Interessen der

Mandanten verbunden, woraus sich bereits ein starker Kontrast zu den anderen Berufsgruppen ergibt. Dazu kommt, dass die Strafverteidigung „nolens volens“ in ihrem „Kampf um’s Recht“ (von Jhering, 1874) nicht selten gegen staatliche Instanzen opponieren muss.

Aus diesen unterschiedlichen Motiven und gesellschaftlichen Positionierungen lässt sich ableiten, dass womöglich auch die sozialen Normen zwischen den hier betrachteten Berufsgruppen in einem gewissen Maße divergieren. Etwaige Verstöße gegen die Strafprozessordnung dürften demzufolge von Richtern als deutlich stärkere Verletzung des eigenen Berufsethos empfunden werden, als dies bei Strafverteidigern der Fall ist. Dies hätte zur Folge, dass die Frage nach informellen Absprachen und damit einem gesetzwidrigen Verhalten von Richtern als weitaus brisanter wahrgenommen würde als durch Strafverteidiger, mit den entsprechenden Konsequenzen für das eigene Antwortverhalten. Damit decken sich auch Erkenntnisse darüber, dass sich die wahrgenommene Erwünschtheit verschiedener Eigenschaften und Verhaltensweisen je nach Personengruppe unterscheiden kann (Dalton & Ortegren, 2011; Furnham, 1986; Johnson & van de Vijver, 2003; Krumpal & Näher, 2012; Krumpal, 2013; Larson & Bradshaw, 2017; Opp, 2001; Phillips & Clancy, 1972; Stocké, 2004; Tourangeau & Yan, 2007). Zudem soll der Effekt der sozialen Erwünschtheit auch damit zusammenhängen, für wie unethisch eine bestimmte Situation oder Handlung eingestuft wird (Chung & Monroe, 2003). Wenn Richter aufgrund ihrer Berufskultur und gesellschaftlichen Stellung informelle Absprachen als besonders unethisch empfinden sollten, wäre auch dadurch eine größere Antwortverzerrung zu erwarten. Vor dem Hintergrund, dass Vorschriften über die Verständigung in Strafverfahren zentral an Richter adressiert sind, lässt sich der beobachtete Effekt der sozialen Erwünschtheit somit gut erklären.

3. Diskussion

In der vorliegenden Studie wurden Belege für das Auftreten des Effekts sozialer Erwünschtheit bei Richtern im Rahmen einer Befragung zur Verständigung im Strafprozess

(Kinzig, Iberl & Koch, 2020) geliefert. Resümierend lässt sich festhalten, dass Richter insbesondere im Vergleich zu Strafverteidigern deutlich stärker dazu neigen, Antworten in sozial erwünschter Richtung zu geben. Als Erklärungsansätze wurden die im Vergleich zu den anderen Berufsgruppen potentiell schwerwiegenderen negativen Konsequenzen für Richter bei Bekanntwerden informeller Absprachen angeführt. Zudem könnte die gesellschaftliche Rolle der Richter mit besonderen sozialen Normen verbunden sein, die das Antwortverhalten der Befragten ebenfalls maßgeblich beeinflussen.

Diesen Ergebnissen schließt sich unmittelbar die Frage an, welche Folgen der allem Anschein nach vorliegende Effekt der sozialen Erwünschtheit für die ermittelten Prävalenzen haben kann. Eine plausible und naheliegende Interpretation ist, dass die wahre Prävalenz informeller Absprachen höher liegt als sich in den Ergebnissen der Richter widerspiegelt. Eine genaue Bezifferung, um wie viel das tatsächliche Aufkommen von den Angaben dieser Berufsgruppe abweicht, ist jedoch nicht möglich.

Eine weitere Konsequenz aus diesem Umstand wäre, dass die Antworten der Strafverteidiger als deutlich valider anzusehen sind als jene der Richter (und Staatsanwälte). Da erstere keine Staatsbedienstete und als sogenannte Freiberufler tätig sind, dürften sie dem geringsten Normdruck ausgesetzt sein. Dies spricht dafür, dass die von dieser Berufsgruppe gegebenen Antworten der Realität am nächsten kommen. Folgt man dieser Hypothese und damit den Angaben der Strafverteidiger, würden sich immerhin 80 % aller justiziellen Akteure bisweilen (mindestens „selten“) an informellen Absprachen beteiligen. Beachtliche 38 % würden dies sogar häufig bis sehr häufig tun (Kinzig, Iberl & Koch, 2020). Jedoch können auch die Angaben von Strafverteidigern durch (andere) systematische Einflüsse verzerrt sein. So ist es nicht auszuschließen, dass sich an der Umfrage überproportional viele Rechtsanwälte beteiligt haben, die das Instrumentarium der illegalen Absprache in besonderer Weise zu nutzen suchen. Zudem erscheint die Frage diskutabel, ob die Berufsgruppe der Richter aufgrund der höheren Zugangsschranken zu diesem Beruf und der bei ihnen im Urteilsspruch liegenden Letztverantwortung die

Rechtmäßigkeit einer Verständigung besser einschätzen kann als das bei den Strafverteidigern der Fall ist.

Für die kriminologische Forschung stellt sich dessen ungeachtet die Frage, wie mit derartigen Antwortverzerrungen umzugehen ist – schließlich kann die Qualität der Daten durch ein systematisches „Underreporting“ merklich in Mitleidenschaft gezogen werden. Insbesondere in der Kriminologie beliebte Täter- und Opferbefragungen enthalten regelmäßig eine Fülle sensibler Fragestellungen, die sozial erwünschte Antworten zur Folge haben können. Wie die vorliegende Studie zeigt, können manche Fragen auch nur von bestimmten Bevölkerungs- oder Berufsgruppen als sensibel wahrgenommen werden, während bei anderen Populationen keine systematischen Antwortverzerrungen zu befürchten sind. Fragen nach der Begehung eigener Bagatellstraftaten, wie etwa kleineren Diebstahlsdelikten, Verkehrsstraftaten oder Drogenvergehen, könnte etwa von den meisten Personen als mäßig problematisch empfunden werden, während in der Strafverfolgung oder bei den Strafgerichten tätige Personen durchaus einen Anlass haben könnten, sich in ihren Antworten als besonders gesetzestreu zu präsentieren. Zur Bestimmung der Größe des Verzerrungsfaktors erforderlich ist also zunächst eine Einschätzung der Sensitivität der betreffenden Frage.

Wenn damit zu rechnen ist, dass eine spezielle Frage als brisant wahrgenommen wird, dürfte der klassische direkte Weg, gewisse Verhaltensweisen oder Merkmale zu eruieren, nicht unbedingt die beste Methode zur korrekten Ermittlung von Prävalenzen sein. Um dem Einfluss sozialer Erwünschtheit in solchen Situationen entgegenzuwirken, wurden daher indirekte Fragemethoden wie die Randomized Response Technique (Warner, 1965) entwickelt. Die Grundidee der Randomized Response Technique (RRT) ist es, die Antworten der Befragten mithilfe eines vorgeschalteten Zufallsexperiments zu verschleiern. Dadurch entsteht eine für die Befragten nachvollziehbare, objektive Anonymität. Diese gewährleistet, dass die Teilnehmenden antworten können, ohne direkte negative Konsequenzen durch die Verletzung ihrer Anonymität befürchten zu müssen. Neben der

RRT wurden mittlerweile mehrere verwandte Modelle entwickelt, die sich auf die Grundidee von Warner stützen (z. B. Clark & Desharnais, 1998; Greenberg et al., 1969; Miller, 1984; Moshagen, Musch & Erdfelder, 2011; Yu, Tian & Tang, 2008). In zahlreichen Befunden wurde bereits bestätigt, dass durch die Anwendung solcher Randomized Response Modelle (RRMs) bei sensiblen Fragestellungen validere Ergebnisse erzielt werden können als das bei direkt formulierten Fragen der Fall ist (Lensvelt-Mulders et al., 2005). Ein Nachteil dieser Methoden ist jedoch, dass die Prävalenzschätzung durch die Einbindung zufälliger Faktoren eine größere Varianz aufweist als bei direkten Fragen; es sind also größere Stichproben nötig, um statistisch belastbare Ergebnisse zu erzielen (Ulrich et al., 2012).

Die Wirkungsweise von RRMs kann anhand des Unrelated Question Model (UQM, s. Abb. 4); Greenberg et al., 1969), einem weit verbreiteten RRM, erläutert werden. Beim UQM wird, wie bei den meisten RRMs, durch die Befragten zunächst ein Zufallsexperiment durchgeführt, dessen Ergebnis nur ihnen selbst bekannt ist. Je nach Ergebnis wird ihnen dann eine von zwei Ja-/Nein-Fragen zugewiesen. Bei einer der Fragen handelt es sich um die sensible Frage nach der zu ermittelnden Prävalenz, wie z. B.: „Haben Sie sich schon einmal an einer informellen Absprache beteiligt?“. Die andere, neutrale Frage dient der Verschleierung der Antwort, wobei die zu erwartende Wahrscheinlichkeitsverteilung der Antworten im Voraus abschätzbar sein muss. Hier wird nicht selten das jeweilige Geburtsdatum erfragt, da dieses in der Bevölkerung in etwa gleich verteilt ist (Statistisches Bundesamt, 2022; Ulrich et al., 2012), z. B.: „Haben Sie in der ersten Jahreshälfte, also vor dem 1. Juli eines Jahres, Geburtstag?“ (ungefähre Wahrscheinlichkeit: 50 %). Beide Fragen werden zusammen nebst einem gemeinsamen Antwortfeld mit den Antwortoptionen „Ja“ und „Nein“ präsentiert. Die Befragten kreuzen also ihre Antwort an, während nur sie selbst wissen, auf welche Frage sich die Antwort bezieht. Über die beobachtbare relative Häufigkeit aller „Ja“-Antworten λ und die a priori bekannten Wahrscheinlichkeiten p, zur sensiblen Frage geleitet zu werden, und q, die neutrale Frage mit „Ja“ zu beantworten, lässt sich die gesuchte Prävalenz π_s dann schätzen über

$$\hat{\pi}_s = \frac{\hat{\lambda} - (1-p) \cdot q}{p}. \tag{1}$$

mit der Varianz

$$\sigma^2 = \frac{\lambda \cdot (1-\lambda)}{n \cdot p^2}. \tag{2}$$

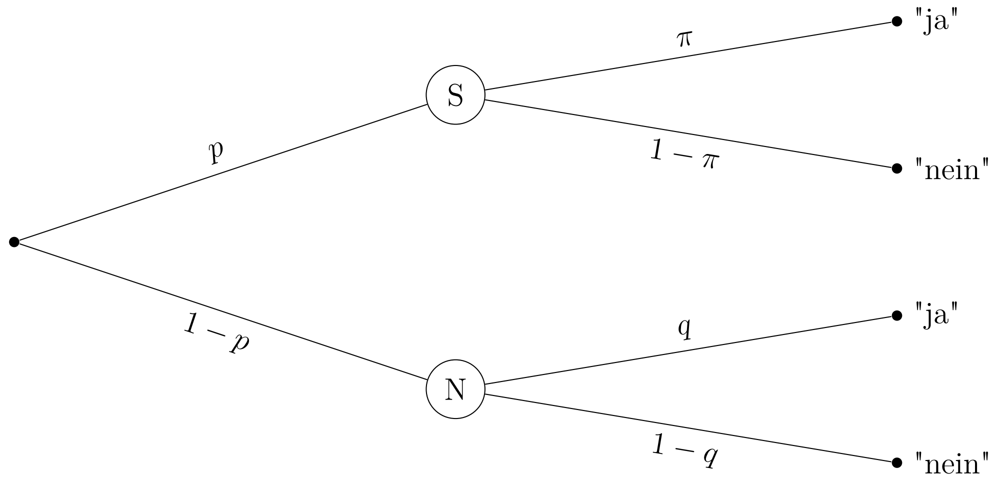
Da es sich bei der Beteiligung justizieller Akteure an informellen Absprachen um ein sensibles Thema handelt, würden sich RRMs grundsätzlich für eine entsprechende Befragung eignen. Wie bereits gesehen, spricht hier – zumindest bei Richtern – viel für eine deutliche Antwortverzerrung durch soziale Erwünschtheit, was die Eignung indirekter Befragungsmethoden zusätzlich unterstreicht. Um eine Vergleichbarkeit einer RRM-gestützten Richterbefragung mit der Erhebung von Kinzig, Iberl und Koch (2020) herzustellen, könnte man etwa nach der Lebenszeitprävalenz bei der Durchführung informeller Absprachen fragen. Diese entspricht den zusammengenommenen Häufigkeiten der Antwortoptionen „sehr häufig“, „häufig“ und „selten“ (s. Abb. 1)), die in der ursprünglichen Befragung präsentiert wurden.

Um die Frage zu beantworten, ob die Anwendung von RRMs zur Ermittlung der Lebenszeitprävalenz bei der Durchführung informeller Absprachen sinnvoll sein könnte, wurden einige Simulationen für das UQM durchgeführt.[2] Dabei wurde vor allem geprüft, wie groß eine Stichprobe aus Richtern sein müsste, um potentiell aussagekräftige Ergebnisse zu erzielen. Einbezogen wurde auch die Möglichkeit, dass einige Befragte immer mit „Nein“ antworten, unabhängig davon welche Frage ihnen eigentlich zugewiesen würde („selbstschützendes Antwortverhalten“). Die Ergebnisse der Simulationen sprechen dabei recht deutlich für eine Eignung des UQM auch im vorliegenden Fall. So würde man unter der Annahme, dass die wahre Prävalenz informeller Absprachen bei 58,7 % liegt (durchschnittliche Lebenszeitprävalenz über alle Berufsgruppen hinweg) und bei einer

[2] Ausführliche Informationen dazu und Ergebnisse der Simulationen sind dem Anhang zu entnehmen, der online unter www.nomos-elibrary.de/10.5771/2365-1083-2022-4-A1 verfügbar ist.

Abbildung 4

Wahrscheinlichkeitsbaum des Unrelated Question Model (UQM) nach Greenberg et al. (1969). "S" markiert den Zweig des Baums für Personen, denen die sensible Frage zugelost wird. "N" markiert den Zweig des Baums für Personen, denen die neutrale Frage zugelost wird.



Häufigkeit pauschaler „Nein"-Antworten von 20 % nur rund 200 Richter befragen müssen, um mit einer zufriedenstellenden Teststärke (β =.8) einen Unterschied zu der Prävalenz von 29,4 % festzustellen, die bei direkter Befragung der Richter ermittelt wurde (Kinzig, Iberl & Koch, 2020). In groß angelegten Forschungsprojekten wie dem von Altenhain, Jahn und Kinzig (2020) dürfte also auch bei zahlenmäßig vergleichsweise kleinen Gesamtpopulationen (z. B. Befragung justizieller Akteure) stets eine Stichprobengröße erreicht werden können, die den Einsatz von RRMs ermöglicht.

Falls für manche Befragungsgruppen eine Gefahr niedriger Compliance besteht, die vermehrte selbstschützende Antworten nach sich ziehen könnte, eignen sich diejenigen RRMs besonders gut, die ein derartiges Antwortverhalten berücksichtigen. Zu nennen ist hier z. B. die „Cheater-Extension" des UQM (UQMC; Reiber, Pope & Ulrich, 2020), eine Erweiterung des UQM, mit der neben der Prävalenz des untersuchten Merkmals auch die Häufigkeit selbstschützenden Verhaltens geschätzt werden kann. Gemäß der hier durchgeführten Simulationen wäre das UQMC unter gewissen Umständen ebenfalls für die Befragung von Richtern zu informellen Absprachen geeignet (s. Anhang).

Abschließend ist festzuhalten, dass die Anwendung von RRMs in der kriminologischen Forschung durchaus als vielversprechend einzustufen ist (s. auch De Puiseau, Hoffmann & Musch, 2015, und Treibel & Funke, 2004). So werden die Anforderungen an für RRMs erforderliche große Stichproben in zahlreichen quantitativen kriminologischen Studien längst erfüllt (z. B. Birkel et al., 2019; Dreißigacker & Riesner, 2018; Ellrich & Baier, 2015; Lutz et al., 2021; Treibel, Dölling & Hermann, 2017; Wegel, 2011), da oftmals ohnehin eine möglichst große Repräsentativität angestrebt wird. Die kriminologische Forschung beschäftigt sich darüber hinaus mit zahlreichen sensiblen Themen und heiklen Fragen, die das Aufkommen von Antwortverzerrungen durch soziale Erwünschtheit begünstigen. Dafür sind die hier im Mittelpunkt stehenden informellen Absprachen nur ein Beispiel. Bereits jetzt existieren zahlreiche Studien, bei denen RRMs im Rahmen kriminologisch relevanter Fragestellungen erfolgreich zum Einsatz gekommen sind (z. B. Dietz et al., 2018; Goodstadt & Gruson, 1975; Houston & Tran, 2001; Iberl, 2021; Musch, Bröder & Klauer, 2001; Reiber, Bryce & Ulrich, 2022; Soeken & Damrosch, 1986; Solomon et al., 2007; Ulrich et al., 2018; Wolter, 2012). Eine breitere Anwendung von RRMs in der kriminologischen Forschung wäre also für die Zukunft wünschenswert. Nur so können indirekte Fragemethoden weiter validiert und für die Anforderungen der Prävalenzforschung optimiert werden, um sich schließlich als gewinnbringende Ergänzung für die methodische Werkzeugkiste der Kriminologie zu etablieren.

Literatur

Altenhain, K., Dietmeier, F. & May, M. (2013). *Die Praxis der Absprachen in Strafverfahren.* Baden-Baden: Nomos.

Altenhain, K., Brandt, T. & Herbst, L. (2020). Leitfadengestützte Interviews mit Richtern, Staats- und Fachanwälten (Modul 5). In K. Altenhain, M. Jahn & J. Kinzig (Hrsg.), *Die Praxis der Verständigung im Strafprozess – Eine Evaluation der Vorschriften des Gesetzes zur Regelung der Verständigung im Strafverfahren vom 29. Juli 2009* (S. 307–480). Baden-Baden: Nomos.

Altenhain, K., Jahn, M. & Kinzig, J. (2020). *Die Praxis der Verständigung im Strafprozess – Eine Evaluation der Vorschriften des Gesetzes zur Regelung der Verständigung im Strafverfahren vom 29. Juli 2009.* Baden-Baden: Nomos.

Birkel, C., Church, D., Hummelsheim-Doss, D., Leitgöb-Guzy, N. & Oberwittler, D. (2019). *Der Deutsche Viktimisierungssurvey 2017: Opfererfahrungen, kriminalitätsbezogene Einstellungen sowie die Wahrnehmung von Unsicherheit und Kriminalität in Deutschland.* Verfügbar unter: https://www.bka.de/SharedDocs/Downloads/DE/Publikationen/Publikationsreihen/Forschungsergebnisse/2018ersteErgebnisseDVS2017.pdf?___blob=publicationFile&v=12 [Zugriff am 22.03.2022]. Institution: Kriminalistisches Institut des Bundeskriminalamtes.

Böning, A. & Schultz, U. (2019). Juristische Sozialisation. In C. Boulanger, J. Rosenstock & T. Singelnstein (Hrsg.), *Interdisziplinäre Rechtsforschung* (S. 193–205). Wiesbaden: Springer VS.

Chung, J. & Monroe, G. S. (2003). Exploring social desirability bias. *Journal of Business Ethics, 44(4)*, 291–302.

Clark, S. J. & Desharnais, R. A. (1998). Honest Answers to Embarrassing Questions: Detecting Cheating in the Randomized Response Model. *Psychological Methods, 3(2)*, 160–168.

Crowne, D. P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of consulting psychology, 24(4)*, 349–354.

Dalton, D. & Ortegren, M. (2011). Gender differences in ethics research: The importance of controlling for the social desirability response bias. *Journal of Business Ethics, 103(1)*, 73–93.

De Puiseau, B. W., Hoffmann, A. & Musch, J. (2015). Soziale Erwünschtheit in Viktimisierungsbefragungen. In N. Guzy, C. Birkel & R. Mischkowitz (Hrsg.), *Viktimisierungsbefragungen in Deutschland – Band 2: Methodik und Methodologie* (S. 187–216). Wiesbaden: Bundeskriminalamt.

Dietz, P., Iberl, B., Schuett, E., van Poppel, M., Ulrich, R. & Sattler, M. (2018). Prevalence Estimates for Pharmacological Neuroenhancement in Austrian University Students: Its Relation to Health-Related Risk Attitude and the Framing Effect of Caffeine Tablets. *Frontiers in pharmacology, 9*, 494.

Dreißigacker, A. & Riesner, L. (2018). *Private Internetnutzung und Erfahrung mit computerbezogener Kriminalität. Ergebnisse der Dunkelfeldstudien des Landeskriminalamtes Schleswig-Holstein 2015 und 2017.* Verfügbar unter: https://www.schleswig-holstein.de/DE/Landesregierung/POLIZEI/DasSindWir/LKA/KFS/_downloads/PIEcK.pdf?__blob=publicationFile&v=1 [Zugriff am 22.03.2022]. Institution: Kriminologisches Forschungsinstitut Niedersachsen e.V..

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of applied Psychology, 37(2)*, 90–93.

Ellrich, K. & Baier, D. (2015). Gewaltausübung durch Polizeibeamte – Ausmaß und Einflussfaktoren. *RPsych Rechtspsychologie, 1(1)*, 22–45.

forsa Politik- und Sozialforschung GmbH (2021). *dbb Bürgerbefragung „Öffentlicher Dienst“ 2021 – Einschätzungen, Erfahrungen und Erwartungen der Bürger.* Verfügbar unter: https://www.dbb.de/fileadmin/user_upload/globale_elemente/pdfs/2021/forsa_2021.pdf [Zugriff am 22.03.2022].

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and individual differences, 7(3)*, 385–400.

Goodstadt, M. S. & Gruson, V. (1975). The randomized response technique: A test on drug use. *Journal of the American Statistical Association, 70(352)*, 814–818.

Greenberg, B. G., Abul-Ela, A. L. A., Simmons, W. R. & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association, 64(326)*, 520–539.

Häcker, H. O. & Stapf, K.-H. (2009). *Dorsch – Psychologisches Wörterbuch* (15. Auflage). Bern: Huber.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30(2)*, 161–172.

Houston, J. & Tran, A. (2001). A survey of tax evasion using the randomized response technique. In T. M. Porcano (Hrsg.), *Advances in Taxation Volume 13* (S. 69–94). Bingley: Emerald Group Publishing Limited.

Iberl, B. (2021). Ein, zwei Bier und ab ans Lenkrad? – Prävalenzschätzung von Alkohol am Steuer durch das Unrelated Question Model. *Kriminologie – Das Online-Journal / Criminology – The Online Journal, 3(3)*, 270–292.

Johnson, T. P. & Van de Vijver, F. J. (2003). Social desirability in cross-cultural research. In J. Harkness, F. Van de Vijver & P. Mohler (Hrsg.), *Cross-cultural survey methods* (S. 195–204). Hoboken: Wiley.

Kinzig, J., Iberl, B. & Koch, J. (2020). Online-Befragung justizieller Akteure (Modul 4). In K. Altenhain, M. Jahn & J. Kinzig (Hrsg.), *Die Praxis der Verständigung im Strafprozess – Eine Evaluation der Vorschriften des Gesetzes zur Regelung der Verständigung im Strafverfahren vom 29. Juli 2009* (S. 191–305). Baden-Baden: Nomos.

Koerth, K. (2019, 3. Mai). *Personalnot – Die Justiz sieht alt aus.* Verfügbar unter: https://www.spiegel.de/karriere/arbeitsueberlastung-im-gericht-warum-die-justiz-alt-aussieht-a-1265194.html [Zugriff am 22.02.2022]. Hamburg: Der Spiegel.

Krumpal, I. & Näher, A. F. (2012). Entstehungsbedingungen sozial erwünschten

Antwortverhaltens: Eine experimentelle Onlinestudie zum Einfluss des Wordings und des Kontexts bei unangenehmen Fragen. *Soziale Welt, 63(1)*, 65–89.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity, 47(4)*, 2025–2047.

Larson, K. E. & Bradshaw, C. P. (2017). Cultural competence and social desirability among practitioners: A systematic review of the literature. *Children and Youth Services Review, 76*, 100–111.

Lee, R. M. (1993). *Doing research on sensitive topics.* London: Sage.

Lensvelt-Mulders, G. J., Hox, J. J., Van der Heijden, P. G. & Maas, C. J. (2005). Metaanalysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research, 33(3)*, 319–348.

Lippert, P. (2021). *Wege zum Prädikatsexamen: Wie jeder seine Chancen auf Top-Jura-Examina durch strukturierte Vorbereitung verbessern kann.* Stuttgart: UTB.

Lutz, P., Stelly, W., Bartsch, T., Thomas, J. & Bergmann, B. (2021). Islamische Seelsorge im Jugendstrafvollzug. *Kriminologie - Das Online-Journal | Criminology-The Online Journal, 3(3)*, 228–248.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior (Doctoral dissertation).* Washington, D.C.: The George Washington University.

Moshagen, M., Musch, J. & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods, 44(1)*, 222–231.

Musch, J., Bröder, A. & Klauer, K. C. (2001). Improving survey research on the World-Wide Web using the randomized response technique. In U.-D. Reips & M. Bosnjak (Hrsg.), *Dimensions of Internet science* (S. 179–192). Lengerich: Pabst Science Publishers.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European journal of social psychology, 15(3)*, 263–280.

Opp, K.-D. (2001). Social networks and the emergence of protest norms. In M. Hechter & K.-D. Opp (Hrsg.), *Social Norms* (S. 234–273). New York: Russell Sage Foundation.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of personality and social psychology, 46(3)*, 598–609.

Phillips, D. L. & Clancy, K. J. (1972). Some effects of "social desirability" in survey studies. *American journal of sociology, 77(5)*, 921–940.

Rasinski, K. A., Willis, G. B., Baldwin, A. K., Yeh, W. & Lee, L. (1999). Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 13(5)*, 465–484.

Reiber, F., Bryce, D. & Ulrich, R. (2022). Self-protecting responses in randomized response designs: A survey on intimate partner violence during the coronavirus disease 2019 pandemic. *Sociological Methods & Research, 1–32.*

Reiber, F., Pope, H. & Ulrich, R. (2020). Cheater detection using the unrelated question model. *Sociological Methods & Research, 1–23.*

Reiber, F., Schnuerch, M. & Ulrich, R. (2020). Improving the efficiency of surveys with randomized response models: A sequential approach based on curtailed sampling. *Psychological Methods, 1–14.*

Soeken, K. L. & Damrosch, S. P. (1986). Randomized response technique: Applications to research on rape. *Psychology of Women Quarterly, 10(2)*, 119–126.

Solomon, J., Jacobson, S. K., Wald, K. D. & Gavin, M. (2007). Estimating illegal resource use at a Ugandan park with the randomized response technique. *Human Dimensions of Wildlife, 12(2)*, 75–88.

Statistisches Bundesamt (Destatis) (2022). *Lebendgeborene: Deutschland, Monate, Geschlecht.* Verfügbar unter: https://www-genesis.destatis.de/genesis//online?operation=table&code=12612-0002&bypass=true&levelindex=0&levelid=1645004567529#abreadcrumb [Zugriff am 16.02.2022].

Stocké, V. (2004). Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit: Ein Vergleich der Prognosen der rational-choice Theorie und des Modells der frame-Selektion. *Zeitschrift für Soziologie, 33(4)*, 303–320.

Stocké, V. & Hunkler, C. (2007). Measures of desirability beliefs and their validity as indicators for socially desirable responding. *Field methods, 19(3)*, 313–336.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin, 133(5)*, 859–883.

Treibel, A., Dölling, D. & Hermann, D. (2017). Determinanten des Anzeigeverhaltens nach Straftaten gegen die sexuelle Selbstbestimmung. *Forensische Psychiatrie, Psychologie, Kriminologie, 11(4)*, 355–363.

Treibel, A. & Funke, J. (2004). Die internetbasierte Opferbefragung als Instrument der Dunkelfeldforschung – Grenzen und Chancen. *Monatsschrift für Kriminologie und Strafrechtsreform, 87(2)*, 146–151.

Ulrich, R., Pope, H. G., Cléret, L., Petróczi, A., Nepusz, T., Schaffer, J., Kanayama, G., Comstock, R. D. & Simon, P. (2018). Doping in two elite athletics competitions assessed by randomized-response surveys. *Sports medicine, 48(1)*, 211–219.

Ulrich, R., Schröter, H., Striegel, H. & Simon, P. (2012). Asking sensitive questions: a statistical power analysis of randomized response models. *Psychological methods, 17(4)*, 623.

von Jhering, R. (1874). *Der Kampf um‘s Recht* (4. Auflage). Wien: Verlag der G. J. Manz’schen Buchhandlung.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60(309)*, 63–69.

Wegel, M. (2011). Gewaltverhalten, Sozialisation und Wertorientierungen in unterschiedlichen Schülermilieus – Ergebnisse der Tübinger Schülerbefragung. In B. Bannenberg & J.-M. Jehle (Hrsg.), *Gewaltdelinquenz, Lange Freiheitsentziehung, Delinquenzverläufe* (S. 85–96). Mönchengladbach: Forum Verlag Godesberg.

Wenninger, G. (2001). *Lexikon der Psychologie in fünf Bänden – Vierter Band (Reg bis Why).* Heidelberg: Spektrum Akademischer Verlag.

Wolter, F. (2012). *Heikle Fragen in Interviews – Eine Validierung der Randomized Response-Technik*. Wiesbaden: Springer VS.

Yu, J. W., Tian, G. L. & Tang, M. L. (2008). Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika, 67(3)*, 251–263.

Appendix

Online-Anhang

Es wurden Powerkurven für das UQM (Greenberg et al., 1969) und das UQMC (Reiber, Pope & Ulrich, 2020) mit unterschiedlichen Parametern berechnet. Dabei wurde entweder getestet, ob die angegebene Prävalenz von 29,4 % (entspricht der angegebenen Lebenszeitprävalenz der Richter in ihrer eigenen Praxis) in einem einseitigen Konfidenzintervall enthalten ist oder nicht (also, ob ein Unterschied zur gemessenen Prävalenz in Kinzig, Iberl und Koch (2020) gefunden werden kann), oder, ob der geschätzte Cheating-Parameter γ sich von 0 unterscheidet.

Der auf der x-Achse variierte Parameter ist entweder die wahre Prävalenz π_s oder die Stichprobengröße N. Es wurde außerdem festgelegt, dass $q = .5$, $p = .67$ und im UQMC $p_1 = .67$ und $p_2 = .33$. Bei π_s als frei variierendem Parameter wurde N abgestuft in 100, 300 und 558 (realistisch zu erhebende Stichprobengrößen mit der tatsächlich erhobenen Stichprobengröße für Richter in Kinzig, Iberl und Koch, 2020). Beim UQMC wurde diese Stichprobe zweigeteilt, um die beiden Stichproben N_1 und N_2 zu bilden.

Bei N als frei variierendem Parameter wurde π_s abgestuft in die möglicherweise dem wahren Wert nahekommenden Werte $\pi_s r = .444$ (Lebenszeitprävalenzangabe der Richter zu Absprachen nach Hörensagen), $\pi_s g = .587$ (Lebenszeitprävalenz der Gesamtstichprobe nach Hörensagen) und $\pi_s v = .804$ (Lebenszeitprävalenz der Strafverteidiger in der eigenen Praxis). Außerdem wurde variiert, wie viele Befragte pauschal mit „Nein“ antworten würden, unabhängig davon, welche Frage ihnen gestellt wird. Dieser Anteil entspricht gleichzeitig dem wahren Cheating-Parameter γ. Er beträgt $\gamma_1 = 0$, $\gamma_2 = .1$, $\gamma_3 = .2$, $\ldots$, $\gamma_6 = .5$.

Die simulierten Häufigkeiten beruhen auf der zufälligen Generierung von Ja- oder Nein-Antworten ausgehend von den o. g. variierenden Parametern. Die binomialverteilten Antworten wurden, um eine effizientere Berechnung zu gewährleisten, durch die Normalverteilung approximiert.

Abbildung A1

UQM, 0-20% pausch. Nein-Antworten, Test, ob sich π_s von π_s0 unterscheidet (abh.: π_s).

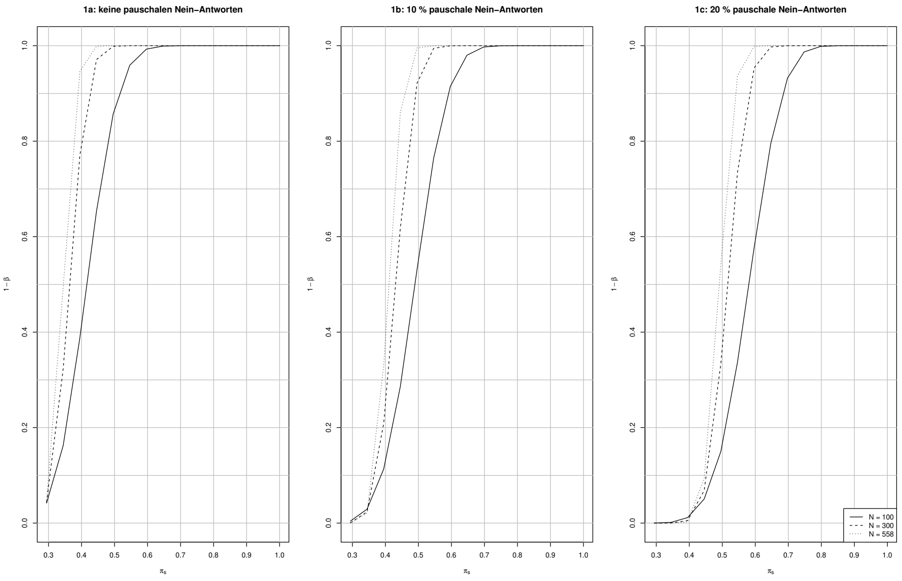


Abbildung A2

UQM, 30-50% pausch. Nein-Antworten, Test, ob sich π_s von π_s0 unterscheidet (abh.: π_s).

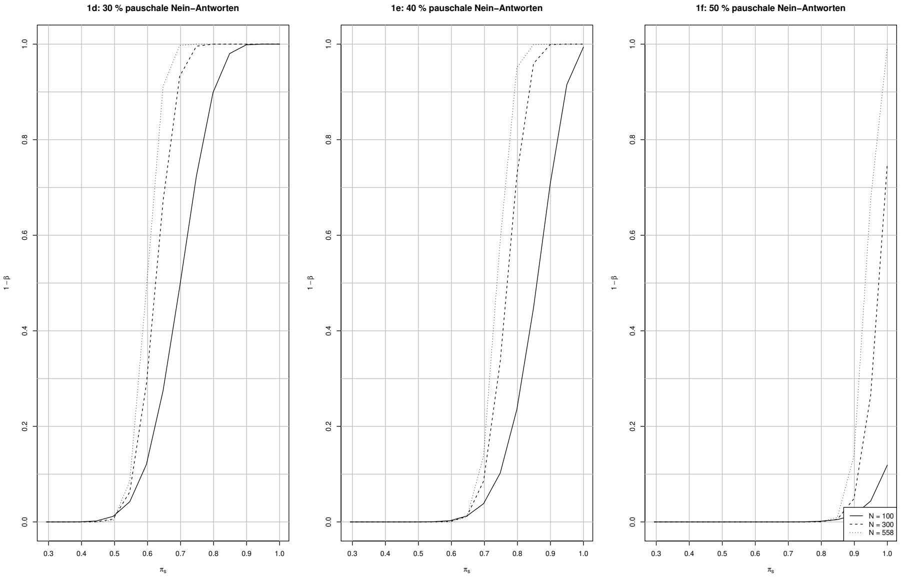


Abbildung A3

UQM, 0-20% pausch. Nein-Antworten, Test, ob sich π_s von $\pi_s 0$ unterscheidet (abh.: N).

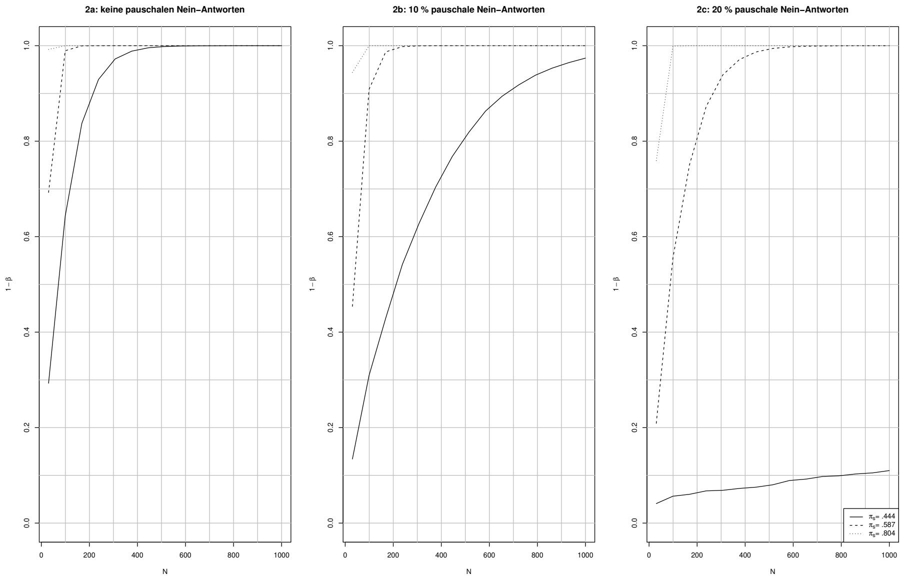


Abbildung A4

UQM, 30-50% pausch. Nein-Antworten, Test, ob sich π_s von $\pi_s 0$ unterscheidet (abh.: N).

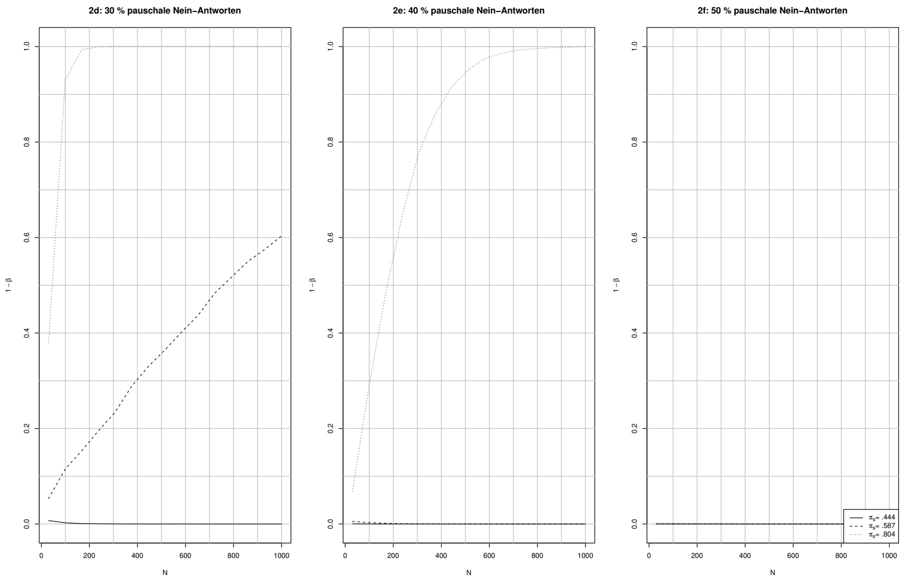


Abbildung A5

UQMC, 0-20% pausch. Nein-Antworten, Test, ob sich π_s von $\pi_s 0$ unterscheidet (abh.: π_s).

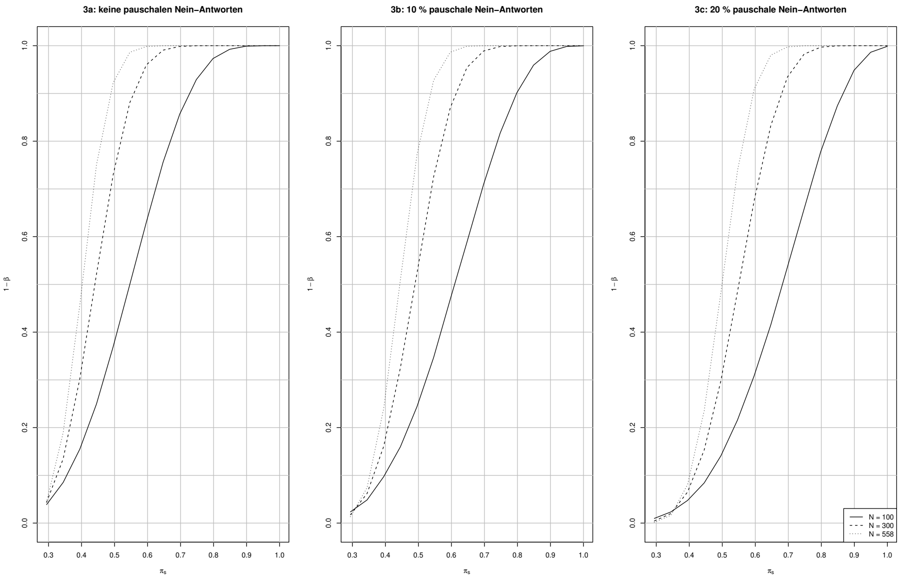


Abbildung A6

UQMC, 30-50% pausch. Nein-Antworten, Test, ob sich π_s von $\pi_s 0$ unterscheidet (abh.: π_s).

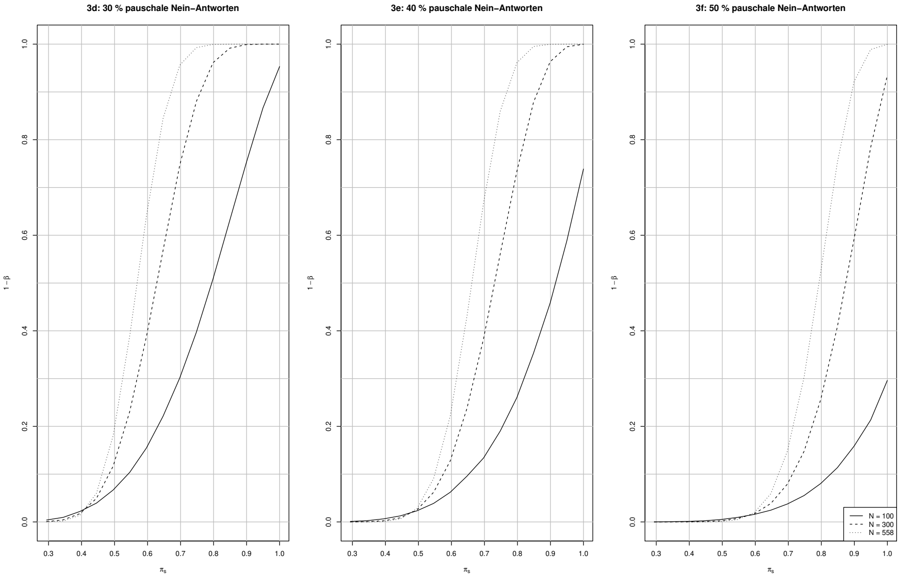


Abbildung A7

UQMC, 0-20% pausch. Nein-Antworten, Test, ob sich π_s von $\pi_s 0$ unterscheidet (abh.: N).

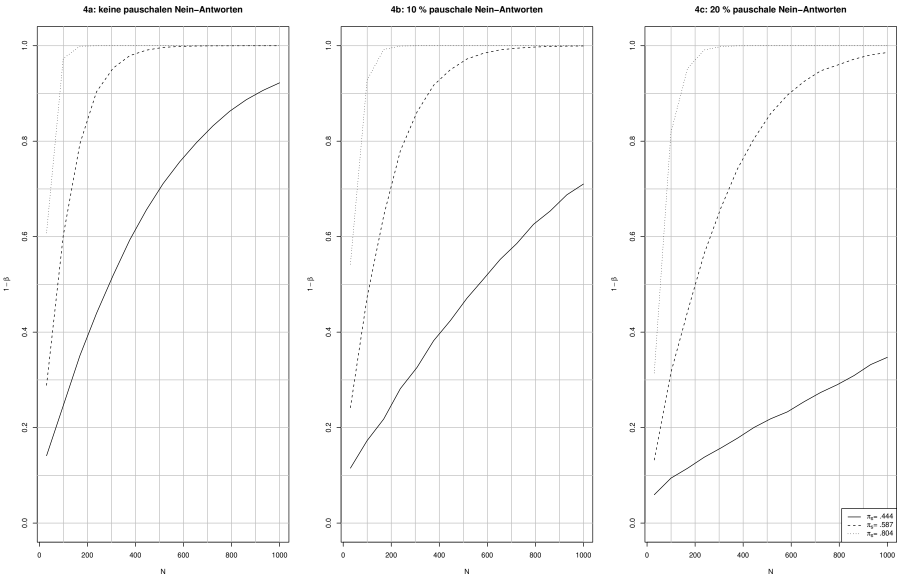


Abbildung A8

UQMC, 30-50% pausch. Nein-Antworten, Test, ob sich π_s von $\pi_s 0$ unterscheidet (abh.: N).

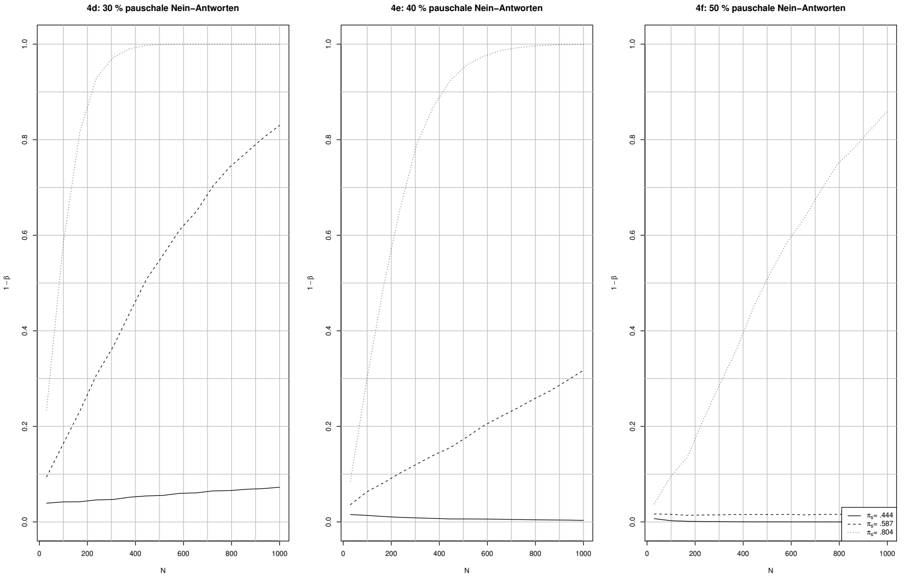


Abbildung A9

UQMC, 0-20% pausch. Nein-Antworten, Test, ob sich γ von 0 unterscheidet (abh.: π_s).

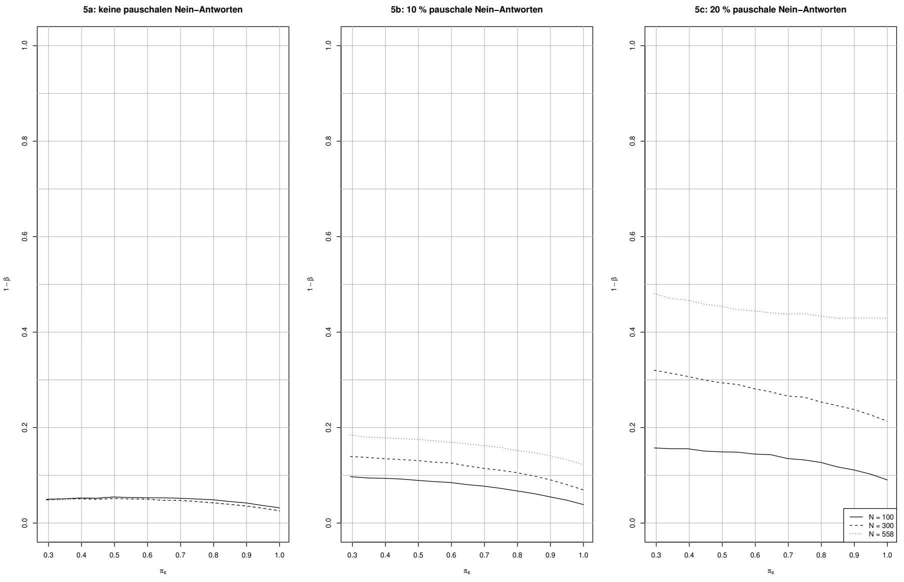


Abbildung A10

UQMC, 30-50% pausch. Nein-Antworten, Test, ob sich γ von 0 unterscheidet (abh.: π_s).

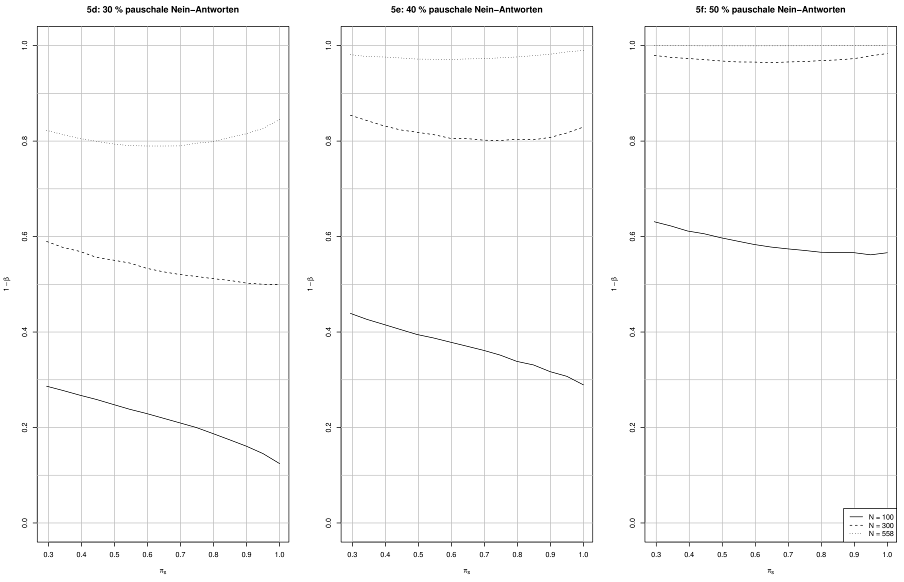
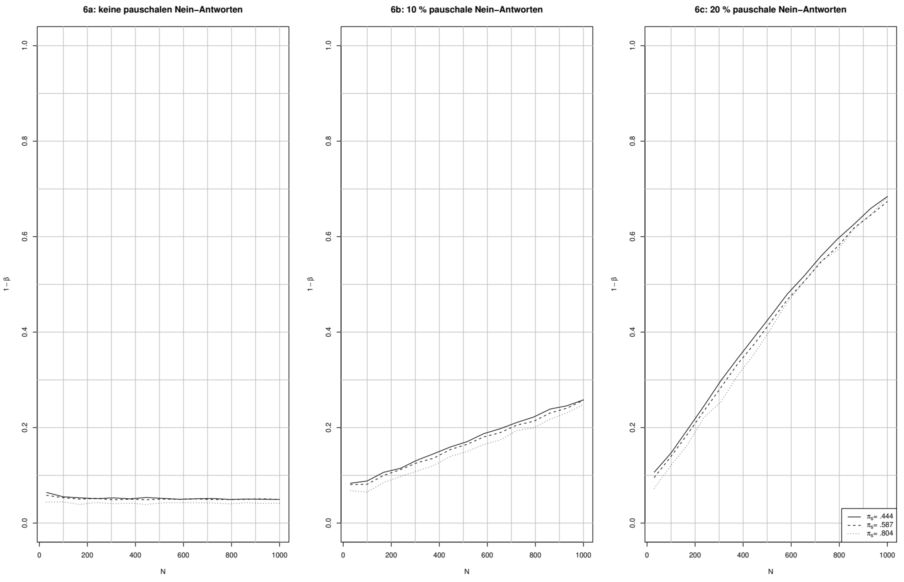


Abbildung A11

UQMC, 0-20% pausch. Nein-Antworten, Test, ob sich γ von 0 unterscheidet (abh.: N).



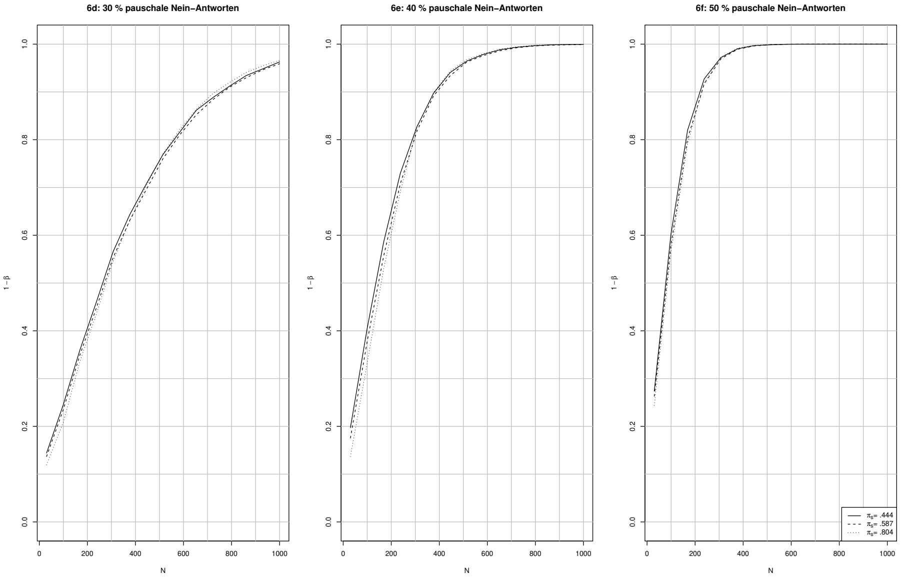
6a: keine pauschalen Nein−Antworten

6b: 10 % pauschale Nein−Antworten

6c: 20 % pauschale Nein−Antworten

Abbildung A12

UQMC, 30-50% pausch. Nein-Antworten, Test, ob sich γ von 0 unterscheidet (abh.: N).



6d: 30 % pauschale Nein−Antworten

6e: 40 % pauschale Nein−Antworten

6f: 50 % pauschale Nein−Antworten

B Paper 2

Iberl, B., & Ulrich, R. (2023). On estimating the frequency of a target behavior from time-constrained yes/no survey questions: A parametric approach based on the Poisson process. *Psychological Methods*. Advance online publication. `https://doi.org/10.1037/met0000588`.

Candidate contributions to the article

Status	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Published	50%	80%	80%	80%

Psychological Methods

On Estimating the Frequency of a Target Behavior From Time-Constrained Yes/No Survey Questions: A Parametric Approach Based on the Poisson Process

Benedikt Iberl and Rolf Ulrich

Online First Publication, July 20, 2023. https://dx.doi.org/10.1037/met0000588

CITATION

Iberl, B., & Ulrich, R. (2023, July 20). On Estimating the Frequency of a Target Behavior From Time-Constrained Yes/No Survey Questions: A Parametric Approach Based on the Poisson Process. *Psychological Methods*. Advance online publication. https://dx.doi.org/10.1037/met0000588

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

© 2023 American Psychological Association
ISSN: 1082-989X

https://doi.org/10.1037/met0000588

On Estimating the Frequency of a Target Behavior From Time-Constrained Yes/No Survey Questions: A Parametric Approach Based on the Poisson Process

Benedikt Iberl[1] and Rolf Ulrich[2]
[1] Faculty of Law, Institute of Criminology, University of Tübingen
[2] Faculty of Science, Department of Psychology, University of Tübingen

Abstract

We propose a novel method to analyze time-constrained yes/no questions about a target behavior (e.g., “Did you take sleeping pills during the last 12 months?”). A drawback of these questions is that the relative frequency of answering these questions with “yes” does not allow one to draw definite conclusions about the frequency of the target behavior (i.e., how often sleeping pills were taken) nor about the prevalence of trait carriers (i.e., percentage of people that take sleeping pills). Here we show how this information can be extracted from the results of such questions employing a prevalence curve and a Poisson model. The applicability of the method was evaluated with a survey on everyday behavior, which revealed plausible results and reasonable model fit.

Translational Abstract

Surveys often address closed questions (e.g., “Did you take sleeping pills in the last 12 months?”), which respondents should answer with “yes” or “no.” Although answering such time-constrained questions is straightforward due to the dichotomous response format, the prevalence obtained does not indicate how frequently the target behavior occurs. We propose a novel method of how to estimate this frequency based on such dichotomous questions. Moreover, this method also allows the determination of the prevalence of trait carriers (e.g., people that take sleeping pills).

Keywords: survey, time-constrained question, prevalence curve, Poisson process

A major goal of surveys on prevalence estimation is to gather knowledge about certain target behaviors. Such surveys often rely on yes/no questions asking participants whether they showed the target behavior within a specific time frame—therefore, we call these questions *time-constrained yes/no questions*. For example, in surveys about the prevalence of drug usage, people are questioned about drug usage within a 12-month time frame (e.g., Beck et al., 2021): “Over the last 12 months, have you taken sleeping pills or drugs for sleep (yes; no; don’t know)?” Time-constrained yes/no questions generate simple, dichotomous count data, from which the relative frequency of “yes” and “no” answers is computed. In our example, Beck et al. (2021) found that 16% of the participants surveyed in 2020 used sleeping pills in the past year. A major problem of time-constrained questions is the ambiguity of the resulting prevalence estimates: Do the results of Beck et al. (2021) imply that there are around 16% regular sleeping pill consumers? Such a conclusion is undoubtedly problematic because even a sleeping pill user might not have consumed the drug in the past year.

Particularly vague is the lifetime prevalence (e.g., Mohebbi et al., 2019) for most behaviors. The problematic interpretations of lifetime-prevalence measures were discussed by Fiedler and Schwarz (2016) in a study on questionable research practices. They argue that lifetime-prevalence estimates do not measure prevalence but are a distinct construct. The authors exemplified this problem in the prevalence of church attendance. Specifically, the percentage of people who visited a church at least once in their lives must be differentiated from the proportion of regular church attendees in the general population. Transferring this logic to our example, a person might have used sleeping pills for a limited time and a long while ago but never since. Including such persons would clearly inflate prevalence estimates of regular sleeping pill consumers. Hence, prevalence estimates relying on time-constrained questions can be ambiguous concerning the true frequency of behaviors—even if the time constraint used is the respondents’ lifetime.

Benedikt Iberl https://orcid.org/0000-0002-4463-7009
Rolf Ulrich https://orcid.org/0000-0001-8443-2705

We thank Jeff Miller for helpful comments on the first version of this manuscript and for helpful suggestions.

This work was not presented elsewhere before this publication. This research received no external financial or non-financial support. There are no conflicts of interest for either author.

The article was preregistered via OSF (see https://osf.io/z35y7). The complete data and analysis code was uploaded at https://osf.io/7ptsd/.

Benedikt Iberl served as the lead for data generation. Benedikt Iberl and Rolf Ulrich contributed equally to scientific ideas, analysis and interpretation, and writing.

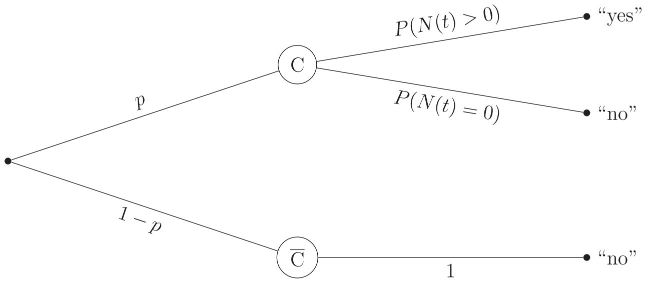
Correspondence concerning this article should be addressed to Benedikt Iberl, Faculty of Law, Institute of Criminology, Eberhard Karls Universität Tübingen, Sand 7, 72076 Tübingen, Germany. Email: [redacted]

There are several cases in the literature where prevalence estimations have been conducted through surveys with time-constrained yes/no questions (e.g., cannabis usage, Atzendorf et al., 2019; cognitive neuroenhancement, Dietz et al., 2018; physical exercise, Lin et al., 2018; intimate partner violence, Reiber et al., 2022; sexual behavior and usage of "dating apps," Sawyer et al., 2018; book reading, Şaşmaz et al., 2014; physical doping, Ulrich et al., 2018; cooking dinner at home, Virudachalam et al., 2014). As noted before, this time constraint hampers the interpretation of results because the relative frequency of "yes" answers is ambiguous with respect to the occurrence of the target behavior of interest. Therefore, some authors have tried to circumvent the ambiguity of these questions by posing multiple-choice questions that include several time frames at once (e.g., Miller et al., 2020; Molinaro et al., 2018; Seitz et al., 2020). For example, Seitz et al. (2020) asked participants if they used sleeping pills in the last 30 days at all, more rarely than once per week, once per week, multiple times per week, or daily. However, such questions still carry ambiguous information because of the specific time frames employed. Other authors use semiopen questions to derive more information, such as asking participants on how many days they showed a particular behavior in the last month (e.g., Cullen et al., 2018; Soga et al., 2021). Such tasks, especially semiopen questions, can be very demanding because participants have to recall more than one occurrence of the behavior in question, which is an undoubtedly effortful memory task.

In this article, we suggest a novel approach to disambiguate the results of time-constrained questions and enhance the interpretations of such results. This approach relies on a model based on a Poisson process. The probability tree in Figure 1 illustrates the proposed *Poisson model.* Assume that we want to investigate the prevalence of sleeping pill consumers in a student sample. The question posed to the students could be simple, like whether they used sleeping pills within a specific time frame. As shown in this figure, the probability tree splits into two main branches, representing *carriers* (i.e., participants who in principle engage in the researched behavior) and *noncarriers* (i.e., participants who will never engage in the studied behavior), respectively. The parameter p thus describes the probability of a randomly selected person in the target population being a sleeping pill user, while $1 - p$ is the probability of them being a nonuser. Consequently, nonusers are assumed to answer "no" regardless of how long the time frame would be.

Figure 1

Illustration of the Proposed Poisson Model



Note. The sample is divided into carriers C and noncarriers $\overline{\text{C}}$ by the parameter p, describing the probability of a random participant being a carrier of the researched attribute. Noncarriers answer "no" with a probability of 1. Carriers answer "yes" with a probability of $P(N(t) > 0)$ or "no" with a probability of $P(N(t) = 0)$.

As can be seen in the upper branch of Figure 1, carriers split into two subgroups: Those carriers that have shown the behavior in question at least once within the specific time frame t, that is, $N(t) > 0$, and carriers that have not shown it, $N(t) = 0$. Hence, the first subgroup would answer "yes," while the second would answer "no," even though the people in the latter group are carriers as well. Therefore, $N(t)$ represents a random variable, which we assume follows a Poisson distribution with intensity parameter λ,

$$P(N(t) = k) = \frac{(\lambda \cdot t)^k \cdot \mathrm{e}^{-\lambda \cdot t}}{k!}, \tag{1}$$

and denotes the probability that the behavior in question occurred k times within the time frame t. Specifically, for $k = 0$, the equation becomes

$$P(N(t) = 0) = \mathrm{e}^{-\lambda \cdot t}. \tag{2}$$

According to our model, the probability of a "yes" response for a randomly drawn participant is then,

$$P(\text{“yes”}|t) = p \cdot P(N(t) > 0), \tag{3}$$

or equivalently,

$$P(\text{“yes”}|t) = p \cdot [1 - P(N(t) = 0)]. \tag{4}$$

Since $N(t)$ is assumed to follow a Poisson process, the preceding equation can be rewritten as

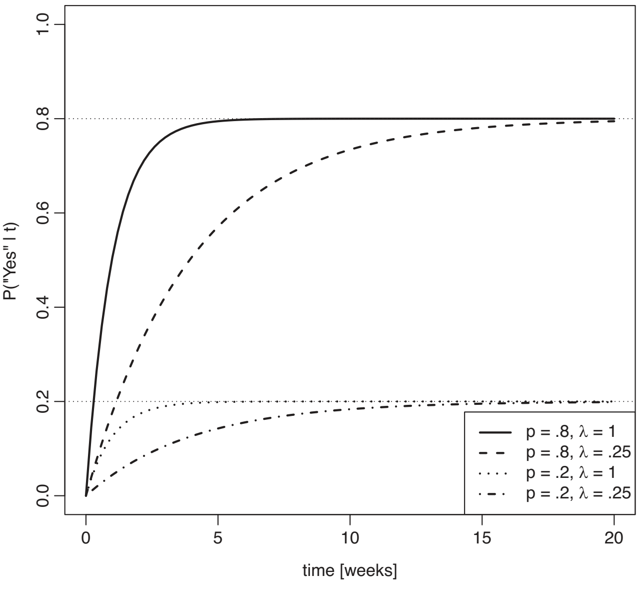
$$P(\text{“yes”}|t) = p \cdot \left(1 - \mathrm{e}^{-\lambda \cdot t}\right), \tag{5}$$

which shows how the predicted prevalence would evolve over time t (hence we call this function the *prevalence curve*).

Figure 2 exemplifies predicted prevalence curves. For this plot, we set the parameter p as .8 or .2 and λ as 1 or 0.25, showing the predicted prevalence curves of all four possible parameter

Figure 2

Examples of Prevalence Curves as a Function of p and λ



combinations. A high value for p, like .8, represents a behavior shown regularly by most of the population. Yet a low value, like .2, represents behavior that only a small part of the surveyed population would potentially engage in. In this case, λ is to be interpreted in proportion to the applied unit of time, that is, weeks. So the higher value of 1 indicates that the behavior occurs once per week on average, while the lower value of 0.25 represents an average occurrence of 0.25 times per week, so once a month. As shown in Figure 2, the resulting prevalence curves rise more steeply with higher values for λ before reaching an asymptote for the maximum prevalence determined by p. This illustrates the model's benefits well: It describes how to estimate the prevalence of a certain behavior (i.e., using two or more different time intervals) without being tied to a specific time interval, thus solving the above problem of ambiguity.[1]

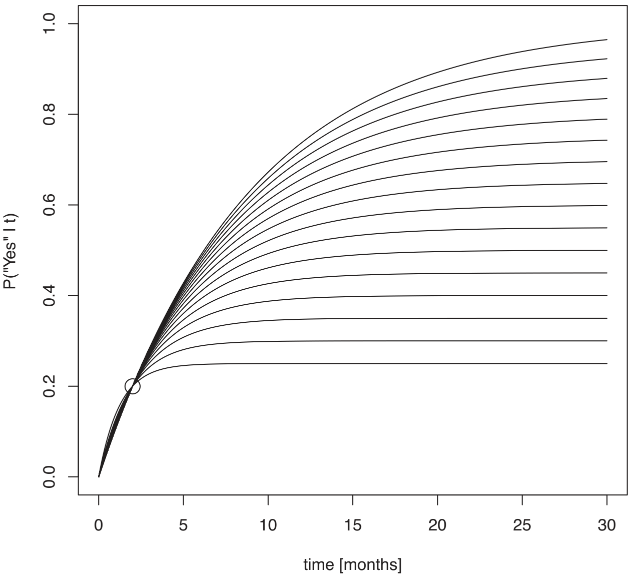
To put the model into practice, we can still ask time-constrained yes/no questions as in every other prevalence survey, but vary the time frame t referred to in the question across several groups of participants. With the observed numbers of "yes" and "no" answers per group, it then becomes possible to estimate the parameters λ and p (see "Appendix A").

A particularly useful property of the Poisson model is that it enhances the interpretation of the results from time-constrained questions. Suppose one is interested in the prevalence of sleeping pill usage in a survey. For example, one might ask here, "have you consumed sleeping pills in the last 2 months?" Let us assume that 20% of all respondents answer "yes" to this question. In this case, there would be an infinite number of ways within the model to interpret this particular result, and Figure 3 presents a subset of these possibilities. Thus, the present approach and the suggested model greatly enhance interpretation since it allows the determination of the empirically appropriate prevalence curve from among these infinite possibilities.

The applicability of the proposed Poisson model for prevalence estimation was tested in a pilot study. For this purpose, we interviewed students at Tübingen University via an online survey on everyday behaviors.

Figure 3

Possible Prevalence Curves for Sleeping Pill Use That Are Consistent With P("yes"|t = 2) = 0.2



Method

Participants

In total, 28,872 students of Tübingen University were invited to the survey by email. In total, 872 students filled out the survey, 858 of whom answered all questions. According to our preregistered data exclusion criteria (see https://osf.io/z35y7), 33 participants were excluded due to failing an attention check. The remaining sample of 839 students consisted of 631 female, 192 male, and 13 nonbinary participants with an average age of 23.5 years ($SD = 5.0$). Most students were enrolled either in the maths and natural science faculty or in the faculty of philosophical sciences (31.6% and 24.4%, respectively). In total, 17.8% of the sample were medical students, 15.8% studied at the faculty of economic and social sciences. The remaining participants studied law (5.8%), theology (4.1%), or "miscellaneous" (0.5%).

Materials

The survey consisted of six prevalence questions, an attention check, and three demographic questions. In the prevalence questions, participants were asked if they engaged in a certain day-to-day behavior within a certain time frame (past week, past month, or past 3 months, depending on the group). They were asked…

- …if they watched a weekly *sports program* (the *Sportschau*) within the past week/month/3 months.
- …if they watched a weekly *crime thriller* (the *Tatort*) within the past week/month/3 months.
- …if they ate *pizza* within the past week/month/3 months.
- …if they drank *coffee* within the past week/month/3 months.
- …if they *congratulated a relative on their birthday* within the past week/month/3 months.
- …if they *participated in another survey* within the past week/month/3 months.

Each question could be answered by clicking a "yes" or "no" button.

In the attention check, they were presented with seven topics. Their task was to select the topics they were asked about in the six previous prevalence questions to identify the decoy option. This item was used as a data exclusion criterion. The topics listed were "other surveys," "the *Sportschau*," "coffee," "crayons" (decoy), "birthday congratulations," "the *Tatort*," and "pizza." In the demographic questions, the participants were asked about their age, sex, and in which faculty they were studying.

Design

The six prevalence questions varied between groups regarding the time frame (1 week, 1 month, and 3 months). Each participant was

[1] The parameter λ might vary across participants. In the discussion below, we show how this can be addressed quantitatively by an elaborated version of the Poisson model.

Table 1

Observed Frequencies of "Yes" and "No" Answers for Each Question

	Time frame		
Question	Last week	Last month	Last 3 months
Sports program			
"Yes"	27	48	65
"No"	256	236	207
Crime thriller			
"Yes"	22	59	69
"No"	261	225	203
Pizza			
"Yes"	131	250	257
"No"	152	34	15
Coffee			
"Yes"	171	190	188
"No"	112	94	84
Birthday congrat.			
"Yes"	77	191	239
"No"	206	93	33
Other surveys			
"Yes"	53	140	194
"No"	230	144	78

randomly assigned to one of the three groups. The questions were presented in randomized order on an individual level to avoid any potential sequence effects.

Procedure

The email invitations were sent to the students on March 23, 2022, starting at 3:20 p.m. In the invitation, the students were briefly told about the goals and contents of the survey. Requirements for participation (at least 18 years old, enrolled as students at Tübingen University, and fluent in German) were mentioned, as well as the basic legal framework (voluntary participation, complete anonymity, and confidential use of all gathered data), and the predicted duration of the survey (2–3 min). As an incentive for participation, the possibility of winning one of five 10 bookstore vouchers was also pointed out to the potential participants. The link to the online survey was contained in the email invitation.

On the first page of the survey, the students were presented with a welcome text, which included information similar to that in the invitation. After continuing, the six randomly ordered prevalence questions were presented to the participants, each question on an individual survey page. For a given participant, the time frame used was the same for all questions. On page 8, the attention check was presented, followed by the demographic questions on the subsequent page. On the last page, the students could state their email addresses if they wanted to participate in the voucher lottery. Throughout the whole survey, no answers were enforced. When participants did not answer a certain item, they were shown a warning and asked if they wanted to skip the question. The survey was online for 2 weeks and taken down on April 6, 2022. Then the lottery winners were drawn randomly on May 10, 2022, and received their vouchers on May 18, 2022.

All computations were executed using the free software *R* (R Core Team, 2018). See https://osf.io/7ptsd/ for the complete data and analysis code.

Results

The observed frequencies of "yes" and "no" answers to all questions are presented in Table 1. The sample sizes are $n_1 = 283$ for the last week group, $n_2 = 284$ for the last month group, and $n_3 = 272$ for the last 3 months group.

The maximum likelihood method was used to estimate the model parameters (see Appendix A). In addition, a parametric bootstrapping procedure (e.g., Boos, 2003) with 1,000 bootstrap samples was employed to estimate the standard error of the estimates along with 95% confidence intervals. The average bootstrapped estimates for p and λ are presented in Table 2. As described above, low values for p, as calculated for the sports program and crime thriller questions, indicate a large portion of noncarriers. The highest estimation for p results for the pizza question, so almost every surveyed participant seems to eat pizza at least once in a while. Unsurprisingly, the λ parameter is highest for coffee consumption, representing a high frequency of this behavior. Since the mean time interval between two occurrences is the inverse of λ (see above), the coffee drinkers among the survey students drink coffee every 0.46 weeks on average, so approximately every three to three and a half days. Meanwhile, participation in other surveys seems to be the rarest behavior with a λ parameter of 0.283, resulting in an average time interval between survey participation of about three and a half weeks. Table 2 also shows 95% confidence intervals for the estimated parameters, simultaneously showing the accuracy of measurement and serving as a means of hypothesis testing.

Table 2

Maximum Likelihood Estimates, SE, and 95% Bootstrap Confidence Intervals for p and λ, and Results of G*-Tests Evaluating the Appropriateness of the Proposed Model*

	p			λ				
Question	Estimate	*SE*	95% CI	Estimate	*SE*	95% CI	*G*	*p* value
Sports program	.230	.028	[.181, .291]	0.502	0.193	[0.241, 0.979]	1.649	.199
Crime thriller	.263	.028	[.211, .323]	0.382	0.104	[0.220, 0.625]	0.123	.726
Pizza	.945	.013	[.919, .968]	0.675	0.055	[0.574, 0.789]	0.001	.976
Coffee	.679	.019	[.642, .718]	2.394	0.862	[1.567, 4.956]	0.311	.577
Birthday congrat.	.891	.024	[.847, .937]	0.360	0.033	[0.300, 0.425]	0.058	.810
Other surveys	.739	.037	[.673, .817]	0.285	0.039	[0.211, 0.369]	0.107	.743

Note. The rate of occurrence λ has the dimension $(\text{week})^{-1}$. The point estimates, standard errors, and confidence intervals were calculated using a parametric bootstrap algorithm with 1,000 bootstrap samples. All *G*-tests were carried out with one degree of freedom ($df = 1$). The *p* values are presented for interpretation of the *G*-tests.

The prevalence curves predicted by the Poisson model are presented in Figure 4 alongside the observed prevalence of "yes" answers. The horizontal lines represent the estimated p values for each question. G-tests were conducted to evaluate model fit, with the results presented in the last two columns of Table 2. None of the tests yield significant results, indicating an appropriate model fit.

As mentioned earlier, with the parameter estimates of p and λ, we can now infer predictions about the prevalence in time frames that were not directly measured in the survey. For example, if we want to make statements about daily pizza consumption, we can insert the estimates $p = .945$ and $\lambda = 0.672$ along with $t = 1/7$ in the prevalence curve, yielding

$$P\left(\text{“yes”}|t = \frac{1}{7}\right) = .945 \cdot \left(1 - \mathrm{e}^{-0.672 \cdot \frac{1}{7}}\right) = .087. \quad (6)$$

According to the assumed Poisson process, the expected number and the standard deviation of events occurring within the time interval t is

$$E[N(t)|\text{carrier}] = \lambda \cdot t, \quad (7)$$

$$SD[N(t)|\text{carrier}] = \sqrt{\lambda \cdot t}\,. \quad (8)$$

Therefore, the estimate of λ can be used to compute the mean number and the variability of carriers' behaviors occurring within a specific time interval t. For pizza eaters, one would calculate their mean frequency of eating pizza per week as $E[N(1)|\text{carrier}] = 0.672$ with $SD[N(1)|\text{carrier}] = 0.820$. Furthermore, the resulting probability mass function of $N(1)$ as shown in Figure 5 is

$$P(N(1) = k|\text{carrier}) = \frac{(0.672 \cdot 1)^k \cdot \mathrm{e}^{-0.672 \cdot 1}}{k!}. \quad (9)$$

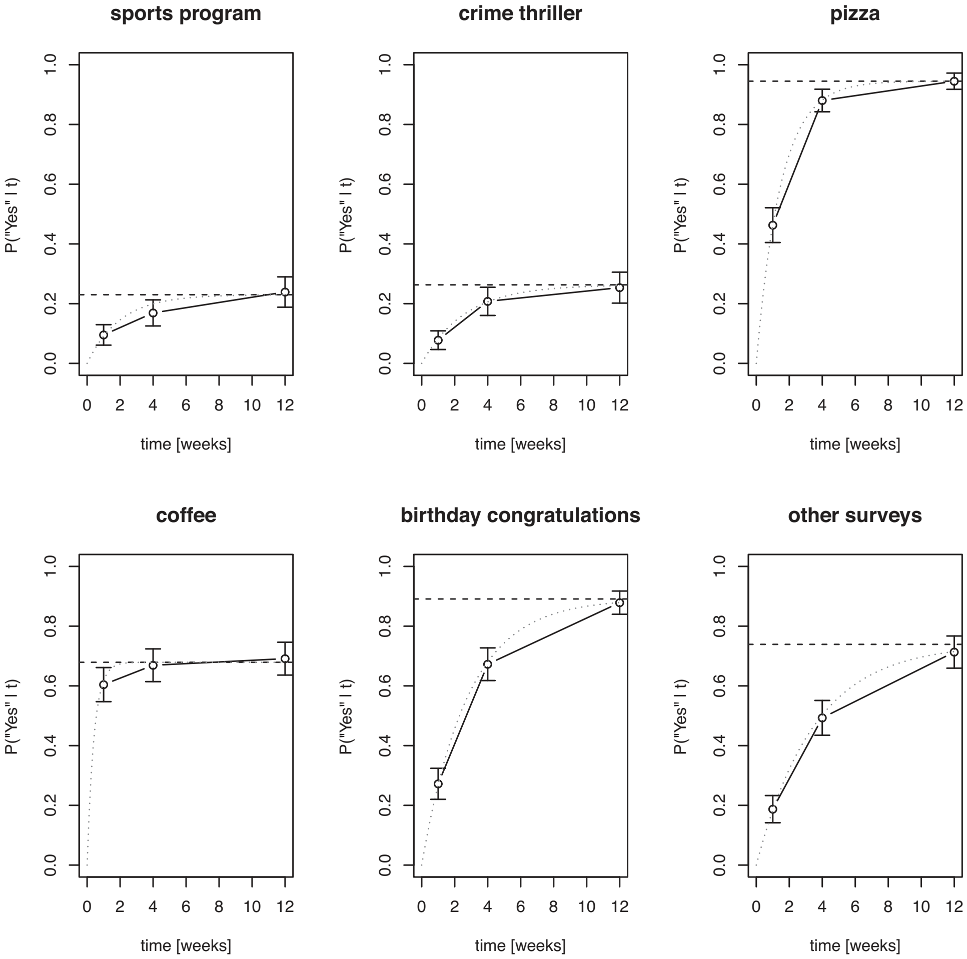
Thus, although we did not ask in the survey about the frequency of the target behavior within a certain time frame, the estimate for λ allows one to infer this frequency. Moreover, this method allows one to identify the true prevalence of trait carriers, p. Neither of these two pieces of information can be inferred directly from time-constrained yes/no questions.

Discussion

In surveys, researchers are often interested in the frequency of certain target behaviors within a specified time frame. For example, a

Figure 4

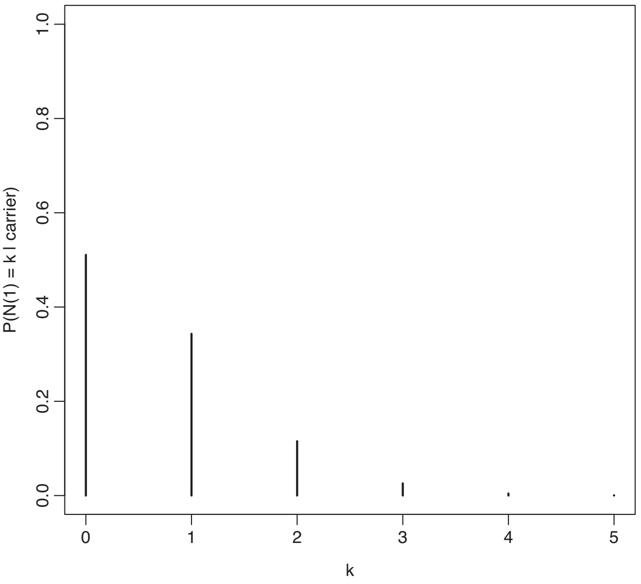
Observed Proportion of "Yes" Answers and Predicted Prevalence Curves According to the Proposed Model



Note. The error bars represent 95% confidence intervals.

Figure 5

Probability Mass Function for Pizza Eaters of Consuming k Pizzas Within a Week (i.e., t = 1)



health psychologist may investigate how frequently respondents take sleeping pills within a month. Asking respondents how often they took such pills within such a time frame is undoubtedly more memory-demanding to answer than merely asking them if they took any sleeping pills within this time frame at all. This may be the reason why researchers usually merely ask whether the target behavior was elicited within a certain time frame or at least once in a person's life. However, one drawback of such dichotomous yes/no questions is that their answers are difficult to interpret. First, a prevalence estimate based on a single time frame is compatible with an infinite number of prevalence curves that differ in the rate of occurrence and the proportion of carriers (compare Figure 3). Second, such a time-constrained prevalence does not offer any information about how frequently a target behavior occurs within a specific period of time.

We proposed a novel methodological approach to overcome these problems. This approach requires that time-constrained prevalence is observed for at least two different time frames (e.g., 1 week and 1 month). In the second step, a Poisson model is employed to estimate the underlying prevalence curve using a standard maximum likelihood procedure. This curve is determined by two parameters. First, the asymptote of this curve reflects the proportion of carriers, that is, respondents who potentially show the behavior in question. Second, the slope of this curve is governed by the rate of occurrence, allowing one to infer information about how often such behavior occurs within a specific time interval (e.g., 2 weeks). If more than two time frames are employed in a survey, the empirical adequacy of the model can be evaluated, for example, with standard goodness of fit tests.

We tested the model in a student sample, using six questions about day-to-day activities and habits, resulting in appropriate model fits. Moreover, the estimated parameters were plausible. For example, it seems natural that only about two-thirds of students are coffee drinkers. According to the predictions of the Poisson model, about 60% of the queried students drink coffee once a week. A survey in 2018 determined that 42% of German people between 18 and 35 describe themselves as daily coffee drinkers (Aral, 2018). Our prevalence estimate for coffee drinkers compares relatively well to these results, considering the age difference between samples—coffee consumption is reported lowest for younger people (Aral, 2018), so there are probably fewer daily coffee drinkers in student samples. Regarding pizza eaters, a recent study revealed that about 26% of students in Germany eat pizza at least weekly and that only 12% of them classified themselves as nonpizza eaters (VuMA Touchpoints, 2021b). However, the sample was only asked about their habit of eating frozen pizza (VuMA Touchpoints, 2021a). Considering that some students might prefer nonfrozen pizza from restaurants or delivery services, or even pizza made by themselves, our estimates also seem quite plausible here. As another example, our *p*-estimate for viewers of the sports program lies at .227. The proportion of people who like watching any sports programs "very much" is about 18% in Germany, with 14% watching less enthusiastically (IfD Allensbach, 2021). It is furthermore notable that this prevalence has been decreasing steadily since 2018 (IfD Allensbach, 2021). So, because we only asked about one specific sports program and our sample is comparatively less interested in such programs due to their young age, this estimate also seems to match the results of other sources rather well. Of course, more future empirical research is required to validate the reliability of this novel approach.

Poisson processes have often been employed to predict human behavior in various research domains. For example, they have been used to model the outcome of soccer games (e.g., Heuer et al., 2010; Nguyen, 2021; Zebari et al., 2021) or of phone calls directed at a call center (e.g., Bonilla-Escribano et al., 2020; Jiang et al., 2016; Weinberg et al., 2007). But also people connecting to wireless networks (Papadopouli et al., 2005; Tyagi et al., 2015) or parking space usage (Peng & Li, 2016) seem to be suitable for Poisson modeling. The Poisson process is one of the most important models for describing random events evenly distributed over time (Dehling & Haupt, 2003). So, it is not surprising that the Poisson model's results match well with the prevalence estimates in our study, resulting in low *G*-values.

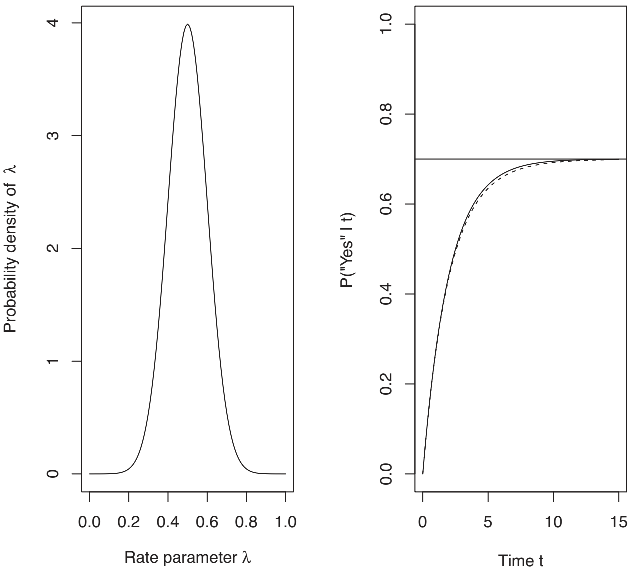
The proposed Poisson model in this article assumes that the rate λ is identical for all carriers. However, this assumption may be unrealistic because it might be too restrictive. To estimate the influence of such variation, let us assume that the rate λ is normally distributed with the mean μ standard deviation σ. Under this assumption, we can show that the prevalence curve takes the following explicit form (see Appendix B),

$$P(\text{“yes”}|t) = p \cdot \left[1 - \mathrm{e}^{-\mu\cdot t + 0.5\cdot\sigma^2\cdot t^2}\right]. \qquad (10)$$

It can be seen that this equation would reduce to the above prevalence curve (i.e., Equation 5) for $\sigma = 0$. Figure 6 shows an example of how the model's predictions change when we assume some variability in rates. The solid prevalence curve shows the prediction for $\sigma = 0$ and the dashed curve for $\sigma = 0.1$ with a variation coefficient of $(\sigma/\mu) = (0.1/0.5) = 0.2$; the theoretical distribution of λ is shown in the left figure. The surprising result is

Figure 6

Random Variation of the rate λ and the Shape of the Prevalence Curve



Note. Left panel: The rate λ is assumed to be normally distributed with mean $\mu = 0.5$ and standard deviation $\sigma = 0.1$. Right panel: Solid prevalence curves with $\sigma = 0$ and dashed curve with $\sigma = 0.1$. The prevalence parameter p is .7 in each case.

that the influence of σ on the prevalence curve appears to be negligible.

To test our impression that the addition of a variability parameter would be of little use, we fitted our data to the compound Poisson model, which includes the parameter σ. As can be seen in Table 3, neither the estimates of p and μ of the compound model, nor the G-values deviate much from their counterparts of the simple model (see Table 2). So, since the influence of σ appears to be quite small, we propose to use the simple model with $\sigma = 0$, unless a priori reasons let one suppose that the variation of the rates is relatively large.

The accuracy of the parameter estimators is crucial to the quality of the model's predictions. To achieve good accuracy with the standard Poisson model, we suggest using at least three time frames. While parameter estimation is already possible with only two time frames and thus two groups, the accuracy of the estimation increases with a third time frame, and the additional degree of freedom enables one to evaluate the empirical adequacy of the Poisson model. Moreover, the time frames should not be chosen arbitrarily.

When the target behavior is frequent, the parameter estimation would be inaccurate for relatively long time frames. For example, using 1 year as the shortest time frame when asking a question about drinking coffee would undoubtedly result in an imprecise estimation for λ. Similarly, when researching an infrequent target behavior, the accuracy of the estimation for p would likely be poor if only short time frames are used. For example, in a survey about hiking, the number of carriers (i.e., hikers) would probably be underestimated if the time-constrained questions used the time frames "last week," "last 2 weeks," and "last month." Parameter estimation for λ and p is best when the longest time frame yields a prevalence close to the asymptote of the prevalence curve and with the two remaining time frames covering the slope of the prevalence curve. Therefore, it seems advisable to perform some piloting before conducting a large-scale survey when little or nothing is known in advance about the behavior in question.

To test the reliability of the model, we simulated "true" occurrences of target behavior across 12 months for different values of λ and p (see Appendix D). We then tested how well the Poisson model recovers these parameters, using different sample sizes and combinations of three time frames as points of measurement. While we found some inaccuracies for certain combinations (e.g., too long time frames and a large true rate λ, or too short time frames and a small true rate λ), the model's estimates are reliable overall—and thus, predictions made on its basis would most likely be accurate. Of course, as mentioned above, further empirical applications are needed to truly assess the performance of predictions made with the Poisson model.

In conclusion, the proposed Poisson model provides valuable information about the number of carriers and noncarriers and the frequencies of the researched behaviors. Furthermore, the present study revealed that the model offers reasonable empirical estimates of target behaviors (i.e., about frequency and prevalence of these behaviors) by using time-constrained yes/no questions. Without such a model, this information cannot be directly inferred from the percentage of "yes" responses to those questions.

Table 3

Maximum Likelihood Estimates and 95% Confidence Intervals for p, μ, and σ and G*-Values of the Compound Poisson Model*

	p			μ			σ			
Question	Estimate	*SE*	95% CI	Estimate	*SE*	95% CI	Estimate	*SE*	95% CI	*G*
Sports program	.229	.029	[.182, .288]	0.525	0.212	[0.240, 1.091]	0.092	0.051	[0.000, 0.198]	1.518
Crime thriller	.264	.028	[.215, .326]	0.382	0.101	[0.220, 0.599]	0.026	0.037	[0.000, 0.102]	0.123
Pizza	.946	.013	[.919, .970]	0.681	0.057	[0.580, 0.808]	0.066	0.067	[0.000, 0.156]	0.000
Coffee	.682	.019	[.644, .720]	2.395	0.766	[1.590, 4.221]	0.251	0.154	[0.000, 0.512]	0.303
Birthday congr.	.892	.023	[.844, .936]	0.362	0.034	[0.301, 0.433]	0.037	0.036	[0.000, 0.085]	0.017
Other surveys	.745	.039	[.668, .822]	0.285	0.040	[0.218, 0.375]	0.035	0.028	[0.000, 0.073]	0.071

Note. The point estimates, standard errors, and confidence intervals were calculated using a parametric bootstrap algorithm with 1,000 bootstrap samples. G-tests are not carried out since there are no degrees of freedom ($df = 0$).

References

Aral. (2018). *Trends beim Kaffee-Genuss 2018* [Trends in coffee consumption 2018] [Brochure]. https://www.aral.de/content/dam/aral/business-sites/de/global/retail/presse/pressemeldungen/2018/kaffeestudie_2018.pdf

Atzendorf, J., Rauschert, C., Seitz, N.-N., Lochbühler, K., & Kraus, L. (2019). The use of alcohol, tobacco, illegal drugs and medicines: An estimate of consumption and substance-related disorders in Germany. *Deutsches Ärzteblatt International*, *116*(35–36), 577–584. https://doi.org/10.3238/arztebl.2019.0577

Beck, F., Léger, D., Fressard, L., Peretti-Watel, P., Verger, P., & Group, C. (2021). Covid-19 health crisis and lockdown associated with high level of sleep complaints and hypnotic uptake at the population level. *Journal of Sleep Research*, *30*(1), Article e13119. https://doi.org/10.1111/jsr.v30.1

Bonilla-Escribano, P., Ramírez, D., & Artés-Rodríguez, A. (2020). *Modeling phone call durations via switching Poisson processes with applications in mental health*. 2020 IEEE 30th international workshop on machine learning for signal processing (MLSP) (pp. 1–6). https://doi.org/10.1109/MLSP49062.2020.9231856.

Boos, D. D. (2003). Introduction to the bootstrap world. *Statistical Science*, *18*(2), 168–174. https://doi.org/10.1214/ss/1063994971

Cullen, K. A., Ambrose, B. K., Gentzke, A. S., Apelberg, B. J., Jamal, A., & King, B. A. (2018). Notes from the field: Use of electronic cigarettes and any tobacco product among middle and high school students—United States, 2011–2018. *Morbidity and Mortality Weekly Report*, *67*(45), 1276–1277. https://doi.org/10.15585/mmwr.mm6745a5

Dehling, H., & Haupt, B. (2003). Der Poisson-Prozess [The Poisson process]. In H. Dehling & B. Haupt (Eds.), *Einführung in die Wahrscheinlichkeitstheorie und Statistik* [Introduction to probability theory and statistics] (pp. 249–261). Springer. https://doi.org/10.1007/978-3-662-06893-9_12

Dietz, P., Iberl, B., Schuett, E., van Poppel, M., Ulrich, R., & Sattler, M. C. (2018). Prevalence estimates for pharmacological neuroenhancement in Austrian university students: Its relation to health-related risk attitude and the framing effect of caffeine tablets. *Frontiers in Pharmacology*, *9*, Article 494. https://doi.org/10.3389/fphar.2018.00494

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Heuer, A., Mueller, C., & Rubner, O. (2010). Soccer: Is scoring goals a predictable Poissonian process?. *EPL (Europhysics Letters)*, *89*(3), Article e38007. https://doi.org/10.1209/0295-5075/89/38007

Iberl, B., & Ulrich, R. (2022a). *Surveys on time-constrained questions (preliminary title)* [OSF Preregistration]. https://osf.io/z35y7

Iberl, B., & Ulrich, R. (2022b). *Surveys on time-constrained questions (preliminary title)* [OSF Project with data and code files]. https://osf.io/7ptsd/

IfD Allensbach. (2021). *Bevölkerung in Deutschland nach Beliebtheit von Sportübertragungen (live) oder Sportsendungen wie ZDF Sportstudio oder Sportschau im Fernsehen von 2017 bis 2021 (Personen in Millionen)* [Population in Germany by popularity of sports broadcasts (live) or sports programs like ZDF Sportstudio or Sportschau in television from 2017 to 2021 (in million people)]. Statista GmbH. https://de.statista.com/statistik/daten/studie/171190/umfrage/interesse-ansportuebertragungen-oder-sportsendungen-im-fernsehen/

Jiang, Z.-Q., Xie, W.-J., Li, M.-X., Zhou, W.-X., & Sornette, D. (2016). Two-state Markov-chain Poisson nature of individual cellphone call statistics. *Journal of Statistical Mechanics: Theory and Experiment*, *7*(7), Article e073210. https://doi.org/10.1088/1742-5468/2016/07/073210

Lin, C.-Y., Park, J.-H., Hsueh, M.-C., Sun, W.-J., & Liao, Y. (2018). Prevalence of total physical activity, muscle-strengthening activities, and excessive TV viewing among older adults; and their association with sociodemographic factors. *International Journal of Environmental Research and Public Health*, *15*(11), Article e2499. https://doi.org/10.3390/ijerph15112499

Miller, C., Ettridge, K., Wakefield, M., Pettigrew, S., Coveney, J., Roder, D., Durkin, S., Wittert, G., Martin, J., & Dono, J. (2020). Consumption of sugar-sweetened beverages, juice, artificially-sweetened soda and bottled water: An Australian population study. *Nutrients*, *12*(3), Article 817. https://doi.org/10.3390/nu12030817

Mohebbi, E., Haghdoost, A. A., Noroozi, A., Vardanjani, H. M., Hajebi, A., Nikbakht, R., Mehrabi, M., Kermani, A. J., Salemianpour, M., & Baneshi, M. R. (2019). Awareness and attitude towards opioid and stimulant use and lifetime prevalence of the drugs: A study in 5 large cities of Iran. *International Journal of Health Policy and Management*, *8*(4), 222–232. https://doi.org/10.15171/ijhpm.2018.128

Molinaro, S., Benedetti, E., Scalese, M., Bastiani, L., Fortunato, L., Cerrai, S., Canale, N., Chomynova, P., Elekes, Z., Feijão, F., & Fotiou, A. (2018). Prevalence of youth gambling and potential influence of substance use and other risk factors throughout 33 European countries: First results from the 2015 ESPAD study. *Addiction*, *113*(10), 1862–1873. https://doi.org/10.1111/add.v113.10

Nguyen, Q. (2021). *Poisson modeling and predicting English premier league goal scoring*. arXiv preprint. https://doi.org/10.48550/arXiv.2105.09881

Papadopouli, M., Shen, H., & Spanakis, M. (2005). *Modeling client arrivals at access points in wireless campus-wide networks*. 2005 14th IEEE workshop on local & metropolitan area networks (pp. 1–7). https://doi.org/10.1109/LANMAN.2005.1541514

Peng, L., & Li, H. (2016). *Searching parking spaces in urban environments based on non-stationary Poisson process analysis*. 2016 IEEE 19th international conference on intelligent transportation systems (ITSC) (pp. 1951–1956). https://doi.org/10.1109/ITSC.2016.7795871

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reiber, F., Bryce, D., & Ulrich, R. (2022). Self-protecting responses in randomized response designs; a survey on intimate partner violence during the coronavirus disease 2019 pandemic. *Sociological Methods & Research*, 1–32. https://doi.org/10.1177/00491241211043138

Şaşmaz, T., Öner, S., Kurt, A. Ö., Yapıcı, G., Yazıcı, A. E., Buğdaycı, R., & Şiş, M. (2014). Prevalence and risk factors of internet addiction in high school students. *The European Journal of Public Health*, *24*(1), 15–20. https://doi.org/10.1093/eurpub/ckt051

Sawyer, A. N., Smith, E. R., & Benotsch, E. G. (2018). Dating application use and sexual risk behavior among young adults. *Sexuality Research and Social Policy*, *15*(2), 183–191. https://doi.org/10.1007/s13178-017-0297-6

Seitz, N.-N., Rauschert, C., Atzendorf, J., & Kraus, L. (2020). *IFT-Berichte Bd. 190: Berlin, Hessen, Nordrhein-Westfalen, Sachsen und Thüringen. Ergebnisse des Epidemiologischen Suchtsurvey 2018* [IFT-Reports Vol. 190: Substance use and substance use disorders in Berlin, Hesse, North Rhine-Westphalia, Saxony and Thuringia. Results of the 2018 Epidemiological Survey of Substance Abuse]. Institut für Therapieforschung München.

Soga, M., Evans, M. J., Tsuchiya, K., & Fukano, Y. (2021). A room with a green view: The importance of nearby nature for mental health during the COVID-19 pandemic. *Ecological Applications*, *31*(2), Article e2248. https://doi.org/10.1002/eap.2248

Tyagi, R. R., Aurzada, F., Lee, K. -D., & Reisslein, M. (2015). Connection establishment in LTE-A networks: Justification of Poisson process modeling. *IEEE Systems Journal*, *11*(4), 2383–2394. https://doi.org/10.1109/JSYST.2014.2387371

Ulrich, R., Pope, H. G., Cléret, L., Petróczi, A., Nepusz, T., Schaffer, J., Kanayama, G., Comstock, R. D., & Simon, P. (2018). Doping in two elite athletics competitions assessed by randomized-response surveys. *Sports Medicine*, *48*(1), 211–219. https://doi.org/10.1007/s40279-017-0765-4

Virudachalam, S., Long, J. A., Harhay, M. O., Polsky, D. E., & Feudtner, C. (2014). Prevalence and patterns of cooking dinner at home in the USA: National Health and Nutrition Examination Survey (NHANES) 2007–2008. *Public Health Nutrition*, *17*(5), 1022–1030. https://doi.org/10.1017/S1368980013002589

VuMA Touchpoints. (2021a). *Den Markt im Blick–Basisinformationen für fundierte Mediaentscheidungen* [Sights on the market–basic information for well-founded media decisions]. ARD Media, RMS & ZDF Werbefernsehen.

VuMA Touchpoints. (2021b). *Studenten in Deutschland nach Häufigkeit des Konsums von Pizza (Baguettes) im Vergleich mit der Bevölkerung im Jahr 2021* [Students in Germany by frequency of pizza (baguette) consume compared to the general public in 2021]. Statista GmbH.

Weinberg, J., Brown, L. D., & Stroud, J. R. (2007). Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association*, *102*(480), 1185–1198. https://doi.org/10.1198/016214506000001455

Zebari, G. M., Zeebaree, S., Sadeeq, M. M., & Zebari, R. (2021). Predicting football outcomes by using Poisson model: Applied to Spanish Primera División. *Journal of Applied Science and Technology Trends*, *2*(4), 105–112. https://doi.org/10.38094/jastt204112

(*Appendices follow*)

Appendix A

Maximum Likelihood Estimation

The likelihood function of the proposed Poisson model is

$$L(p, \lambda) = \prod_{j=1}^{m} \prod_{i=1}^{n_j} \left[P(\text{“yes”}|t_j)^{a_j} \cdot P(\text{“no”}|t_j)^{b_j} \right] \tag{A1}$$

with the number of groups m, the sample size per group n_j, the time frame per group t_j that the question refers to, the observed number of “yes” answers per group a_j and the observed number of “no” answers per group b_j. Taking the log of Equation A1 gives

$$\log L(p, \lambda) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left[a_j \cdot \log\left[P(\text{“yes”}|t_j)\right] + b_j \cdot \log\left[P(\text{“no”}|t_j)\right] \right]. \tag{A2}$$

Maximizing Equation A2 by a numerical search routine yields the maximum likelihood estimators for the parameters λ and p.

Appendix B

Compound Poisson Model

Let f be the probability density function associated with the rate of occurrence λ. The compound Poisson model is then generally given as

$$P(\text{“yes”}|t) = \int P(\text{“yes”}|t, \lambda) f(\lambda)\, d\lambda \tag{B1}$$

$$= p \cdot \left[1 - \int e^{-\lambda \cdot t} f(\lambda)\, d\lambda \right]. \tag{B2}$$

Note that the integral resembles the moment-generating function $M(t)$ of f, that is,

$$M(t) = \int e^{\lambda \cdot t} f(\lambda)\, d\lambda \tag{B3}$$

with the modification that t must be replaced by $-t$ in the respective moment-generating function. Therefore, we can generally write

$$P(\text{“yes”}|t) = p \cdot [1 - M(-t)]. \tag{B4}$$

For example, the moment-generating function of a normal distribution with μ and σ is given by

$$M(t) = e^{\mu \cdot t + 0.5 \cdot \sigma^2 \cdot t^2} \tag{B5}$$

and thus

$$P(\text{“yes”}|t) = p \cdot \left[1 - e^{-\mu \cdot t + 0.5 \cdot \sigma^2 \cdot t^2} \right]. \tag{B6}$$

Clearly, $\mu > 0$ should be relatively large with respect to σ, such that f is virtually zero at $\mu = 0$. Therefore, in all computations for the compound model, we restricted σ such that $\sigma < 0.2 \cdot \mu$ but >0.

Another candidate for f could be a rectangular distribution with $0 < a \le \lambda \le b$ and the moment-generating function

$$M(t) = \frac{e^{b \cdot t} - e^{a \cdot t}}{t \cdot (b - a)} \tag{B7}$$

yielding

$$P(\text{“yes”}|t) = p \cdot \left[1 - \frac{e^{-a \cdot t} - e^{-b \cdot t}}{t \cdot (b - a)} \right]. \tag{B8}$$

Appendix C

Statistical Analysis With R

In this appendix, we give an example of the statistical analysis with the proposed (two-parameter) Poisson model. For this purpose, we use the data of the “pizza” question. Note that the parameter estimates will differ slightly with each computation due to the sampling in the parametric bootstrap method.

```
## data ##
lim          <- 1e-10             # lower limit for p and lambda (0)
up_lim_lam   <- 7                 # upper limit for lambda (7)
t0           <- c( 1, 4, 12)      # time frame per group [weeks]
N.t          <- c(283, 284, 272)  # sample size per group
a            <- c(131, 250, 257)  # observed yes-answers (pizza)
b            <- N.t - a           # observed no-answers
```

(*Appendices continue*)

```
## functions ##
# function of the prevalence curve #
pc <- function(t, p, lam) {p*(1-exp(-lam*t))}
# log-likelihood function for maximum likelihood estimation
MLE <- function(par, a, b){
        p <- par[1]
        lam <- par[2]
        pyes <- pc(t0, p, lam)
        lL <- a*log(pyes) + b*log(1-pyes)
        MLE <- -sum(lL)
        }
# G function for testing model fit
Gf <- function(par, a, b){
        p <- par[1]
        lam <- par[2]
        E.t.yes <- pc(t0, p, lam)*N.t
        E.t.no <- N.t - E.t.yes
        G <- 2*sum(a*log(a/E.t.yes) +
        b*log(b/E.t.no))
        return(G)
        }
## parameter estimation via bootstrap sampling (1000 samples) ##
# creating vectors for the observed yes-/no-answers in each group
obs.t1 <- c(rep(1, a[1]), rep(0, b[1]))
obs.t2 <- c(rep(1, a[2]), rep(0, b[2]))
obs.t3 <- c(rep(1, a[3]), rep(0, b[3]))
# bootstrap sampling
p.b    <- numeric(1000)
lam.b  <- numeric(1000)
for(i in 1:1000){
        # resampling a and b from observed data
        a.b <- c(sum(sample(x=obs.t1, size=N.t[1], replace=TRUE)),
                 sum(sample(x=obs.t2, size=N.t[2], replace=TRUE)),
                 sum(sample(x=obs.t3, size=N.t[3], replace=TRUE)))
        b.b <- N.t - a.b
        # maximum likelihood estimation of the redrawn sample
        ML <- optim(par = c(0.5, 0.8), a = a.b, b = b.b,
                    fn = MLE, method = 'L-BFGS-B',
                    lower = c(lim, lim),
                    upper = c(1-lim, up_lim_lam))
        # extracting p and lambda estimates
        p.b[i]   <- ML$par[1]
        lam.b[i] <- ML$par[2]
        }
# extracting parameter estimators, SEs and 95%-CIs
p.m    <- mean(p.b)                      # point estimate for p
p.se   <- sd(p.b)                        # standard error for p
p.ci   <- quantile(p.b, c(.025,.975))    # 95%-CI for p
lam.m  <- mean(lam.b)                    # point est. for lambda
am.se  <- sd(lam.b)                      # se for lambda
lam.ci <- quantile(lam.b, c(.025,.975)) # 95%-CI for lambda
```

(*Appendices continue*)

```
## G-test ##
# G estimation
Gest <- optim(c(0.5,0.8), a = a, b = b,
              fn = Gf, method='L-BFGS-B',
              lower = c(lim, lim),
              upper = c(.999, up_lim_lam))
# extracting G- and p-values
Gval <- Gest$value
pval <- pchisq(Gest$value, df = 1, lower.tail=FALSE)
```

Appendix D

Parameter Recovering Performance

The Monte-Carlo study in this appendix investigates the performance of the Poisson model in recovering the parameters p and λ that were used to simulate time-constrained yes/no survey data. More specifically, the study simulated the occurrence of such behavior over the past 12 months for a sample consisting of n virtual participants. Each virtual participant was a carrier with the probability p and a noncarrier with the complementary probability $1 - p$.

Table D1 illustrates a simulated data set. For each of the preceding 12 months, this set shows the simulated occurrences of the target behavior for each virtual participant. For example, for Participant 5 (a carrier) the target behavior occurred thrice during the past 3 months. Moreover, each simulation allotted a participant to one of three time frames (e.g., the past 2, 6, or 12 months). Then it was registered whether the target behavior occurred within this time frame or not. For example, for the past 2 months, Participants 1, 2, 3, 4, and 5 would respond with "no," "yes," "no," "no," and "yes," respectively. In each simulation, such a data set with yes/no answers was generated, and the parameters p and λ were estimated with the R code described in Appendix C.

The simulation study orthogonally combined three sets of time frames (Set 1: 1, 3, and 9 months; Set 2: 2, 6, and 12 months; Set 3: 1, 2, and 6 months), p (.5 or .7), λ (0.2, 0.5, or 1) and sample size n (250 or 500 participants per group/time frame, so total sample size is 750 or 1,500). The measurement unit of λ is 1/month, that is, λ represents the average number of monthly occurrences. For each of the 36 factorial combinations, 1,000 data sets were simulated, and for each set, the parameters p and λ were estimated. The results are presented in Tables D2, D3, and D4.

The simulations revealed that the model's performance in recovering the parameters p and λ is rather reliable. However, estimation accuracy depends on the specified time frames. For example, λ estimation in simulations 13 and 14 (see Table D3) is somewhat imprecise, since the true underlying rate is high, but there is no point of measurement for the shortest time frame. So, the slope of the underlying prevalence curve is not captured precisely. In simulations 29 and 30 as well as in simulations 35 and 36 (see Table D4) on the other hand, the estimator for p is relatively inaccurate. This is due to the low rate of the simulated behavior and the short time frames used for measurement. In this case, the asymptote of the prevalence curve is not adequately captured, which worsens the estimation of p. In summary, this simulation study has revealed that in most cases the parameter recovering is excellent.

Table D1

Example for a Data Set Containing Simulated Occurrences of Behavior in Each of the Preceding 12 Months

		*i*th preceding month							
ID	Carrier	1	2	3	4	5	6	…	12
1	Yes	0	0	0	1	1	0	…	1
2	Yes	1	0	0	2	0	1	…	0
3	No	0	0	0	0	0	0	…	0
4	No	0	0	0	0	0	0	…	0
5	Yes	1	2	0	0	0	2	…	1
6	No	0	0	0	0	0	0	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	Yes	0	3	0	0	1	0	…	1

Table D2

Results of Simulations 1–12

				p Estimate		λ Estimate	
Simulation no.	True p	True λ	n	*M*	*SE*	*M*	*SE*
1	.5	1.0	250	.501	.020	1.037	0.220
2	.5	1.0	500	.501	.014	1.011	0.144
3	.5	0.5	250	.502	.029	0.508	0.093
4	.5	0.5	500	.501	.021	0.502	0.063
5	.5	0.2	250	.526	.103	0.201	0.060
6	.5	0.2	500	.508	.063	0.203	0.042
7	.7	1.0	250	.701	.019	1.017	0.152
8	.7	1.0	500	.699	.013	1.014	0.104
9	.7	0.5	250	.703	.027	0.506	0.068
10	.7	0.5	500	.702	.019	0.501	0.048
11	.7	0.2	250	.717	.093	0.202	0.045
12	.7	0.2	500	.709	.063	0.201	0.032

Note. Simulations for time frame set 1, 3, and 9 months.

(*Appendices continue*)

Table D3
Results of Simulations 13–24

				p Estimate		λ Estimate	
Simulation no.	True *p*	True λ	*n*	*M*	*SE*	*M*	*SE*
13	.5	1.0	250	.501	.014	1.221	0.962
14	.5	1.0	500	.500	.010	1.101	0.601
15	.5	0.5	250	.502	.022	0.510	0.104
16	.5	0.5	500	.501	.015	0.509	0.077
17	.5	0.2	250	.509	.062	0.206	0.054
18	.5	0.2	500	.504	.040	0.204	0.037
19	.7	1.0	250	.701	.013	1.073	0.469
20	.7	1.0	500	.700	.009	1.029	0.156
21	.7	0.5	250	.702	.019	0.501	0.072
22	.7	0.5	500	.701	.014	0.502	0.052
23	.7	0.2	250	.709	.060	0.203	0.040
24	.7	0.2	500	.702	.039	0.203	0.027

Note. Simulations for the time frame set 2, 6, and 12 months.

Table D4
Results of Simulations 25–36

				p Estimate		λ Estimate	
Simulation no.	True *p*	True λ	*n*	*M*	*SE*	*M*	*SE*
25	.5	1.0	250	.501	.026	1.037	0.214
26	.5	1.0	500	.501	.017	1.014	0.139
27	.5	0.5	250	.508	.043	0.502	0.104
28	.5	0.5	500	.502	.030	0.503	0.074
29	.5	0.2	250	.563	.184	0.203	0.083
30	.5	0.2	500	.529	.133	0.206	0.064
31	.7	1.0	250	.701	.024	1.017	0.149
32	.7	1.0	500	.702	.016	1.001	0.099
33	.7	0.5	250	.703	.043	0.508	0.081
34	.7	0.5	500	.702	.028	0.501	0.051
35	.7	0.2	250	.729	.148	0.206	0.065
36	.7	0.2	500	.720	.116	0.203	0.047

Note. Simulations for the time frame set 1, 2, and 6 months.

Received June 14, 2022
Revision received March 11, 2023
Accepted March 31, 2023 ■

C Paper 3

Iberl, B., Aljovic, A., Ulrich, R., & Reiber, F. (2024). The Poisson extension of the Unrelated Question Model - Improving surveys with time-constrained yes/no questions on sensitive topics. *Survey Research Methods, 18*(1), 21-38. https://doi.org/10.18148/srm/2024.v18i1.8252.

Candidate contributions to the article

Status	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Published	30%	35%	30%	50%

Survey Research Methods (2024)
Vol. 18, No. 1, pp. 21-38
https://doi.org/10.18148/srm/2024.v18i1.8252
European Survey Research Association

The Poisson Extension of the Unrelated Question Model: Improving Surveys with Time-Constrained Questions on Sensitive Topics

Benedikt Iberl[1] · Anesa Aljovic[1] · Rolf Ulrich[1] · Fabiola Reiber[2]
[1]Eberhard Karls University of Tübingen
[2]University of Mannheim

The Poisson model (Iberl & Ulrich, 2023) is a new survey technique that enables the estimation of how frequently a certain behavior occurs, while employing easy-to-answer yes/no-questions that refer to a specific time frame (e.g., "Did you participate in gambling during the last 12 months?"). In this paper, this model is combined with the unrelated question model (UQM) by Greenberg et al. (1969). The UQM is another survey technique that guarantees complete and objective anonymity to participants in order to achieve more valid survey results when asking sensitive questions (e.g., about drug use). The resulting Poisson extension of the UQM (UQMP) is expected to yield valid estimations for how many participants engage in a researched sensitive behavior, and how regularly they do so. The performance of the UQMP was compared to the performance of the standard Poisson model, employing direct questions, in a survey on drinking and driving. While prevalence estimates differ greatly between the UQMP and the standard Poisson model, the results of both models indicate a high rate of drinking and driving among those German traffic participants who generally engage in this behavior. The different prevalence estimates could be due to the fact that some participants in online studies read instructions superficially, lowering the quality of results; we discuss possible causes for these problems and why the UQMP or similar approaches can be valuable nonetheless.

Keywords: Randomized Response Technique; Unrelated Question Model; survey research; time-constrained questions; prevalence curve; Poisson process

1 Introduction

In survey research, one is often interested in obtaining prevalence estimates describing a certain target behavior. Prevalence estimates can be useful in many politically or socially important fields, such as for the assessment of public opinion or to evaluate the frequency of criminal or risky behavior like drug abuse. Oftentimes, these prevalence estimates are produced by posing yes/no questions that refer to a particular time frame, such as "Did you gamble in the past 12 months" (e.g., Andrie et al., 2019; Atzendorf et al., 2019; Beck et al., 2021; Birkel et al., 2022; Burr et al., 1989; Ferrante et al., 2012; Han et al., 2015; Isolauri & Laippala, 1995; Linton et al., 1998; McCabe et al., 2006; McKetin et al., 2006; Şaşmaz et al., 2014; Sawyer et al., 2018; Virudachalam et al., 2014). In this paper, we will call such questions *time-constrained yes/no questions*.

However, one might not only be interested in whether the concerning behavior has occurred within a certain time frame, but also how often the behavior is shown. So, besides the information about whether someone was gambling in the last year, a researcher might be interested in the rate of this behavior (i.e., the average frequency of the concerning behavior per time unit). To measure this rate, one could simply ask participants how often they have engaged in the behavior in question within a certain time frame (e.g., "How often did you gamble in the past 12 months?"). This kind of questioning technique is also widely used in prevalence research (e.g., Cullen et al., 2018; Miller et al., 2020; Molinaro et al., 2018; Seitz et al., 2020; Soga et al., 2021). Responding to questions that require more than a simple yes or no answer may present some challenges compared to time-constrained yes/no questions. Answering time-constrained yes/no questions might be quicker and less demanding for participants since they only need to recall one instance of the behavior in question. Although there is no direct re-

Corresponding author: Benedikt Iberl, Eberhard Karls University of Tübingen, Tübingen, Germany (Email: [redacted])

search comparing the effort needed to answer time-constrained yes/no questions with those asking about behavior frequency, studies suggest that retrieving multiple memories of events or behavior instances can be more taxing for participants (e.g., Aarts & Dijksterhuis, 1999; Bousfield & Sedgewick, 1944; Echterhoff & Hirst, 2006; Janssen et al., 2011; Schwarz et al., 1991).

In conclusion, these questions share a fundamental weakness: The resulting prevalence estimates are ambiguous and do not yield reliable information about the number of people regularly engaging in the behavior, or *trait carriers*. For example, in a study on addictive behavior, Andrie et al. (2019) asked students in several European countries whether they gambled in the past 12 months. According to the results, 12.5% of the surveyed participants gambled in the past year (Andrie et al., 2019). Obviously, these results are not conclusive regarding the number of trait carriers, that is, regular gamblers within the student population. Instead, they only yield a punctually relevant prevalence estimation. For instance, there might be gamblers that did not gamble in the past year; so, this past-year prevalence of 12.5% is obviously not the same as the prevalence of gamblers in the underlying population. Assuming otherwise would result in an underestimation of the prevalence one wants to measure. One might try to circumvent this ambiguity by expanding the time frame in the posed question, measuring the lifetime prevalence in the most extreme example. However, with such broad time frames, some respondents who are not gambling on a regular basis, but only did so once or twice a long time ago, might be included in the prevalence estimate, despite one would not describe them as gamblers (Fiedler & Schwarz, 2016). Thus, an inflated estimate would result. Another solution might be to ask the participants directly whether they consider themselves to be gamblers. While this would undoubtedly be the most straight forward approach, self-assessments might yield problematic results as well (e.g., due to social desirability bias).

In the following, we introduce a recently proposed method that can solve both mentioned problems of time-constrained yes/no questions (no information about the rate of the behavior and ambiguity of prevalence estimates due to punctual information) while still using the same kind of questions (Iberl & Ulrich, 2023). Based on a *Poisson process*, this method might be an efficient solution to these problems compared to the mentioned traditional alternatives.

1.1 The Poisson model: A solution for the problems of time-constrained questions?

This *Poisson model* (Iberl & Ulrich, 2023) yields prevalence estimates of *trait carriers* (and, in turn, of *non-carriers*). Additionally, it becomes possible to estimate the rate of the behavior in question. Nonetheless, nothing changes for the participants — they still get asked simple time-constrained yes/no questions; however, they are split into multiple groups. Between groups, the questions are varied slightly: For each group, the respective question refers to a different time-frame t. Since the Poisson model is based on a Poisson process, it can be used to describe any form of behavior that can be assumed to occur regularly and periodically, for example, driving a car, drinking coffee, or smoking cigarettes.

In Fig. 1, the Poisson model is depicted as a probability tree. This tree shows the probability of answering “yes” or “no” to any question on whether a respondent behaved in a certain way in a specific time frame t. According to the model, the probability of being a carrier is π, with the probability of being a non-carrier being defined by $1–\pi$.

Non-carriers, who are represented in the lower branch of the tree, would always answer “no” to a question asking whether they behaved in a certain way (e.g., whether they gambled) in a certain time frame t. The probability of a no-answer would always be 1 for non-carriers, regardless of the time frame, because they do not engage in the behavior in question. For carriers, on the other hand, two answers are possible. One group of carriers could answer “no”, because they did not show the behavior in the time frame t referred to in the question ($N(t) = 0$). The other group of carriers might have engaged in the behavior at least once in the time frame (so $N(t) = 0$), and would thus answer “yes”.

Since we assume the behavior to be Poisson distributed, $N(t)$ represents a random variable with the rate parameter λ, which denotes the average number of occurrences of the target behavior per time unit. In other words, the reciprocal of λ is the average interoccurrence time. In addition, the probability of k occurrences of the target behavior within the time frame t is given by

$$P(N(t) = k) = \frac{(\lambda \cdot t)^k \cdot e^{-\lambda \cdot t}}{k!}. \tag{1}$$

Thus, the probability of a no-answer is

$$P(N(t) = 0) = e^{-\lambda \cdot t}. \tag{2}$$

A random participant would answer with “yes” to the time-constrained prevalence question with the probability

$$P(\text{"yes"} \mid t) = \pi \cdot P(N(t) > 0) \tag{3}$$

or

$$P(\text{"yes"} \mid t) = \pi \cdot [1 - P(N(t) = 0)]. \tag{4}$$

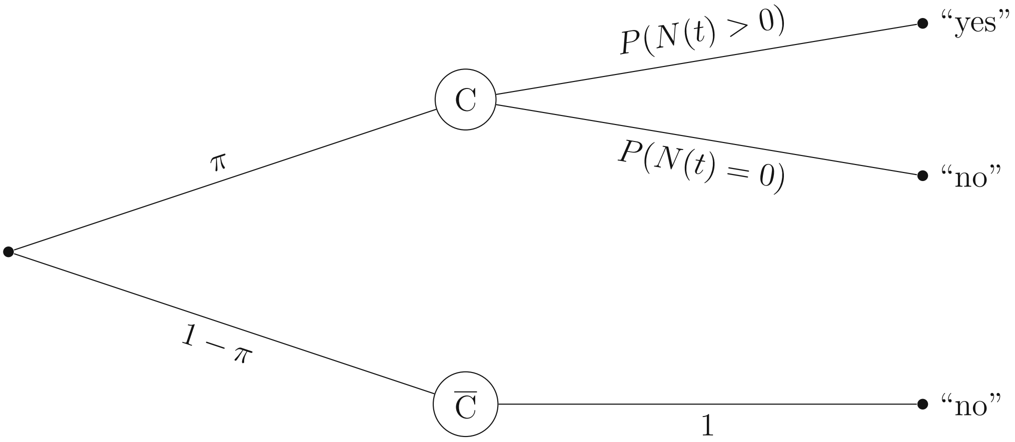


Fig. 1

Probability tree of the Poisson model. The sample is divided into carriers C and non-carriers $\overline{\mathrm{C}}$ by the parameter π, describing the probability of a random participant being a carrier of the researched attribute. Non-carriers answer "no" with a probability of 1. Carriers answer "yes" with a probability $P(N(t) > 0)$ or "no" with a probability of $P(N(t) = 0)$

Inserting the formula of the Poisson process, one gets the *prevalence curve*,

$$P(\text{"yes"} \mid t) = \pi \cdot \left(1 - e^{-\lambda \cdot t}\right), \tag{5}$$

depicting the prevalence of the behavior as a function of time, with the parameters π and λ determining the asymptote and the slope of the curve, respectively. Fig. 2 shows exemplary prevalence curves and the effects of different parameter values for π and λ.

The estimation of the parameters π and λ is enabled by using multiple groups of participants. As mentioned before, the time frame t of the question is varied between groups. For example, one group of participants would be asked if they gambled in a time frame of $t_1 = 1$ week, while another would be asked the same question referring to the time frame of $t_2 = 4$ weeks, and so on. With at least two time frames t_i, it is possible to estimate π and λ and thus determine the prevalence curve describing the probability of occurrence over time for the researched behavior. Parameter estimation is performed with the maximum likelihood procedure (see Supplementary Material).

Iberl and Ulrich (2023) have shown that the Poisson model can be applied to questions about everyday behavior, like drinking coffee, watching sports, and eating pizza. While the Poisson model has some weaknesses compared to traditional methods (e.g., the strict assumption of the researched behavior being Poisson-distributed and the need for larger sample sizes), it offers a novel approach to the mentioned problems in prevalence research. Of course, the model can theoretically also be used for any other behavioral prevalence measurement. In this regard, it would be particularly interesting to apply the model to sensitive topics, like drug usage or violent behavior. In this context, however, another problem arises, which the Poisson model does not address, that is, the problem of social desirability bias. Especially for research about the prevalence of crime, victimization, drug use or other socially relevant topics, the more indicative prevalence information provided by the

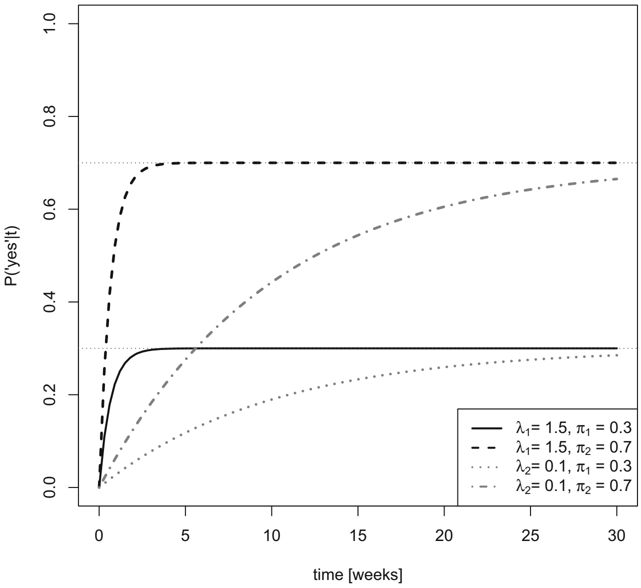


Fig. 2

Examples of prevalence curves as a function of π and λ with varying values for both parameters

Poisson model could be of special interest. Even the example of gambling mentioned above might be seen as a sensitive topic by some, since this topic is oftentimes associated with addiction. Unfortunately, it is well-documented that asking direct questions about sensitive topics can lead to higher amounts of socially desirable answers, mostly resulting in an underestimation of the prevalence of interest and thus a loss of validity (for reviews of social desirability research see, e.g., Krumpal, 2013; Nederhof, 1985).

1.2 Asking sensitive questions with the randomized response technique

To solve this problem, Warner (1965) designed a then-novel questioning approach, the *randomized response technique* (RRT). The basic idea of this approach is that the connection between the question of interest (about a sensitive topic, e.g., drug use) and the corresponding answer is masked by a random component, enabling anonymity for the participants, thus leading to more honest answers in turn. Over time, plenty of related models (which can be summarized under the term *randomized response models* or RRMs) emerged, each building on this basic idea. One relatively widely used model is the *unrelated question model* (UQM) by Greenberg et al. (1969). While some RRMs, for example, the *forced response model* (Boruch, 1971), require participants to lie under certain circumstances, which could be socially undesirable in itself, participants are required to always answer honestly in the UQM. Because of this, the UQM might be psychologically acceptable to participants (Höglinger et al., 2016; Reiber, Bryce, & Ulrich, 2022; Reiber et al., 2020; Ulrich et al., 2018).

The probability tree for this model is presented in Fig. 3. In the UQM, participants of a survey on sensitive topics are asked one of two questions; a sensitive question (e.g., drug use) or a neutral (or *unrelated*) question. A Bernoulli experiment (e.g., a dice roll), with the probability p set by design, is conducted by the participants themselves, and precedes the question. It is important that the result of this random experiment is kept secret by the participants and is only known to them. In the case of the first outcome, with the probability p, a participant is confronted with the sensitive question. In the case of the other outcome, with the counter probability 1–p, the participants are meant to answer the neutral question. The participants' answer ("yes" or "no") is recorded afterwards, while only they know which question they answered to. Due to the masking via the random experiment, the resulting yes- or no-answer of any participant could refer to either the sensitive or the neutral question. The sensitive question, under the assumption of honest answers by participants, will be answered with "yes" with the unknown probability π, or with "no" (and the probability 1–π). The neutral question, on the other hand, has to regard a topic of which the prevalence is known or can be estimated. In practice, birth dates, which are roughly uniformly distributed, are frequently used for this purpose. For example, a question like "Is your birth date in the first half of the year, so before the 1st of July?" can be used. Thus, the probability q of answering this neutral question with "yes" is set by design — in the aforementioned example, $q \approx 0.5$. Moreover, birth dates have also been used as a randomization device for the parameter p (e.g., Dietz et al., 2018).

In summary, the model consists of two design parameters, that is, the probabilities p, to be assigned the sensitive question, and q, to answer the neutral question with "yes", and one unknown parameter of interest, the prevalence π of the sensitive attribute or behavior. The probability of a yes-answer, γ (we renamed this parameter to avoid confusion since it is originally labeled λ like the rate in the Poisson model) is then

$$\gamma = p \cdot \pi + (1 - p) \cdot q, \tag{6}$$

according to the model. γ can be estimated via the observable relative frequency of yes-answers. With $\widehat{\gamma}$, π can be estimated by

$$\widehat{\pi} = \frac{\widehat{\gamma} - (1 - p) \cdot q}{p}. \tag{7}$$

The variance of π is

$$\sigma^2_\pi = \frac{\gamma \cdot (1 - \gamma)}{n \cdot p^2}, \tag{8}$$

and 95% confidence intervals can be formed by

$$\widehat{\pi} \pm 1.96 \cdot \sqrt{\widehat{\sigma}^2_\pi}. \tag{9}$$

Notably, other than in the Poisson model, π is defined with respect to the time frame posed by the question. Thus, if the question states, "Did you gamble in the last year?", π refers to the one-year prevalence of gambling (i.e., anyone who gambled during this time), not the proportion of gamblers (i.e., anyone who gambles regularly, independent of the exact time frame). Consequently, like any other RRM, the UQM faces the same problems of ambiguity and inability to estimate rates of occurrence as traditional *direct questioning* techniques (DQ) when it comes to measuring prevalence of behavior due to time-constrained questions. While multiple authors have already designed RRMs that can be used for sensitive quantitative variables (e.g., Greenberg et al., 1971; Himmelfarb & Edgell, 1980; Huang et al., 2006; Kumar, 2022; Liu & Chow, 1976), these solutions still comprise the problem that those kinds of questions might be more difficult to answer, as explained above.

1.3 The UQMP: A new approach for time-constrained questions on sensitive topics

In this paper, we propose a new approach, combining the benefits of the Poisson model and RRMs, enabling a questioning technique which is both independent of time constraints and valid for questions about sensitive topics. Since the UQM has some qualities that distinguishes it from other RRMs, the paper at hand will focus on this particular model. This is because, for one, the UQM is regarded as more psychologically acceptable than several other RRMs, as already mentioned. Additionally, it is one of the most efficient RRMs (Ulrich et al., 2018).

Our proposed extension of the UQM via the Poisson model — let us call it the *UQMP* — is depicted in Fig. 4. As can easily be seen when comparing it to the UQM presented in Fig. 3, the proposed UQMP basically extends the classic UQM with the possibility to distinguish carriers from non-carriers, independent of time constraints. Like in the UQM, the probability tree of the UQMP spreads into two main branches. The upper branch represents participants led to the sensitive question, the lower one represents participants getting assigned the neutral question. The lower branch is identical to that of the UQM, leading to the possibilities of participants answering "yes" or "no" (with the probabilities q or 1–q, respectively). However, in the upper branch, the parameter π does not represent the probability of giving a positive answer to the sensitive question, like it is the case in the UQM. Instead, it is defined as the probability that a random participant drawing the sensitive question is a carrier of the researched attribute, like in the Poisson model. From there on out, like in the standard Poisson model, the non-carriers are assumed to always answer "no", while the carriers might answer either "yes" or "no", depending on the time frame t that the sensitive question refers to.

The probability of a yes-answer to a question referring to the time frame t in the UQMP is

$$P(\text{"yes"} \mid t) = p \cdot \pi \cdot \left(1 - e^{-\lambda \cdot t}\right) + (1 - p) \cdot q, \qquad (10)$$

with λ representing the average rate of occurrence of the researched behavior, like in the standard Poisson model.

Similar to the Poisson model, we can estimate the parameter values of π and λ by varying the time frames t_i that the sensitive question is referring to (the neutral question has to be invariant between groups, so that q is constant). At least two time frames t_i are needed for parameter estimation. To test model fit, a third time frame is needed. Additional time frames might be helpful to increase the accuracy of parameter estimation. As in the Poisson model,

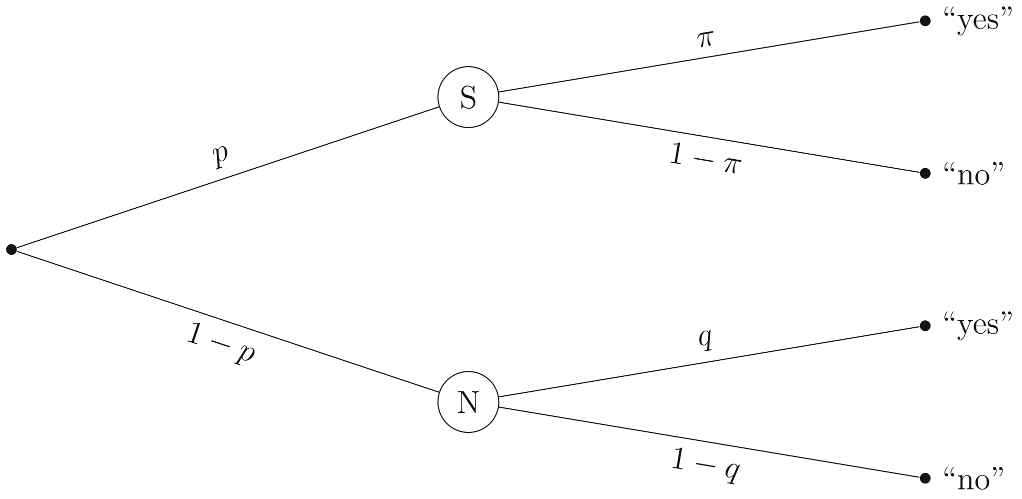


Fig. 3

Probability tree of the unrelated question model. The sample is divided into participants drawing the sensitive question S and those drawing the neutral question N (with the probabilities p and 1–p, respectively). The probability of a yes-answer to the sensitive question is π, for a no-answer it is 1–π. Participants drawing the neutral question answer "yes" or "no" with the probabilities of q and 1–q, respectively

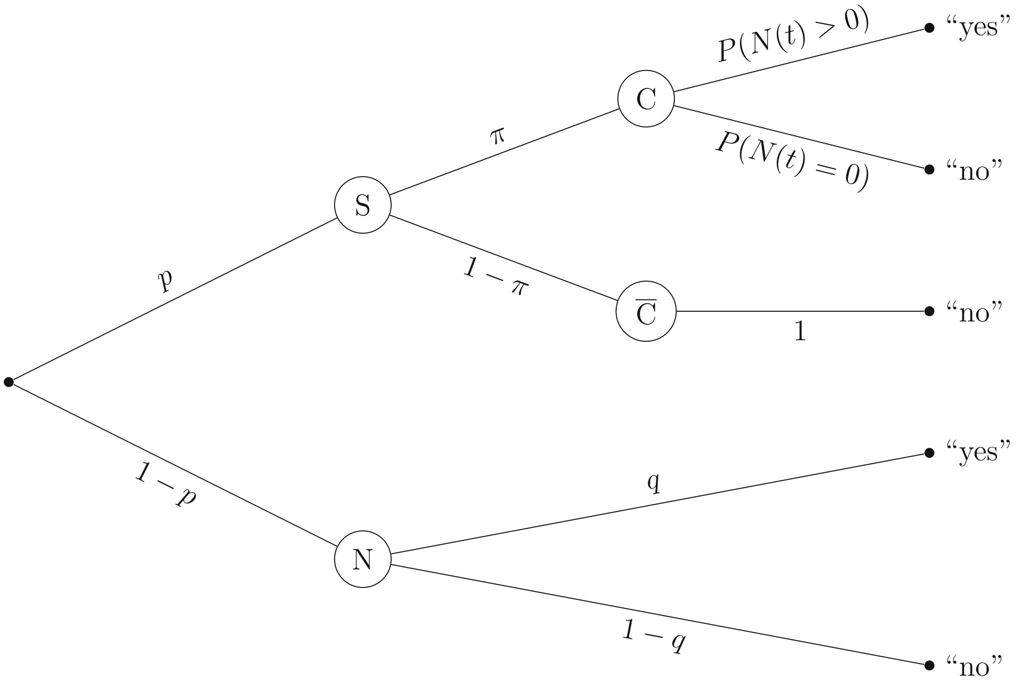


Fig. 4

Probability tree of the Poisson extension of the unrelated question model. The sample is divided into participants drawing the sensitive question S and those drawing the neutral question N (with the probabilities p and 1–p, respectively). The probability of a participant drawing the sensitive question being a carrier C is π, for them being a non-carrier $\overline{C}$ is 1–π. Carriers answer "yes" to the sensitive question with the probability $P(N(t) > 0)$, or "no" with the probability $P(N(t) = 0)$. Non-carriers are assumed to answer "no" in all cases. Participants drawing the neutral question answer "yes" or "no" with the probabilities of q and 1–q, respectively

the maximum likelihood procedure can be used to estimate the parameters (see Supplementary Material).

Differently to the standard Poisson model, the probability P("yes" $|t$) is not equivalent to the prevalence curve, since not every yes-answer in the UQMP is related to the topic of interest. Instead, the probability distribution of P("yes" $|t$) includes the probability of answering "yes" to the neutral question as well. This can clearly be seen in Fig. 5, since the curve does not start at an intercept of 0, but at $0 + (1 - p) \cdot q$ and since the asymptote is not located at π, but at $p \cdot \pi + (1 - p) \cdot q$. Additionally, the slope of the curve is stretched by the parameter p.

Consequently, the prevalence curve must be represented by the conditional probability of answering "yes", given the sensitive question. This conditional probability is calculated by

$$P(\text{"yes"} \mid t, \text{sensitive question}) = \frac{P(\text{"yes"} \mid t) - (1 - p) \cdot q}{p}. \tag{11}$$

Inserting the probability of answering "yes" in the UQM procedure (see Eq. 10) yields a function that is equivalent to Eq. 5.

1.4 The study at hand

In this study, we tested the applicability of the proposed model, the UQMP. To do so, we used the UQMP to estimate the prevalence of drinking and driving, defined as "driving while drunk", in a sample of Germans regularly participating in motorized traffic. Additionally, we applied the standard Poisson model, using a direct question, to measure the same prevalence. Thus, a comparison between the UQMP and the standard Poisson model, using the DQ technique, is enabled. To control whether the UQM method works as intended, we also asked a non-sensitive question in the DQ and UQM format; the prevalence estimates for non-sensitive attributes should not differ between both methods. Finally, we asked some questions regarding the perception

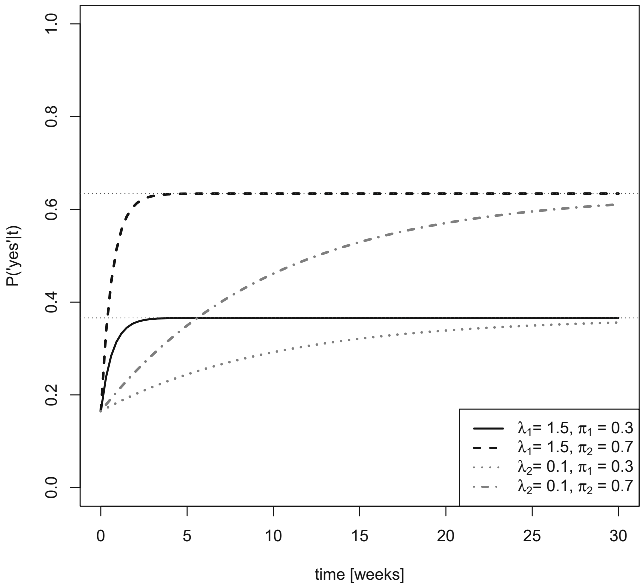


Fig. 5

Examples of the probability distribution of $P(\text{“yes”} \mid t)$ as a function of π and λ with varying values for both parameters. The UQM design parameters are set to $p = 0.67$ and $q = 0.5$, thus the intercept at $t = 0$ is located at $(1 - 0.67) \cdot 0.5 = 0.165$

of the survey, e.g., if the participants felt anonymous during the survey process.

Only some research exists regarding the prevalence of drinking and driving in Germany. While the German police and the Federal Office for Motorized Traffic publish some statistics about traffic violations involving alcohol, those numbers are not indicative of the true prevalence of drinking and driving. This is because not every person gets caught driving under the influence, thus a substantial *dark figure* (i.e., the cases not known by the authorities) of drinking and driving is to be assumed. The likely most valid measurement for this dark figure was provided by Krüger and Vollrath (1998), who measured the prevalence with a *roadside survey*: In cooperation with the police, they pulled drivers over randomly and measured their blood alcohol level. As a result, 1% of the drivers violated the allowed maximum level of *blood alcohol concentration* (BAC), which is 0.05% according to German law. However, it is still possible that Krüger and Vollrath (1998) underestimated the prevalence of drunk drivers, as some drivers may choose to travel on less-monitored roads after alcohol consumption.

Unfortunately, a more up-to-date roadside survey has not been conducted in Germany since. In a more recent study, Goldenbeld et al. (2020) used direct questions in an online survey to measure the prevalence of drinking and driving in multiple European and non-European countries. For Germany, they found that 9% of drivers admitted that they might have violated the legal BAC-level in the last month. In another study, Iberl (2021) compared the UQM and DQ in an online survey to measure the lifetime prevalence of drinking and driving for German university students, finding no difference between the prevalence estimates in both methods. Drinking and driving was defined similarly as in Goldenbeld et al. (2020), as “driving under the influence of alcohol while accepting the possibility of a rule violation” (Iberl, 2021, p. 277), resulting in an estimation of $\pi = 0.44$. This UQM lifetime prevalence estimate of 0.44 was used as a point of orientation in the study at hand. This prevalence is most likely lower in a student sample compared to the general population, as students are younger and less likely to own motorized vehicles (younger people were also less likely to be drunk drivers in the roadside survey by Krüger & Vollrath, 1998). This could indicate that the proportion of trait carriers would be higher than 0.44 in a more representative sample. However, the definition of drinking and driving in Iberl (2021) is much broader than the one used in our study, which is probably the main reason why the estimate of 0.44 is much higher than in other studies. We therefore assumed that the proportion of true trait carriers should be lower than 0.44 in our sample.

As the non-sensitive question for validating the UQM, we used a question about the eye color of the participants, assuming eye color to be a non-sensitive attribute. To be precise, we estimated the prevalence of blue eye color via the DQ and UQM methods.

Our preregistered hypotheses (see https://osf.io/nh6e9) were:[1]

1. The proposed model (UQMP) fits the data well. Thus, it might be suitable for application in prevalence research about sensitive topics.
2. (a) The prevalence of drinking and driving (i.e., π) is higher in the UQMP than in the standard Poisson model based on direct questioning, which may indicate a more accurate estimate.
   (b) The UQM should result in participants in the first group (UQMP) feeling more anonymous compared to participants in the second group (DQ based on Poisson model).
3. The proportion of trait carriers is expected to be lower than 0.44 (the lifetime prevalence for drinking and driving in students in Iberl, 2021).
4. The prevalence estimate of the non-sensitive eye color question does not differ between questioning via the UQM and via a direct question.

[1] As proposed by the anonymous reviewer, we slightly altered the wording and structure of the hypotheses from the preregistration to increase comprehensibility.

Table 1

Distribution of demographics in the sample compared to those of the German population owning a driver's license

		Distribution	
	Demographic	Sample (%)	Population (%)
Gender	Female	42.8	43.1
	Male	56.9	56.9
	Non-binary	0.3	0.0
Age	18–29 years	15.8	16.8
	30–39 years	19.5	20.1
	40–49 years	14.1	14.2
	50–59 years	17.5	16.7
	60 years and older	33.2	31.8

The reference distribution of demographics is based on data by the Kraftfahrt-Bundesamt [Federal Office for Motor Traffic]

2 Method

2.1 Design

The study at hand is built as a 2 (DQ vs. UQM group) × 4 (drinking and driving in the past week/month/six months/ year) between-subjects design. Participants were randomly assigned to one of the eight resulting groups. The questions in the survey were presented in a fixed order for all participants regardless of the group. Quotas regarding age and gender were set in advance. Those were derived from the data of the Kraftfahrt-Bundesamt [Federal Office for Motor Traffic] (2022) and were applied to the aspired sample size set in advance in the preregistration.

2.2 Participants

For our survey, we aimed for a sample representative of regular motorized road users in Germany. To reach this goal, the market research company *Bilendi S.A.* was commissioned to recruit a sample of $N = 3680$ German participants with the same demographic properties as the population of Germans with a driver's license (see Kraftfahrt-Bundesamt [Federal Office for Motor Traffic], 2022).

The sample size rationale for the study was based on simulations, which in turn, were based on parameter values that seemed realistic. For the number of carriers, we assumed a prevalence of $\pi = 0.30$. This assumption was based on the prevalence in Iberl (2021) and the hypothesis that the π estimate in the study at hand would be smaller due to different wording of the questions posed. For the mean rate of drinking and driving we assumed $\lambda = 1$ (i.e., one instance of drunk driving per month[2]) to be a somewhat realistic value. Assuming these values, good accuracy for the maximum-likelihood-estimation of both parameters is achieved in the UQMP with a sample size of 600 participants for four groups and time frames t_i (past week/ month/six months/year; the mean standard deviation for π and λ was 0.021 and 0.244, respectively). For the standard Poisson model, 200 participants per group were sufficient for good estimation accuracy (mean standard deviation of 0.023 for π and 0.231 for λ). In total, the simulations pointed toward a sample size of 3200 participants as adequate. To assure a sufficient sample size after data exclusion, we increased the aspired sample size by 15%, yielding a final goal sample size of 3680 participants.

In total, 5739 potential participants followed the invitation link to the online survey. Participants who did not drive a motor vehicle at least once per week at the time of the study were screened out at the beginning of the survey. 279 participants who failed an implemented attention check were screened out as well (5% of the potential participants). After screen-outs, $N = 3682$ completed surveys remained, fulfilling the aspired sample size. Furthermore, we used the relative speed index (RSI) approach of Leiner (2019b) to identify participants who answered the survey substantially faster than average. The RSI was computed according to Leiner (2019b) and calculated separately for each group, to take possible differences in completion time into account. In total, after applying the described and preregistered exclusion criteria (see Iberl et al., 2022a), a sample of $N = 3529$ participants remained.

Of the 3529 participants, 1512 or 43% stated their gender as female, 2007 or 57% as male and 10 or 0% as non-binary. The mean age in the sample was 48.9 years

[2] In this case and throughout the rest of the paper, the unit of the parameter λ is 1/month.

Table 2

Observed frequencies of responses for each subgroup to the question of drinking and driving

Group	Time frame	*n*	"yes"	"no"
UQM	One week	672	283	389
	One month	670	272	398
	Six months	654	276	378
	One year	655	277	378
DQ	One week	210	20	190
	One month	221	23	198
	Six months	215	17	198
	One year	232	22	210

(SD = 16.3) with a minimum age of 18 and a maximum age of 89. The distribution of gender and age in the sample matches well with the one for Germans with a driver's license according to the Kraftfahrt-Bundesamt [Federal Office for Motor Traffic] (2022), see Table 1. Thus, we believe to have achieved a sample approximately representative of the motorized road users in Germany with respect to age and gender (2022).

2.3 Material and procedure

After preparation of the survey, using the software *SoSciSurvey* (Leiner, 2019a) and preregistration of the study, the recruitment phase started on July 27th, 2022 via *Bilendi S.A.* First, the participants received the link to the online questionnaire from the aforementioned market research company. Upon following this link, they were presented an introductory text, explaining the legal framework of the survey (voluntary participation, guarantee of anonymity and contact information of the responsible party). At the same time, conditions for participation were determined (at least 18 years of age and fluency in the German language). Lastly, it was announced that they will be able to create a personal code with which they would be able to delete their data if they wanted to. The participants created this code on the following page.

On the third page, information about demographics was inquired. At later stages of the sampling phase, some of the preset demographic quotas were already fulfilled (e.g., the aspired number of male participants was complete). In this case, any participant of the same demographic (e.g., any male participant) would be screened out after this page and redirected to another website appointed by the market research company. The demographic questions were followed by the question about traffic participation on the next page ("Do you drive a motor vehicle (e.g., a car, motorbike, motor scooter, etc.) at least once per week?"), screening out any participants who drove more rarely than once a week.

Next, the participants were queried about drinking and driving. At this point, it was explained to the UQM group that a specific questioning method would be used in this survey and that this method would guarantee their complete anonymity. On the next page, they were instructed to think about the birth date of a friend or relative and to remember this birth date for the next page. Then, they were presented the UQM question design:

Is the birthday of the person you thought about between the 1st and 10th day of the respective month? Then please answer question A honestly.

Is the birthday of the person you thought about between the 11th and 31st day of the respective month? Then please answer question B honestly.

Question A: Is the birthday of the person you thought about in the first half of the year, so before the 1st July of a year?

Question B: Did you drive a motorized vehicle (a car, motorcycle, scooter, etc.) in the last week/month/six months/year while being drunk or knowing that you had too much to drink?

So, the participants could be led to the neutral *Question A* or the sensitive *Question B* about drinking and driving, depending on the birthday they thought of. Then, they should answer honestly, regardless of the question they were led to.

The time frame that *Question B* referred to varied, depending on the group of participants. The intro question and *Question A* concerned the birth date the participants were instructed to think about. They were designed so that the probability to be assigned to the sensitive *Question B* was $p \approx 0.67$ and that the probability to answer "yes" to *Question A* was $q \approx 0.5$.

Meanwhile, participants in the DQ group were told that on the next page, there would be a question regarding drinking and driving, followed by an independent extra question.

Table 3

Maximum likelihood estimates, standard errors, and 95% bootstrap confidence intervals for π and λ, and results of G-tests for the UQMP and the standard Poisson model

Group	π			λ				
	Estimate	SE	95% – CI	Estimate	SE	95% – CI	G	p
UQM	0.388	0.015	[0.358; 0.418]	9.810	0.617	[7.761; 10.000]	1.855	0.173
DQ	0.096	0.010	[0.078; 0.116]	8.756	2.012	[3.732; 10.000]	1.041	0.308

The rate of occurrence λ has the dimension $[month]^{-1}$. The point estimates, standard errors, and confidence intervals were calculated using a parametric bootstrap algorithm with 1000 bootstrap samples. All G-tests were carried out with two degrees of freedom ($df = 2$). p values are presented for interpretation of the G-tests

They were also guaranteed anonymity. After continuing, the DQ group was presented the direct question about drinking and driving. This question was identical to *Question B* of the UQM design, also with varying time frames depending on group, but posed directly.

Afterwards, it was announced to the participants in the UQM group that another question using the same method, but tackling another topic, would be asked. However, the attention check followed on the next page. In this attention check, participants were asked which one of six cities was not located in Germany. While five of the named cities were German, *London* was included as the odd one out. Participants who failed to answer the attention check correctly were screened out, as mentioned above.

The attention check was succeeded by the eye color question. The participants in the UQM group were again requested to think about a certain birth date. On the following page, the questions regarding eye color were posed to them in the same way as the question about drinking and driving, but with *Question B* being worded as "Do you have blue eyes?" (obviously without referring to a time frame). The same question was presented directly to participants in the DQ group after they completed the attention check.

Participants who completed the eye color question were confronted with questions about survey impression on the last page of the survey. These questions inquired, using a five-point Likert scale, how anonymous the participants felt during survey completion and how reprehensible they thought drinking and driving was. Subsequently, participants were redirected to a website of *Bilendi S.A.*

Completing the survey took the participants in the final sample 3 min and 38 s on average (SD = 89.58s). Unsurprisingly, participants in the UQM group took longer on average (3 min and 56 s, SD = 89.40s) than participants in the DQ group (2 min and 45 s, SD = 65.65s). Data acquisition ended on August 8th, 2022.

3 Results

All computations were executed with the free software *R* (R Core Team, 2018). See Iberl et al. (2022b) for the complete data and analysis code.

The sample sizes as well as the observed yes- and no-answers to the drinking and driving question are presented in Table 2 for each subgroup. The combined sample sizes are n = 2651 for the four UQM groups and n = 878 for the four DQ groups.

The prevalence π and mean rate λ for drinking and driving were estimated via the maximum likelihood method both for the UQM group, using the UQMP, and the DQ group, using the standard Poisson model (see Supplementary Material). For reliable calculation of standard errors and 95% confidence intervals, a *parametric bootstrapping procedure* with 1000 bootstrap samples was employed (see, e.g., Boos, 2003). Table 3 contains the parameter estimates for the UQM group via UQMP and the DQ group via the Poisson model.

In line with Hypothesis 1, the G-tests are non-significant for both models. As predicted in Hypothesis 2a, the proportion of carriers is considerably higher in the UQMP method. While $\pi = 0.096$ in the DQ group, meaning around 10% of the sample can be described as drunk drivers, the estimate resulting in the UQMP is as high as $\pi = 0.388$ (but, as expected, lower than 0.44, see Hypothesis 3). The λ-estimates are very high in both groups, which indicates a high rate of drinking and driving among the carriers. But, since the upper boundary of the 95% confidence intervals for λ reach the set upper limit for parameter estimation (10), those estimates are to be interpreted cautiously.

The graphics in Fig. 6 show similar resulting prevalence curves for both the UQMP and the standard Poisson model, with the UQMP's prevalence curve having a higher asymptote (as determined by π). The curves rise very steeply, reaching the asymptote already on the first point of measurement. This kind of fast-rising curve is a result of the high λ-values estimated in both models. According to these results, carriers of the "drinking and driving"- attribute show

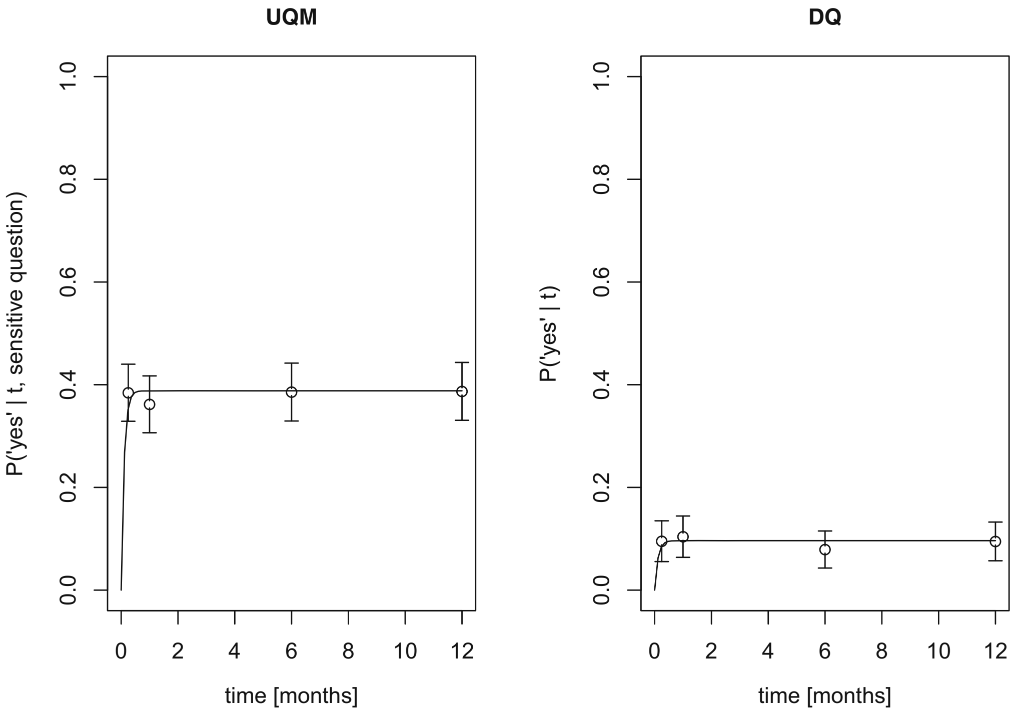


Fig. 6

The prevalence curves for the UQMP and the standard Poisson model resulting from parameter estimation in both methods. The points indicate the prevalence estimates per time frame. The error bars represent 95% confidence intervals

this behavior regularly, since its probability of occurrence does not seem to change over time. With a behavioral pattern like this, the specific value of λ could theoretically be infinite in the Poisson model, and should thus not be interpreted.

Regarding the control question of eye color, 95% confidence intervals were calculated via the standard procedures for the respective type of question (see Eq. 9) for the UQM group, and the standard binomial 95%-CI for the DQ group. In the UQM group ($n = 2651$), a prevalence for blue eyes of 0.517 (95% CI [0.489, 0.546]) was estimated. The prevalence in the DQ group ($n = 878$) is significantly lower with an estimate of 0.355 (95% CI [0.324, 0.387]). Thus, these results contradict Hypothesis 4 regarding the equality of both prevalence measures for eye color.

Most participants, regardless of group, felt that their anonymity was well protected in the survey. On the 5-point Likert scale, the mean score was 4.213 ($SD = 0.918$). Still, the feeling of anonymity differed significantly between groups (*Welch Two Sample t-test*; $t(1478) = 4.410$, $p < 0.001$), with the UQM group showing slightly higher scores $\left(M_{\mathrm{UQM}} = 4.252\right)$ than the DQ group $\left(M_{DQ} = 4.093\right)$. The damnability of drinking and driving was rated highly by both groups ($M_{\mathrm{UQM}} = 4.683$, $M_{DQ} = 4.710$), with no statistically significant differences between mean scores (*Welch Two Sample t-test*; $t(1560) = -0.968$, $p = 0.333$). The results of the questions about the participants' impression of the questionnaire concur with the preregistered Hypothesis 2b.

4 Discussion

In this paper, we presented a novel method, the UQMP, combining the Poisson model (Iberl & Ulrich, 2023) with the unrelated question model by Greenberg et al. (1969).

Through the approach of the Poisson model, unambiguous prevalence estimation and the estimation of the mean rate of a behavior's occurrence are rendered possible. Additionally, the UQM is designed to solve the problem of socially desirable answers to sensitive questions by providing the participants with complete and transparent anonymity. The model was applied to the sensitive topic of drinking and driving in motorized traffic, and compared to the Poisson model, another recently proposed method (Iberl & Ulrich, 2023). For this purpose, a sample representative of German drivers in terms of gender and age was queried via an online survey. Although the model appears to fit the data based on the G-tests, the obtained flat prevalence curves were unexpected based on this model. Thus, Hypothesis 1 can only be conditionally confirmed. Regarding the proportion of trait carriers, we get an estimate as high as 39% for drunk drivers in Germany using the UQMP. With direct questioning, in the standard Poisson model, the amount of carriers is estimated to be lower, as expected (see Hypothesis 2a), with 10% of the participants being identified as carriers. As anticipated, these percentages are lower than the 44% lifetime-prevalence that resulted for drinking and driving in the student survey of Iberl (2021) (see Hypothesis 3). Also, evidence is found for participants in the UQMP group to feel somewhat more protected regarding their anonymity, compared to the DQ group (see Hypothesis 2b). An unexpected result can be found in the neutral question about eye color (Hypothesis 4): While we assumed no difference in estimation for blue eye prevalence between UQM and DQ methods, the results differ significantly.

In the following, we will first interpret the values resulting from the parameter estimations in the Poisson model and the UQMP methods. Then, we discuss the unexpected results in the blue eye color prevalence estimation, proposing some possible explanations and summarizing the results of a follow-up study we conducted to test one of those explanations. Afterwards, we conclude the applicability of the UQMP. We finish the discussion with an assessment of our findings and possible further research regarding the uses of the Poisson model.

4.1 Comparing the Poisson model and the UQMP

Interestingly, the prevalence curves for both the standard Poisson model and the UQMP are similar in shape, rising very steeply and reaching the asymptote already on the first point of measurement (one week). In turn, the λ parameters assume very high values in both models, with 9.810 for the UQM group and 8.756 for the DQ group. However, as mentioned above, both 95% confidence intervals include the preset upper limit for estimation, thus the values can not be interpreted as the amount of times the behavior occurred in the reference time unit (in this case, one month). This is a consequence of all four measurement points, in both groups, yielding the same relative frequency of yes-answers. Thus, the data truly support a straight line for a prevalence curve, instead of an actual curve. With such a result, theoretically, λ could be infinite.

When it comes to behaviors like drinking and driving, one would expect to see a concave prevalence curve, meaning that the proportion of people engaging in the behavior should increase with the length of the time frame. Due to this, the rather flat prevalence curves found in this study are surprising. One possible explanation for the unexpected curves is that many participants may have misread or misunderstood the questions, resulting in equal proportions of yes-answers regardless of the time frame. Some studies (see e.g., Lannoy et al., 2021; Maurage et al., 2020; National Institute of Alcohol Abuse and Alcoholism [NIAAA], 2018) have identified *binge drinkers*, who consume large amounts of alcohol on rare occasions. This also suggests that the prevalence curve for drinking and driving should indeed be concave (especially with past-month-prevalence estimates as high as 26%, see National Institute of Alcohol Abuse and Alcoholism [NIAAA], 2018). However, binge drinkers might not drive during their times of consumption, which would not have influenced our measurement of drinking and driving prevalence. While the validity of the resulting prevalence curves remains unclear, such flat functions can be compatible with the model's assumptions: If drinking and driving occurs very regularly for some individuals, at least once a week, while the rest of the population does not engage in this behavior, flat prevalence curves would be expected in the Poisson model.

According to the UQMP, the carriers account for 39% of the sample, while the π estimate for the standard Poisson model is much lower with around 10%. Thus, our DQ estimation for Germans who drive under the influence is remarkably close to the one of Goldenbeld et al. (2020), who found a 30-days-prevalence of 9%, also by DQ. The much higher prevalence of carriers resulting from the UQMP could theoretically indicate a more valid estimation since some respondents in the DQ group might not have answered truthfully due to social desirability bias. The results in the eye color question, however, clearly point towards problems with overestimation in the UQM group.

4.2 Unexpected results: Blue eye color prevalence

The expectation for estimation of blue eye color prevalence was for both methods to yield the same results (Hypothesis 4). Since this question should not be perceived as sensitive, no social desirability bias should influence the answers, resulting in equal prevalence estimates for the UQM and DQ groups. The value of 52% blue-eyed respondents resulting for the UQM group not only seems high compared to the 36% in the DQ group, but also when looking at the (sparse) corresponding literature. In a 19th-century study of *Virchow*, which has been deemed still relevant by Katsara and Nothnagel (2019), a prevalence of almost 40% for blue eyes in the German population was found. On a German website about "rapid facts", a non-published study is cited to have found a prevalence of 30% for blue eyes in Germany; additionally, the users of the website can report their own eye color, resulting in a prevalence of 31% (kurzwissen.de, 2019). While the latter source is of questionable validity, it also points towards our UQM estimate being too high and towards the DQ estimate as the more valid one.

Regardless of the actual prevalence, the difference between both methods' estimates is unexpected. The most logical explanation seems to be an unknown effect related to the questioning technique used in both groups.

There are multiple possible explanations for how a supposed effect of the questioning technique could have come to pass. First, since we did not randomize the order of the alcohol-related and eye-color-related questions, some kind of order effect could explain the results. Maybe the participants in the UQM group did not pay as much attention to the instructions after they already answered both the question about drinking and driving and the attention check. Or maybe the birth date they thought about while answering the first question influenced the birth date they used for the second question about eye color, distorting the design probabilities of p and q. Either way, if the position of the eye color question caused the suspected inflation in the prevalence estimate of blue eye color, the question regarding drinking and driving could be unaffected by this, since it was asked before. To test this explanation of order effects, we conducted a follow-up a study using the curtailed sampling approach (Reiber, Schnuerch, & Ulrich, 2022; Wetherill, 1975). In this follow-up study, we switched the order of the UQM questions, asking the eye color question before the attention check and the drinking and driving question. If the possible order effect caused the high value for blue eye color prevalence in the UQM group, an estimate in the realm of the value in the DQ group, 36%, should be expected. However, the follow-up study led to an estimate for blue eye prevalence via UQM similar to the one of 52% in the main study, even though the questions' positioning was swapped. So, order effects alone do not seem to have influenced the results of the neutral question, pointing towards different explanations (for a more detailed description of the follow-up study see Iberl, Aljovic, Ulrich, Reiber (2022c) and the Supplementary Material).

A second possible explanation lies in some kind of random responding by the participants (independent of the order of questions). Some respondents might not follow the instructions thoroughly enough (either by unwillingness or due to comprehension issues), in turn answering randomly. This would, of course, influence not only the neutral question of eye color, but also the results for the prevalence of drinking and driving in the UQM group. To test this explanation post-hoc, we calculated, assuming the true prevalence of blue eye color to be 36% like in the DQ group, how many participants would have to answer "yes" randomly (with the probability of 0.5) in order to produce the result of 52%. Surprisingly, more than 100% of randomly responding participants would be needed for this result to occur. So, a truly random pattern of responding can not be the only reason for the unexpected results. The explanation gets more likely if one assumes a non-equal distribution for the probabilities of "randomly" answering "yes" or "no". Potentially, the yes-answer is chosen more often than the no-answer in random responding (because it is more appealing for some reason or just because it is read first). If we assume a probability of random yes-answers of 0.75, about one-third of the participants would have to respond randomly to get our result for blue eye prevalence. This seems more plausible than random responding with equal probabilities for yes- and no-answers. But, even if we assume an uneven probability for both answers, random responding by itself seems to be unlikely as an explanation for the results. A recent study by Meisters et al. (2022) supports this claim, finding that while random responding exists in the researched RRM, it only has a minor influence on the resulting prevalence. However, there also seems to be some contrary evidence pointing towards random responding as a substantial factor in RRMs (Walzenbach & Hinz, 2019).

A third explanation could lie in the nature of the surveyed sample. Since the sample consisted of most likely highly experienced participants in regards to online surveys, it is reasonable to assume that privacy concerns were less common compared to the general population. This is supported by the result that the participants in the DQ group felt almost as anonymous as those in the UQM group. As some studies show, RRMs work best when the question is perceived as sensitive, so when a social desirability bias is to be expected when using DQ instead (e.g., Lensvelt-Mulders et al., 2005; Tourangeau & Yan, 2007). With highly survey-experienced participants who feel very anonymous, a substantial social desirability bias might be less likely. So, for this specific sample, the UQM might only be perceived as confusing and annoying, instead of a protective question design, leading to the unexpected results through random responding or even noncompliance. Generally, it could be a problem of RRM application in online surveys that it is hard to control whether the participants comprehend the instructions of the method or whether they understand that the RRM provides a high level of anonymity. Unfortunately, it is not yet well understood what role comprehension plays for the validity of RRMs (e.g., Bullek et al., 2017; Hoffmann et al., 2017; Höglinger & Jann, 2018; Meisters et al., 2020). Thus, it may be crucial to conduct RRT surveys in person, with interviewers explaining the procedure to respondents before running the UQM survey (e.g., Striegel et al., 2010).

Multiple other explanations are imaginable as well. For example, the eye color question might not be as neutral as supposed, or it might have been difficult to answer for participants with ambiguous eye colors (like "blue-green" or "grey-blue"), leading to unforeseen response behavior. However, none of these possible effects would seem strong enough to explain the high prevalence of 52% on its own, but some of them might contribute to an inflation of the estimate. So, although we can not identify a single explanation, we might have identified some possible effects that could have caused the high estimate in the UQM group. Regardless, the UQM method seems to be the cause of this inflated estimate. While we can not exclude the reasons to be of a nature regarding the content of the eye color question, which would not influence the estimation for drinking and driving by UQM, we can also not confidently assume the UQM to have worked as intended. Consequently, the DQ estimate seems to be the more accurate result, and the UQMP estimate should be viewed with caution. This, however, is due to the UQM—and not due to the Poisson model used in combination.

In future research, one could test the combination of the Poisson model with different RRMs, such as the crosswise model (CWM, Yu et al., 2008). These combinations would be easy to realize and might yield more plausible results.

5 Conclusion

Albeit we can not rule out problems of the UQM method in view of our findings, the Poisson model seems to work as intended. While the resulting prevalence curves are unexpected in shape, this is caused by the answers of the participants, showing no variation of the proportion of yes-answers between time-frames. Also, the π estimate for drunk drivers resulting by DQ is similar to the results of another study using DQ to measure prevalence of drinking and driving in Germany (Goldenbeld et al., 2020). Additionally, even though the λ estimate seems off at first glance, it behaves as expected given there is no variance between the proportion of yes-answers in the four points of measurement. A "prevalence line" of some sort between the four points of measurement, as resulted in this study, does not contradict the model's core assumptions: If there exists only a rather small population of carriers that is showing the behavior very frequently, a line is to be expected. To be precise, if the behavior's rate of occurrence is as high or higher than $1/t_1$, with t_1 being the time frame of the first point of measurement (here: one week), prevalence curves like those in Fig. 6 are likely. This applies to both the Poisson model and the UQMP. Unfortunately, in case of drinking and driving, a shorter reference time frame than one week would probably not have been suitable for usage in the questionnaire. This is because driving under the influence is supposed to be more frequent during the weekend, especially weekend nights, according to some studies (e.g., Krüger & Vollrath, 1998; Vanlaar, 2005). Thus, the assumption of drinking and driving as a Poisson-distributed variable might be invalid for time frames smaller than one week. Regardless of the assumed overestimation due to unforeseen bias caused by the UQM method, the form of the UQM prevalence curve is similar to the one in the DQ group. In these cases, the estimated value for the λ parameter should not be interpreted. Instead, the rate of occurrence of the researched behavior can be assumed to be at least 1 referring to t_1, that is, once a week.

To sum up the results of our study regarding the problem of alcohol in motorized traffic, we found that at least 10% of drivers in Germany are frequently, at least once a week, driving under the influence of alcohol. The other (at maximum) 90%, on the other hand, seem to be non-carriers, who essentially never engage in drinking and driving. While the UQMP estimation does not seem reliable, we can not rule out an underestimation of drinking and driving by the standard Poisson model using the DQ method, so the proportion of 10% for drunk drivers can be seen as a lower border for the true amount of carriers. A substantial part of those carriers can be assumed to be participants with problematic or even pathological alcohol consumption. This claim is supported by an older study by Selzer and Barton (1977), which showed that about two-thirds of the drunk drivers in their sample were pathological drinkers. Also, this hypothesis seems valid due to the high amount of problematic drinkers in Germany: According to Atzendorf et al. (2019), 18% of a sample of 9267 Germans between the age of 18 and 64 years reported the use of alcohol in hazardous quantities for the time frame of 30 days before the survey. It stands to reason to assume that many people with such high alcohol usage still rely heavily on motorized traffic in their everyday lives, as this is generally by far the most used form of transportation in Germany (Bundesministerium für Digitales und Verkehr [Federal Ministry for Digital and Transport], 2021). The fact that the majority of the German population uses a car or another motorized vehicle to commute to their workplace, school or university seems particularly relevant in this context (Bundesministerium für Digitales und Verkehr [Federal Ministry for Digital and Transport], 2021).

In conclusion, the Poisson model used in our study seems suited for practical application, even if the shape of the resulting prevalence curve contradicts this statement at first glance. Still, the Poisson model should be tested and compared with traditional methods in future studies to determine when it is suited best. Furthermore, we showed that the Poisson model can be combined with indirect questioning techniques such as the UQM. There are still open questions regarding the validity of the UQM and RRMs in general, as we have to assume an unexpected inflation of false-positive yes-answers due to the results of our control question about eye color prevalence. Thus, more research is needed to understand the validity of RRMs further and to identify scenarios of when RRMs are (not) to be preferred over DQ methodology. Plus, while the sample size needed for a satisfactory level of statistical power is already high in RRMs compared to DQ, the additional participants needed for implementing a Poisson extension to an RRM is considerable. In further studies, it might be reasonable to test Poisson model extensions and their applicability for different RRMs, like the cheater detection model (CDM, Clark & Desharnais, 1998) or the crosswise model (CWM, Yu et al., 2008). Since they enable time-independent prevalence estimation for objectively anonymous survey procedures, Poisson model extensions for RRMs might be worth the effort.

Acknowledgements We thank an anonymous reviewer for many constructive and detailed comments.

This work was not presented elsewhere before this publication. The data presented in this work was used in the unpublished master thesis of one of the authors (Aljovic, 2022).

The paper was preregistered via OSF (see https://osf.io/nh6e9). The complete data and analysis code was uploaded at https://osf.io/5pkm4/.

References

Aljovic, A. (2022). Zeitunabhängige Prävalenzschätzung von Alkohol am Steuer: Eine Einführung des Unrelated Question Models mit Poisson-Erweiterung [Time-independent prevalence estimation of drinking and driving: An introduction of the Poisson extension to the Unrelated Question Model]. Unpublished master thesis.

Aarts, H., & Dijksterhuis, A. (1999). How often did I do it? Experienced ease of retrieval and frequency estimates of past behavior. *Acta Psychologica*, *103*(1–2), 77–89. https://doi.org/10.1016/S0001-6918(99)00035-9.

Andrie, E.K., Tzavara, C.K., Tzavela, E., Richardson, C., Greydanus, D., Tsolia, M., & Tsitsika, A.K. (2019). Gambling involvement and problem gambling correlates among European adolescents: Results from the European Network for Addictive Behavior study. *Social Psychiatry and Psychiatric Epidemiology*, *54*(11), 1429–1441. https://doi.org/10.1007/s00127-019-01706-w.

Atzendorf, J., Rauschert, C., Seitz, N.-N., Lochbühler, K., & Kraus, L. (2019). The use of alcohol, tobacco, illegal drugs and medicines: An estimate of consumption and substance-related disorders in Germany. *Deutsches Ärzteblatt International*, *116*(35–36), 577–584. https://doi.org/10.3238/arztebl.2019.0577.

Beck, F., Léger, D., Fressard, L., Peretti-Watel, P., Verger, P., & Group, C. (2021). Covid-19 health crisis and lockdown associated with high level of sleep complaints and hypnotic uptake at the population level. *Journal of Sleep Research*, *30*(1), e13119. https://doi.org/10.1111/jsr.13119.

Birkel, C., Church, D., Erdmann, A., Hager, A., & Leitgöb-Guzy, N. (2022). Sicherheit und Kriminalität in Deutschland – SKiD 2020: Bundesweite Kernbefunde des Viktimisierungssurvey des Bundeskriminalamts und der Polizeien der Länder. https://www.bka.de/DE/UnsereAufgaben/Forschung/ForschungsprojekteUndErgebnisse/Dunkelfeldforschung/SKiD/Ergebnisse/Ergebnisse_node.html [Safety and crime in Germany—SKiD 2020: Nationwide findings of the Victimization Survey by the German Federal Office of Criminal Investigation and by the Police of the federal states].

Boos, D.D. (2003). Introduction to the bootstrap world. *Statistical Science*, *18*(2), 168–174. https://doi.org/10.1214/ss/1063994971.

Boruch, R.F. (1971). Assuring confidentiality of responses in social research: a note on strategies. *The American Sociologist*, *27701807*, 308–311.

Bousfield, W.A., & Sedgewick, C.H.W. (1944). An analysis of sequences of restricted associative responses. *The Journal of General Psychology*, *30*(2), 149–165. https://doi.org/10.1080/00221309.1944.10544467.

Bullek, B., Garboski, S., Mir, D.J., & Peck, E.M. (2017). *Towards understanding differential privacy: When do people trust randomized response technique?* Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. (pp. 3833–3837). https://doi.org/10.1145/3025453.3025698.

Bundesministerium für Digitales und Verkehr [Federal Ministry for Digital and Transport] (2021). Verkehr in Zahlen 2021/2022 [Traffic in numbers 2021/2022]. https://bmdv.bund.de/SharedDocs/DE/Publikationen/G/verkehr-in-zahlen-2021-2022-pdf.pdf?__blob=publicationFile

Burr, M.L., Butland, B., King, S., & Vaughan-Williams, E. (1989). Changes in asthma prevalence: two surveys 15 years apart. *Archives of Disease in Childhood*, *64*(10), 1452–1456. https://doi.org/10.1136/adc.64.10.1452.

Clark, S.J., & Desharnais, R.A. (1998). Honest answers to embarrassing questions: detecting cheating in the randomized response model. *Psychological Methods*, *3*(2), 160–168. https://doi.org/10.1037/1082-989X.3.2.160.

Cullen, K.A., Ambrose, B.K., Gentzke, A.S., Apelberg, B.J., Jamal, A., & King, B.A. (2018). Notes from the field: Use of electronic cigarettes and any tobacco product among middle and high school students—United States, 2011–2018. *Morbidity and Mortality Weekly Report*, *67*(45), 1276–1277. https://doi.org/10.15585/mmwr.mm6745a5.

de kurzwissen (2019). Augenfarben Häufigkeit in Deutschland & weltweit [Eye colour prevalence in Germany & worldwide]. kurzwissen.de. https://kurzwissen.de/augenfarben-haeufigkeit/

Dietz, P., Iberl, B., Schuett, E., van Poppel, M., Ulrich, R., & Sattler, M.C. (2018). Prevalence estimates for pharmacological neuroenhancement in Austrian university students: Its relation to health-related risk attitude and the framing effect of caffeine tablets. *Front. Pharmacol.*, *9*(494), 1–9. https://doi.org/10.3389/fphar.2018.00494.

Echterhoff, G., & Hirst, W. (2006). Thinking about memories for everyday and shocking events: Do people use ease-of-retrieval cues in memory judgments? *Memory & Cognition*, *34*(4), 763–775. https://doi.org/10.3758/BF03193424.

Ferrante, T., Castellini, P., Abrignani, G., Latte, L., Russo, M., Camarda, C., Veronesi, L., Pasquarella, C., Manzoni, G.C., & Torelli, P. (2012). The pace study: Past-year prevalence of migraine in parma's adult general population. *Cephalalgia*, *32*(5), 358–365. https://doi.org/10.1177/0333102411434811.

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150.

Goldenbeld, C., Torfs, K., Vlakveld, W., & Houwing, S. (2020). Impaired driving due to alcohol or drugs: International differences and determinants based on E-Survey of Road Users' Attitudes first-wave results in 32 countries. *IATSS Research*, *44*(3), 188–196. https://doi.org/10.1016/j.iatssr.2020.07.005.

Greenberg, B.G., Abul-Ela, A.-L.A., Simmons, W.R., & Horvitz, D.G. (1969). The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, *64*(326), 520–539.

Greenberg, B.G., Kuebler Jr, R.R., Abernathy, J.R., & Horvitz, D.G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, *66*(334), 243–250. https://doi.org/10.2307/2283916.

Han, B., Compton, W.M., Gfroerer, J., & McKeon, R. (2015). Prevalence and correlates of past 12-month suicide attempt among adults with past-year suicidal ideation in the united states. *The Journal of Clinical Psychiatry*, *76*(3), 15414. https://doi.org/10.4088/JCP.14m09287.

Himmelfarb, S., & Edgell, S.E. (1980). Additive constants model: A randomized response technique for eliminating evasiveness to quantitative response questions. *Psychological Bulletin*, *87*(3), 525. https://doi.org/10.1037/0033-2909.87.3.525.

Hoffmann, A., Waubert de Puiseau, B., Schmidt, A.F., & Musch, J. (2017). On the comprehensibility and perceived privacy protection of indirect questioning techniques. *Behavior Research Methods*, *49*, 1470–1483. https://doi.org/10.3758/s13428-016-0804-3.

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *PLOS ONE*, *13*(8), e201770. https://doi.org/10.1371/journal.pone.0201770.

Höglinger, M., Jann, B., & Diekmann, A. (2016). Sensitive questions in online surveys: an experimental evaluation of different implementations of the randomized response technique and the crosswise model. *Survey Research Methods*, *10*(3), 171–187. https://doi.org/10.18148/srm/2016.v10i3.6703.

Huang, K.-C., Lan, C.-H., & Kuo, M.-P. (2006). Estimation of sensitive quantitative characteristics in randomized response sampling. *Journal of Statistics and Management Systems*, *9*(1), 27–35. https://doi.org/10.1080/09720510.2006.10701191.

Iberl, B. (2021). Ein, zwei Bier und ab ans Lenkrad? – Prävalenzschätzung von Alkohol am Steuer durch das Unrelated Question Model [One or two drinks before going for a ride?—Prevalence estimation of driving under the influence via the unrelated question model]. *Kriminologie – Das Online-Journal [Criminology-The Online Journal]*. https://doi.org/10.18716/ojs/krimoj/2021.3.5.

Iberl, B., & Ulrich, R. (2023). On estimating the frequency of a target behavior from time-constrained yes/no survey questions: a parametric approach based on the Poisson process. *Psychological Methods*. https://doi.org/10.1037/met0000588.

Iberl, B., Aljovic, A., Ulrich, R., & Reiber, F. (2022a). *Application of a Poisson extension of the unrelated question model to drinking and driving [OSF preregistration]*. https://doi.org/10.17605/OSF.IO/NH6E9.

Iberl, B., Aljovic, A., Ulrich, R., & Reiber, F. (2022b). Application of a Poisson extension of the Unrelated Question Model to drinking and driving [OSF Project with data and code files]. https://osf.io/5pkm4/

Iberl, B., Aljovic, A., Ulrich, R., & Reiber, F. (2022c). *Follow-up study to "application of a Poisson extension of the unrelated question model to drinking and driving"—testing for possible order effects [OSF preregistration]*. https://doi.org/10.17605/OSF.IO/ZV63D.

Isolauri, J., & Laippala, P. (1995). Prevalence of symptoms suggestive of gastroesophageal reflux disease in an adult population. *Annals of Medicine*, *27*(1), 67–70. https://doi.org/10.3109/07853899509031939.

Janssen, J., Müller, P. A., & Greifeneder, R. (2011). Cognitive processes in procedural justice judgments: the role of ease-of-retrieval, uncertainty, and experience. *Journal of Organizational Behavior*, *32*(5), 726–750. https://doi.org/10.1002/job.700.

Katsara, M.-A., & Nothnagel, M. (2019). True colors: a literature review on the spatial distribution of eye and hair pigmentation. *Forensic Science International: Genetics*, *39*, 109–118. https://doi.org/10.1016/j.fsigen.2019.01.001.

Kraftfahrt-Bundesamt [Federal Office for Motor Traffic] (2022). Fahrerlaubnisbestand im ZFER (Zentrales Fahrerlaubnis-Register) 2022 [Registry of driver's licenses in the CDLR (Central Driver's License Registry) 2022]. https://www.kba.de/DE/Statistik/Kraftfahrer/Fahrerlaubnisse/Fahrerlaubnisbestand/fahrerlaubnisbestand_node.html

Krüger, H.-P., & Vollrath, M. (1998). Fahren unter Alkohol in Deutschland: Die Ergebnisse des Deutschen Roadside Surveys [Driving under the influence in Germany: The results of the German roadside survey]. In H.-P. Krüger (Ed.), *Fahren unter Alkohol in Deutschland [Driving under the influence in Germany]* (pp. 33–57). Gustav Fischer.

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.

Kumar, A. (2022). Estimation of means of two quantitative sensitive variables using randomized response. *Computational Statistics and Applications*. https://doi.org/10.5772/intechopen.101269.

Lannoy, S., Duka, T., Carbia, C., Billieux, J., Fontesse, S., Dormal, V., Gierski, F., López-Caneda, E., Sullivan, E. V., & Maurage, P. (2021). Emotional processes in binge drinking: a systematic review and perspective. *Clinical Psychology Review*, *84*, 101971. https://doi.org/10.1016/j.cpr.2021.101971.

Leiner, D. J. (2019a). SoSci Survey (Version 3.1.06) [Computer software]. https://www.soscisurvey.de

Leiner, D. J. (2019b). Too fast, too straight, too weird: nonreactive indicators for meaningless data in internet surveys. *Survey Research Methods*, *13*(3), 229–248. https://doi.org/10.18148/srm/2019.v13i3.7403.

Lensvelt-Mulders, G. J., Hox, J. J., Van der Heijden, P. G., & Maas, C. J. (2005). meta-analysis of randomized response research: thirty-five years of validation. *Sociological Methods & Research*, *33*(3), 319–348. https://doi.org/10.1177/0049124104268664.

Linton, S. J., Hellsing, A.-L., & Halldén, K. (1998). A population-based study of spinal pain among 35-45-year-old individuals: prevalence, sick leave, and health care use. *Spine*, *23*(13), 1457–1463.

Liu, P., & Chow, L. (1976). A new discrete quantitative randomized response model. *ACM SIGSIM Simulation Digest*, *7*(3), 30–31. https://doi.org/10.1145/1102746.1102750.

Maurage, P., Lannoy, S., Mange, J., Grynberg, D., Beaunieux, H., Banovic, I., Gierski, F., & Naassila, M. (2020). What we talk about when we talk about binge drinking: Towards an integrated conceptualization and evaluation. *Alcohol and Alcoholism*, *55*(5), 468–479. https://doi.org/10.1093/alcalc/agaa041.

McCabe, S. E., Cranford, J. A., & Boyd, C. J. (2006). The relationship between past-year drinking behaviors and nonmedical use of prescription drugs: Prevalence of co-occurrence in a national sample. *Drug and Alcohol Dependence*, *84*(3), 281–288. https://doi.org/10.1016/j.drugalcdep.2006.03.006.

McKetin, R., McLaren, J., Lubman, D.I., & Hides, L. (2006). The prevalence of psychotic symptoms among methamphetamine users. *Addiction*, *101*(10), 1473–1478. https://doi.org/10.1111/j.1360-0443.2006.01496.x.

Meisters, J., Hoffmann, A., & Musch, J. (2020). Can detailed instructions and comprehension checks increase the validity of crosswise model estimates? *PLOS ONE*, *15*(6), e235403. https://doi.org/10.1371/journal.pone.0235403.

Meisters, J., Hoffmann, A., & Musch, J. (2022). More than random responding: empirical evidence for the validity of the (extended) crosswise model. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01819-2.

Miller, C., Ettridge, K., Wakefield, M., Pettigrew, S., Coveney, J., Roder, D., Durkin, S., Wittert, G., Martin, J., & Dono, J. (2020). Consumption of sugar-sweetened beverages, juice, artificially-sweetened soda and bottled water: An Australian population study. *Nutrients*, *12*(3), 817. https://doi.org/10.3390/nu12030817.

Molinaro, S., Benedetti, E., Scalese, M., Bastiani, L., Fortunato, L., Cerrai, S., Canale, N., Chomynova, P., Elekes, Z., Feijão, F., et al. (2018). Prevalence of youth gambling and potential influence of substance use and other risk factors throughout 33 European countries: first results from the 2015 ESPAD study. *Addiction*, *113*(10), 1862–1873. https://doi.org/10.1111/add.14275.

National Institute of Alcohol Abuse and Alcoholism [NIAAA] (2018). Alcohol facts and statistics. Report. https://www.niaaa.nih.gov/sites/default/files/AlcoholFactsAndStats.pdf

Nederhof, A.J. (1985). Methods of coping with social desirability bias: a review. *European Journal of Social Psychology*, *15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303.

R Core Team (2018). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/

Reiber, F., Pope, H., & Ulrich, R. (2020). Cheater detection using the unrelated question model. *Sociological Methods & Research*. https://doi.org/10.1177/0049124120914919.

Reiber, F., Bryce, D., & Ulrich, R. (2022). Self-protecting responses in randomized response designs: a survey on intimate partner violence during the coronavirus disease 2019 pandemic. *Sociological Methods & Resarch*. https://doi.org/10.1177/00491241211043138.

Reiber, F., Schnuerch, M., & Ulrich, R. (2022). Improving the efficiency of surveys with randomized response models: a sequential approach based on curtailed sampling. *Psychological Methods*, *27*(2), 198–211. https://doi.org/10.1037/met0000353.

Şaşmaz, T., Öner, S., Kurt, A.Ö., Yapıcı, G., Yazıcı, A.E., Buğdaycı, R., & Şiş, M. (2014). Prevalence and risk factors of internet addiction in high school students. *The European Journal of Public Health*, *24*(1), 15–20. https://doi.org/10.1093/eurpub/ckt051.

Sawyer, A.N., Smith, E.R., & Benotsch, E.G. (2018). Dating application use and sexual risk behavior among young adults. *Sexuality Research and Social Policy*, *15*(2), 183–191. https://doi.org/10.1007/s13178-017-0297-6.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: another look at the availability heuristic. *Journal of Personality and Social psychology*, *61*(2), 195. https://doi.org/10.1037/0022-3514.61.2.195.

Seitz, N.-N., Rauschert, C., Atzendorf, J., & Kraus, L. (2020). *IFT-Berichte Bd. 190: Berlin, Hessen, Nordrhein-Westfalen, Sachsen und Thüringen. Ergebnisse des Epidemiologischen Suchtsurvey 2018*. München: Institut für Therapieforschung. [IFT-Reports Vol. 190: Substance use and substance use disorders in Berlin, Hesse, North Rhine-Westphalia, Saxony and Thuringia. Results of the 2018 Epidemiological Survey of Substance Abuse]

Selzer, M.L., & Barton, E. (1977). The drunken driver: a psychosocial study. *Drug and Alcohol Dependence*, *2*(4), 239–253. https://doi.org/10.1016/0376-8716(77)90002-3.

Soga, M., Evans, M.J., Tsuchiya, K., & Fukano, Y. (2021). A room with a green view: the importance of nearby nature for mental health during the COVID-19 pandemic. *Ecological Applications*, *31*(2), e2248. https://doi.org/10.1002/eap.2248.

Striegel, H., Ulrich, R., & Simon, P. (2010). Randomized response estimates for doping and illicit drug use in elite athletes. *Drug and Alcohol Dependence*, *106*(2–3), 230–232. https://doi.org/10.1016/j.drugalcdep.2009.07.026.

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.

Ulrich, R., Pope, H.G., Cléret, L., Petróczi, A., Nepusz, T., Schaffer, J., Kanayama, G., Comstock, R.D., & Simon, P. (2018). Doping in two elite athletics competitions assessed by randomized-response surveys. *Sports Medicine*, *48*(1), 211–219. https://doi.org/10.1007/s40279-017-0765-4.

Vanlaar, W. (2005). Drink driving in Belgium: results from the third and improved roadside survey. *Accident Analysis & Prevention*, *37*(3), 391–397. https://doi.org/10.1016/j.aap.2004.12.001.

Virudachalam, S., Long, J.A., Harhay, M.O., Polsky, D.E., & Feudtner, C. (2014). Prevalence and patterns of cooking dinner at home in the USA: National Health and Nutrition Examination Survey (NHANES) 2007–2008. *Public Health Nutrition*, *17*(5), 1022–1030. https://doi.org/10.1017/S1368980013002589.

Walzenbach, S., & Hinz, T. (2019). Pouring water into wine: revisiting the advantages of the crosswise model for asking sensitive questions. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2019-00002.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63–69.

Wetherill, G.B. (1975). *Sequential methods in statistics* (2nd edn.). Chapman and Hall.

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, *67*(3), 251–263. https://doi.org/10.1007/s00184-007-0131-x.

Appendix A

Maximum Likelihood Estimation

The likelihood function for both the standard Poisson model and the Poisson extension to the unrelated question model (UQMP) is

$$L(\pi, \lambda) = \prod_{j=1}^{m} \prod_{i=1}^{n_j} \Big[P(\text{“yes”}\,|t_j)^{a_j} \cdot P(\text{“no”}\,|t_j)^{b_j} \Big] \tag{A1}$$

with the number of groups m, the sample size per group n_j, the time frame per group t_j that the question refers to, the observed number of yes-answers per group a_j and the observed number of no-answers per group b_j. For the two models, $P(\text{“yes”}\,|t_j)$ is computed via the Equations 5 or 10, for the standard Poisson model or the UQMP, respectively. $P(\text{“no”}\,|t_j)$ can be computed per $1 - P(\text{“yes”}\,|t_j)$. Taking the log of Equation A1 gives

$$\log L(\pi, \lambda) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \Big[a_j \cdot \log[P(\text{“yes”}\,|t_j)] + b_j \cdot \log[P(\text{“no”}\,|t_j)] \Big]. \tag{A2}$$

Maximizing equation A2 by a numerical search routine yields the maximum likelihood estimators for the parameters π and λ.

Appendix B

Statistical analysis with R

In this appendix, we give an example for the statistical analysis of our data with the proposed UQMP. The full analysis script and data set can be downloaded at https://osf.io/5pkm4/.

```
## options
# setting RNG seed to a fixed value, so the bootstrap procedure
# yields the same results in every run
set.seed(123)


## design parameters and observed variables
# probability p of getting sensitive question
p <- 245.25/365.25
# probability q of a yes-answer to the neutral question
q <- 181.25/365.25

# values for parameter estimation
lim         <- 1e-10            # lower limit for lambda
up_lim_lam <- 10                # upper limit for lambda

# time frames t_j for all groups j = 1, 2, 3, 4
t0 <- c(.25, 1, 6, 12)

# number of bootstrap samples
nb <- 1000

# group sizes N_t
```

```
N.t.uqm <- c(672, 670, 654, 655)
# yes-answers a_t
a.uqm   <- c(283, 272, 276, 277)
# no-answers b_t
b.uqm   <- c(389, 398, 378, 378)

## functions
# function of the UQMP
pc.uqm  <- function(t,p,q,PI,lam){p*PI*(1-exp(-lam*t))+(1-p)*q}

# log-likelihood-function
MLE.uqm <- function(par, a, b){
        PI      <- par[1]
        lam     <- par[2]
        pyes    <- pc.uqm(t0,p,q,PI,lam)
        lLu     <- a*log(pyes) + b*log(1-pyes)
        MLE.uqm <- -sum(lLu)
}

# G function for testing model fit
Gf.uqm <- function(par, a, b){
        PI      <- par[1]
        lam     <- par[2]
        N.t     <- a + b
        E.t.yes <- pc.uqm(t0,p,q,PI,lam)*N.t
        E.t.no  <- N.t - E.t.yes
        G.uqm   <- 2*sum(a*log(a/E.t.yes) +
                         b*log(b/E.t.no))
        return(G.uqm)
}

## parameter estimation via bootstrap sampling
# "vectorizing" observed yes- and no-answers for every group
obs.t1.uqm <- c(rep(1, a.uqm[1]), rep(0, b.uqm[1]))
obs.t2.uqm <- c(rep(1, a.uqm[2]), rep(0, b.uqm[2]))
obs.t3.uqm <- c(rep(1, a.uqm[3]), rep(0, b.uqm[3]))
obs.t4.uqm <- c(rep(1, a.uqm[4]), rep(0, b.uqm[4]))

# bootstrap sampling
PI.b.uqm    <- numeric(nb)
lam.b.uqm   <- numeric(nb)

for(i in 1:nb){
  # resampling a and b from observed data
  a.b.uqm <- c(sum(sample(x = obs.t1.uqm, size = N.t.uqm[1],
                    replace = TRUE)),
               sum(sample(x = obs.t2.uqm, size = N.t.uqm[2],
                    replace = TRUE)),
               sum(sample(x = obs.t3.uqm, size = N.t.uqm[3],
                    replace = TRUE)),
               sum(sample(x = obs.t4.uqm, size = N.t.uqm[4],
                    replace = TRUE)))
  b.b.uqm <- N.t.uqm - a.b.uqm
  # maximum likelihood estimation of the redrawn sample
  ML.uqm <- optim(par = c(0.5, 0.8), a = a.b.uqm, b = b.b.uqm,
                  fn = MLE.uqm, method = 'L-BFGS-B',
                  lower = c(lim, lim),
                  upper = c(1-lim, up_lim_lam))

  # extracting pi and lambda estimates
  PI.b.uqm[i]  <- ML.uqm$par[1]
  lam.b.uqm[i] <- ML.uqm$par[2]
}

# extracting parameter estimators, SEs and 95%-CIs
# point estimate for pi
PI.m.uqm   <- mean(PI.b.uqm)
# standard error for pi
PI.se.uqm  <- sd(PI.b.uqm)
# 95%-CI for pi
PI.ci.uqm  <- quantile(PI.b.uqm, c(.025, .975))

# point est. for lambda
lam.m.uqm  <- mean(lam.b.uqm)
# se for lambda
lam.se.uqm <- sd(lam.b.uqm)
# 95%-CI for lambda
lam.ci.uqm <- quantile(lam.b.uqm, c(.025, .975))

# G-test
```

```
Gest.uqm    <- optim(c(0.5,0.8), a = a.uqm, b = b.uqm,
                     fn = Gf.uqm,
                     method='L-BFGS-B',
                     lower = c(lim, lim),
                     upper = c(1-lim, up_lim_lam))

# extracting G-value
Gval.uqm <- Gest.uqm$value
# computing p-value for G-test decision
pval.uqm <- pchisq(Gest.uqm$value, df = 1, lower.tail = FALSE)
```

Appendix C

Follow-up study: Estimating blue eye color prevalence with a UQM curtailed sampling approach

Introduction

To investigate a possible explanation for the unforeseen results of the main study, namely the unexpected difference between prevalence estimates of blue eye color via DQ and UQM (see Hypothesis 4 of the main study), a follow-up study was preregistered (see Iberl et al., 2022c) and conducted. To be precise, we tested whether the unexpected results could have emerged due to order effects of the posed questions. The basic idea of the follow-up study was to collect further data from a sample confronted with the UQM method, but while exchanging the order of the drinking and driving and eye color questions. If an order effect was responsible for the unforeseen results, swapping the order of the questions should lead to a prevalence estimate of blue eye color via UQM that is similar to the one generated by the DQ method in the main study.

Because of this rather simple premise and to reduce sample sizes, we opted for a *sequential sampling application* for the UQM as proposed by Reiber, Schnuerch, and Ulrich (2022). Contrary to classical statistical methods, the basic idea of sequential testing is to stop the sampling process as soon as sufficient information is generated to align with a preset hypothesis. One variant of sequential sampling, *curtailed sampling*, was proposed by Wetherill (1975) (for a detailed description, see Reiber, Schnuerch, & Ulrich, 2022). In this method, certain stopping rules are set before the sampling process. Those are defined by a maximum sample size, N_{max}, defining the maximum needed sample size to align with one of the hypotheses, and c_s, a limited number of observed "successes" (here: yes- answers) needed to reject the null hypothesis. In turn, the null hypothesis is confirmed if $c_f = N_{max} - c_s + 1$ "failures" (here: no-answers) are observed. N_{max} and c_s are determined via power analysis, depending on the preset hypotheses H_0 and H_1 and on the preset error probabilities α (to falsely reject H_0) and β (to falsely reject H_1). Reiber, Schnuerch, and Ulrich (2022) created applications for curtailed sampling to RRMs, including the UQM, which we used for the follow-up study.

In the follow-up study, we calculated a sequential binomial hypothesis test according to Reiber, Schnuerch, and Ulrich (2022). We set $\pi \leq 0.387$ as the null hypothesis, stating that the prevalence for blue eye color prevalence is lower than or equal to the upper boundary of the 95% confidence interval for the blue eye color estimated via DQ in the main study. As alternative hypothesis, we set $\pi \geq 0.489$, stating that the mentioned prevalence is at least equal to the lower boundary of the 95% confidence interval for the prevalence estimated via UQM in the main study. α and β were set to 0.05 each. The UQM variables p and q were set to the same values as in the main study, so to circa 0.67 and 0.5, respectively. With these variables, the power analysis resulted in the maximum possible sample size of $N_{max} = 579$. After $c_s = 265$ yes-answers or $c_f = 315$ no-answers would be observed, sampling would be stopped.

We predicted that there would be an order effect in accordance to our explanation for the unexpected results. So, we hypothesized that H_1 would be rejected, meaning that the estimate of blue eye color prevalence is in line with the corresponding DQ estimate of the main study.

Method

The follow-up study amounted to a simple one-group-design without experimental manipulations.

Like in the main study, we commissioned *Bilendi S.A.* to generate a sample with demographics as similar as possible to those of German citizens with a driver's license (see Kraftfahrt-Bundesamt [Federal Office for Motor Traffic], 2022).

The procedure and materials used in the follow-up study were identical to those in the main study. The only differences were that in the follow-up study, there was no DQ group, and the order of the eye colour question and the drinking and driving question was switched. Additionally, since it was not necessary to vary the time constraints in this design, the drinking and driving question always referred to the past year. The data exclusion procedure was similar to the one used in the main study as well. But, to make sure that the sampling process was not stopped prematurely, participants who answered the survey too fast were marked as fast respondents immediately after completing the survey. To do this, we assumed the average response timings in the follow-up study to be similar to the response times of the UQM group in the main study. This seemed like a valid assumption, since the surveys were identical except for the order of two questions. Thus, we compared the RSI values (according to Leiner, 2019b) of the participants directly with the corresponding RSI values in the UQM group in the main study.

Sampling started on September 16th, 2022 and was completed on September 20th, 2022.

Results

Sampling was stopped after $N = 498$ observations. Because the stopping criterion of 265 yes-answers was reached, the null hypothesis was rejected.

The sample was similarly distributed in terms of demographic data compared with the data for Germans with driver's licenses (Kraftfahrt-Bundesamt [Federal Office for Motor Traffic], 2022), as illustrated in Table C1.

To estimate π, we used an estimator for the parameter λ that corrects for bias induced by the sequential sampling procedure (Reiber, Schnuerch, & Ulrich, 2022),

$$\hat{\gamma} = \frac{c_s - 1}{N - 1}. \tag{C1}$$

Inserting the corrected estimate $\hat{\gamma}$ into the standard formulae for calculating π in the UQM (see Equations 7 and 9) yields the point estimate $\hat{\pi} = 0.548$ (95% CI [0.483, 0.614]).

Discussion

In conclusion, the hypothesis test points toward the follow-up study's UQM estimate for the prevalence of blue eye color not being significantly lower than than the one in the main study. Thus, an order effect seems unlikely as the explanation for the unexpected differences between blue eye color prevalence estimates using DQ and UQM

Table C1

Distribution of demographics in the sample of the follow-up study compared to those of the German population owning a driver's license

Demographic		Distribution	
		sample	population
Gender	female	46.2%	43.1%
	male	53.6%	56.9%
	non-binary	0.2%	0.0%
Age	18-29 years	15.1%	16.8%
	30-39 years	19.7%	20.1%
	40-49 years	16.7%	14.2%
	50-59 years	19.1%	16.7%
	60 years and older	29.5%	31.8%

Note. The reference distribution of demographics is based on data by the Kraftfahrt-Bundesamt [Federal Office for Motor Traffic] (2022).

methods in the main study.

D Paper 4

Iberl, B. (2021). Ein, zwei Bier und ab ans Lenkrad? - Prävalenzschätzung von Alkohol am Steuer durch das Unrelated Question Model [One or Two Drinks Before Going for a Ride? - Prevalence Estimation of Driving Under the Influence via the Unrelated Question Model]. *Kriminologie-Das Online-Journal | Criminology-The Online Journal, 3*(3). https://doi.org/10.18716/ojs/krimoj/2021.3.5.

Candidate contributions to the article

Status	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Published	100%	100%	100%	100%

No. 3/2021

KrimOJ Kriminologie – Das Online-Journal
Criminology – The Online Journal

Benedikt Iberl

Ein, zwei Bier und ab ans Lenkrad? – Prävalenzschätzung von Alkohol am Steuer durch das Unrelated Question Model

Wie häufig in Deutschland die bewusste Inkaufnahme von Alkoholverstößen im Straßenverkehr auftritt, ist bislang fast gänzlich unerforscht. Die entsprechende Prävalenz wurde in der vorliegenden Studie mit Hilfe des Unrelated Question Model (UQM; Greenberg et al., 1969), einer indirekten Fragemethode für heikle Themen, ermittelt. Dabei wurde auch der Einfluss des Instruktionsverständnisses untersucht. In einer Online-Umfrage wurden Studierende der Eberhard Karls Universität Tübingen entweder mittels direkter Fragemethode oder via UQM (in zwei Varianten) befragt. Es resultiert eine Schätzung der Lebenszeitprävalenz von 44 % für alle drei Gruppen. Erklärungsansätze für die Ergebnisse werden diskutiert und zukünftige Forschungsfragen bezüglich Alkohol im Straßenverkehr und der Funktionsweise des UQM aufgestellt.

Schlagwörter: Alkohol am Steuer; bewusste Inkaufnahme von Alkoholverstößen; Direkte und Indirekte Fragemethoden; Heikle Fragen; Randomized Response Technique; Straßenverkehr; Unrelated Question Model

One or Two Drinks Before Going for a Ride? – Prevalence Estimation of Driving Under the Influence via the Unrelated Question Model

Until now, there has been almost no research as on the prevalence of the willful acceptance of driving under the influence in Germany. In this study, the respective prevalence was determined by employing the Unrelated Question Model (UQM; Greenberg et al., 1969), which is an indirect questioning technique for sensitive topics. During this process, the influence of comprehension instruction was also examined. In an online survey, students of the University of Tübingen were questioned, either via direct questioning or via UQM (in two variations). A lifetime prevalence of 44 % was estimated for all three groups. Explanations for the results are discussed and future research questions regarding both driving while intoxicated and the mechanisms of the UQM are presented.

Keywords: direct and indirect questioning techniques, driving while intoxicated, Randomized Response Technique, road traffic, sensitive questions, Unrelated Question Model, willful acceptance of driving under the influence

1. Einleitung

Alkohol im Straßenverkehr ist schon lange ein Thema, das die Gesellschaft bewegt und beschäftigt. Dabei ist es durchaus facettenreich: Es beinhaltet sowohl eindeutig rechtswidrige und gesellschaftlich geächtete Verhaltensweisen wie das Fahren im Vollrausch, gleichzeitig aber auch grenzwertige Handlungen, die nicht zwingend mit Regelverstößen einhergehen

https://doi.org10.18716/ojs/krimoj/2021.3.5

müssen. Situationen wie die folgenden mögen vielen bekannt vorkommen: Man trinkt am frühen Abend beim Grillen ein oder zwei Gläser Bier oder Wein und setzt sich einige Stunden später für den Heimweg hinters Lenkrad. Man hat auf einer Hochzeit etwas (oder deutlich) über die Stränge geschlagen und muss am nächsten Vormittag nach einer zu kurzen Nacht und einem Kater nach Hause fahren. In derartigen Situationen ist man zwar meist nicht betrunken, ob man die Regelungen zu Promillegrenzen im Straßenverkehr noch einhält, ist dagegen eine andere Frage. Dabei kann manchmal der Eindruck entstehen, dass eine solche bewusste Inkaufnahme von Alkoholverstößen im Gegensatz zum Fahren bei Volltrunkenheit in einigen Kreisen als eine Art „Kavaliersdelikt“ betrachtet wird; jedoch gibt es seit jeher auch Stimmen, die derartiges Verhalten entschieden verurteilen. In diesem Zusammenhang sei an die durch den ehemaligen bayerischen Ministerpräsidenten Günther Beckstein ausgelöste Diskussion erinnert, welche die Frage zum Gegenstand hatte, ob man nach dem Konsum von zwei Maß Bier noch fahrtüchtig sei oder nicht (DER SPIEGEL, 2008; Süddeutsche Zeitung, 2008).

1.1 Alkohol am Steuer

1.1.1 Zahlen und Trends im Hellfeld

Unbestritten ist wohl, dass Alkohol am Steuer schwere Unfälle verursacht. Diese sogenannten „Alkoholunfälle“ sind überdurchschnittlich schwer. So war 2019 nach Angaben des statistischen Bundesamtes jede/r 13. Verkehrstote (7,5 %) auf Unfälle unter dem Einfluss von Alkohol zurückzuführen (Statistisches Bundesamt, 2020a; Statistisches Bundesamt, 2020b). Verheerend ist dabei, dass Verkehrsteilnehmer*innen den Einfluss von Alkohol bisweilen unterschätzen. Bereits ab 0,3 Promille (entspricht ca. einem Glas Wein oder Bier) treten Beeinträchtigungen auf, die mit einer Erhöhung der Unfallwahrscheinlichkeit einhergehen, etwa die Einschränkung der Bewegungskoordination, des Sehfeldes und der Fähigkeit, Entfernungen einzuschätzen (Bund gegen Alkohol und Drogen im Straßenverkehr e.V., 2011; Alkohol? Kenn dein Limit., 2021). Ab 1,0 Promille ist die Unfallgefahr rund zehnmal höher als im nüchternen Zustand (Bund gegen Alkohol und Drogen im Straßenverkehr e.V., 2011).

In der jüngeren Vergangenheit fand in Deutschland eine Sensibilisierung für die mit Alkoholkonsum verbundenen Gefahren statt. Initiativen wie „Alkohol? Kenn dein Limit.“ (Bundeszentrale für gesundheitliche Aufklärung) sorgen mitunter gezielt bei Jugendlichen für eine frühe Aufklärung. Insgesamt ist der pro-Kopf-Konsum von Alkohol in Deutschland seit den 50er Jahren stark angestiegen, erreichte in den 70er Jahren einen Höhepunkt und ist seither kontinuierlich gesunken (John & Hanke, 2018). Parallel zu dieser Entwicklung wurden auch die entsprechenden im Straßenverkehr geltenden Regelungen zunehmend verschärft: 1973 wurde erstmals ein fester Grenzwert von 0,8 Promille Blutalkoholkonzentration (BAK) eingeführt (Statistisches Bundesamt, 2020b). Seit 1998 gelten die heute weithin bekannten „0,5 Promille“ als BAK-Grenzwert, seit 2007 gilt ein absolutes Alkoholverbot für Fahranfänger*innen und Verkehrsteilnehmende unter 21 Jahren (Statistisches Bundesamt, 2020b). Verstöße gegen diese Regelungen werden je nach Höhe des Alkoholpegels und Schwere des Vergehens als Ordnungswidrigkeit mit hohen Bußgeldern und Punkten in Flensburg oder als Straftat mit Geld- oder Freiheitsstrafen geahndet; in beiden Fällen kann ein Entzug der Fahrerlaubnis (auch dauerhaft) erfolgen (Statistisches Bundesamt, 2020b).

Problematisch gestaltet sich die Beantwortung der Frage nach der Prävalenz von Alkohol am Steuer. Die Verkehrsunfall-Statistiken des Statistischen Bundesamtes sind hier leider nur bedingt von Belang, da mutmaßlich der Großteil der Alkoholfahrten nicht mit einem Unfall endet. Aufschlussreicher sind die Zahlen des Fahreignungsregisters (FAER, Kraftfahr-Bundesamt), in das sämtliche bekannt gewordene Verkehrsverstöße einfließen. Im Jahr 2019 wurden 115 623 Alkoholverstöße (79 727 Straftaten und 35 896 Ordnungswidrigkeiten) festgestellt (Kraftfahr-Bundesamt, 2021). Am 1.1.2020 waren im FAER 11 134 475 Personen mit gültiger Fahrerlaubnis registriert (ebd.). Setzt man diese Zahlen in Beziehung zueinander, lässt sich die Jahresprävalenz von Alkoholverstößen grob einschätzen: Demnach hat 2019 ca. 1 % (Verhältnis der Alkoholverstöße zu Fahrerlaubnissen) der deutschen Fahrzeugführenden gegen die Regelungen verstoßen. Diese Schätzung muss freilich als Obergrenze verstanden werden, denn einzelne Personen können auch für mehrere Verstöße in diesem Jahr verantwortlich sein. Da aber Verkehrskontrollen nicht lückenlos durchgeführt werden und Alkoholkontrollen bei Verkehrsbeteiligten meist anlassbezogen (z. B. nach einem Unfall oder auffälligem Fahrverhalten) oder zu gewissen Zeiten an gewissen Orten (z. B. samstagabends in der Nähe einer Disco oder eines Jahrmarktes) stattfinden, muss hier von einem großen Dunkelfeld ausgegangen werden (Kulemeier, 1991).

1.1.2 Das Dunkelfeld von Alkohol im Straßenverkehr

Überraschenderweise existieren insbesondere in Deutschland recht wenige Dunkelfeldstudien zur Häufigkeit von Alkohol am Steuer. Die Methode der Wahl bei derartigen Untersuchungen ist das sogenannte „Roadside-Survey“ – zu Forschungszwecken und stichprobenartig durchgeführte Alkoholkontrollen im Straßenverkehr, häufig in Zusammenarbeit mit der Polizei. Nach den hier durchgeführten Recherchen wurde das jüngste und einzige Roadside-Survey in Deutschland zwischen 1992 und 1994 durchgeführt („Das Deutsche Roadside-Survey“, DRS; Krüger, 1998; Schöch, 2001). Bei diesem wurden in Unterfranken und Thüringen im Rahmen polizeilicher Verkehrskontrollen 24 000 Kraftfahrer*innen untersucht. Bei ca. 5 % der Stichprobe wurde ein Alkoholkonsum festgestellt, in Unterfranken lag die Prävalenz um über einen Prozentpunkt höher als in Thüringen. Der damals geltende Grenzwert von 0,8 Promille wurde bei etwa 0,5 % der Fahrten überschritten. Bei rund 1 % der Messungen trat ein Wert über 0,5 Promille auf, bei etwas unter 2 % ein Wert über 0,3 Promille. Diese Ergebnisse stehen dem Anschein nach den Zahlen des Hellfeldes zwar nicht entgegen, jedoch ist fraglich, wie aktuell die von Krüger und Kolleg*innen ermittelten Prävalenzen knapp 30 Jahre später noch sind.

Als Anhaltspunkte für die Schätzung des Dunkelfelds können auch einige jüngere Roadside-Surveys aus anderen Ländern dienen. Beispielsweise wurde 2003 in Belgien ein Roadside-Survey durchgeführt, bei dem insgesamt bei 3 % der Kontrollierten eine BAK über 0,5 Promille festgestellt wurde; bei dem Teil der Messungen, der an Wochenendnächten durchgeführt wurde, waren es allerdings 8 % (Vanlaar, 2005). In Thailand wurde 2005-2006 ein Alkoholfahrtenanteil von 5,5 % festgestellt, davon etwas weniger als die Hälfte mit einer BAK über 0,5 Promille (Ingsathit et al., 2009). Zwischen 2006 und 2011 wurden in 13 europäischen Ländern unter dem Banner des Projekts „DRUID“ (Driving under the influence of drugs, alcohol and medicines) diverse Roadside-Surveys durchgeführt. Dabei wurde für Europa insgesamt eine Prävalenz von ca. 3,5 % für Alkohol im Straßenverkehr geschätzt (Houwing et al., 2011a; Houwing et al., 2011b). Bei einem Roadside-Survey in British Columbia, Kanada, wurden 2012 bei nächtlichen Kontrollen zwischen Mittwoch und Samstag Alkoholfahrtenanteile von 7 % und Überschreitungen der 0,5 Promille bei ca. 2 % der kontrollierten Personen festgestellt (Beasley et al., 2012). Eine Studie aus den USA fand ähnliche Zahlen: bei 8 % wurde Alkoholkonsum nachgewiesen, bei 3 % über 0,5 Promille (Berning, Compton & Wochinger, 2015). Diese Erkenntnisse lassen einerseits eine relativ homogene Prävalenz über Ländergrenzen hinweg vermuten und unterstützen andererseits die These, dass die deutschen Hellfelddaten die tatsächliche Häufigkeit von Alkohol im Straßenverkehr unterschätzen.

Die wohl neueste Dunkelfelduntersuchung zu Alkohol im Straßenverkehr in Deutschland wurde 2012 durchgeführt. Dabei wurde bei regelmäßig autofahrenden Personen zwischen 18 und 39 Jahren die Häufigkeit von Fahrten unter dem Einfluss psychoaktiver Substanzen mit Hilfe von Smartphones und Selbstberichten ermittelt (Walter, 2012). Aus den berichteten Mengen konsumierten Alkohols und weiteren Angaben wurde schließlich näherungsweise die BAK zum Zeitpunkt der Fahrten berechnet. Verglichen wurden außerdem regelmäßige Nutzer*innen illegaler Drogen und als Kontrollgruppe Personen, die mindestens ein Jahr lang keine illegalen Drogen zu sich genommen hatten. Der Anteil an Alkoholfahrten war bei Konsumierenden illegaler Drogen deutlich höher als in der Kontrollgruppe. Für die Gesamtpopulation der 18- bis 24-Jährigen schätzte Walter einen Wert für Alkoholfahrten in Höhe von 1,6 %, für die 25- bis 39-Jährigen resultierten 3,3 %. Aufschlussreich sind hier die deutlichen Unterschiede zwischen den Altersgruppen und der Effekt des regelmäßigen Konsums illegaler Drogen. Aufgrund der kleinen Stichprobengröße (*N* = 100 Personen in der Kontrollgruppe) und der Methode der Selbstberichte können aus der Studie allerdings nur eingeschränkt Schlüsse über die Prävalenz der Alkoholfahrten in der Gesamtbevölkerung geschlossen werden.

Diese Studien – und dieser Mangel an Studien in Deutschland – zeigen den Bedarf für aktuelle Untersuchungen in dieser Thematik auf. Die aktuellen Zahlen für Deutschland geben lediglich Aufschluss über *entdeckte Verstöße*. Nach Angaben des statistischen Bundesamtes sind jedoch immerhin an 7 % der Alkoholunfälle mit Personenschaden Fahrzeugführer*innen beteiligt, die 0,5 Promille oder weniger im Blut haben, einen etwas größeren Anteil haben Personen mit BAK-Werten zwischen 0,5 und 0,8 Promille (Statistisches Bundesamt, 2020b). Wie diese Zahlen zeigen, sind Alkoholfahrten im „Grenzbereich des Erlaubten" durchaus relevant für die Verkehrssicherheit, gleichzeitig bleiben sie aber wahrscheinlich oft unentdeckt. Wie häufig Fahrzeugführer*innen im Anschluss an Alkoholkonsum bewusst in Kauf nehmen, die geltenden Regeln zu verletzen – an dieser Stelle darf an die Beispiele der Grillparty und der Hochzeit erinnert werden – ist gänzlich unbekannt. Gerade weil die so handelnden Personen nicht volltrunken am Straßenverkehr teilnehmen oder mit hundertprozentiger Sicherheit die Regeln absichtlich verletzen, könnte dieses Verhalten gesellschaftlich bisweilen als „Kavaliersdelikt" wahrgenommen werden, obwohl möglicherweise bereits Regelverstöße vorliegen und das Unfallrisiko deutlich gesteigert ist.

Die Prävalenz dieser *bewussten Inkaufnahme von Alkoholverstößen im Straßenverkehr* ist daher Gegenstand dieser Untersuchung. Aus forschungsökonomischen Gründen wurde eine studentische Stichprobe befragt und statt eines aufwendigen Roadside-Surveys die bedeutend sparsamere Methode der Online-Umfrage angewendet.

1.2 Das Stellen heikler Fragen – Die „Randomized Response Technique"

1.2.1 Hintergrund und Funktionsweise

Bei der Erstellung von Fragebögen zur Erforschung sensibler Thematiken existiert ein grundlegendes Problem: Der Effekt der sozialen Erwünschtheit kann bei Umfragen zu systematischen Antwortverzerrungen führen, wodurch das zu ermittelnde Merkmal unterschätzt wird (z. B. Krumpal, 2013; Nederhof, 1985). So könnten Befragte strafrechtliche oder soziale Konsequenzen durch eine ehrliche Antwort befürchten, insbesondere wenn sie sich nicht anonym fühlen (Krumpal, 2013). Dass es sich bei Alkohol am Steuer um ein sensibles Umfragethema handelt, zeigen Befragungen, die im Rahmen des DRS (Krüger, 1998) durchgeführt wurden. Es kann vermutet werden, dass diese wahrgenommene Sensibilität beim Thema „Alkohol im Straßenverkehr" auch heute noch besteht.

Spezielle Fragemethoden wurden entwickelt, um diese befürchteten Antwortverzerrungen bei der Erfassung heikler Merkmale zu minimieren. Eine solche Fragemethode stellt die *Randomized Response Technique* (RRT) von Warner (1965) dar. Die Grundidee der RRT besteht darin, dass die individuellen Antworten der Befragten durch eine Art „statistische Störvariable" verschleiert werden. In der Regel wird hierfür ein von den Teilnehmenden eigenständig durchgeführtes Zufallsexperiment der eigentlichen Frage nach dem sensiblen Merkmal vorgeschaltet. Durch diese „Verschleierung" entsteht eine vollständige Anonymität, wodurch die Befragten ohne Angst vor negativen Konsequenzen zu ehrlichen Antworten bewegt werden sollen. Mittlerweile gibt es zahlreiche verschiedene Fragemethoden, die sich an die RRT von Warner anlehnen (sog. *Randomized Response Models*, RRMs). In der vorliegenden Studie kam das *Unrelated Question Model* (UQM; Greenberg et al., 1969), ein recht einfaches und weit verbreitetes RRM, zur Anwendung. Das UQM gilt als eine etablierte indirekte Fragetechnik, die in Hinblick auf statistische Eigenschaften und „psychologische Akzeptanz" im Vergleich zu anderen RRMs besser abschneidet (Reiber, Pope & Ulrich, 2020; Ulrich et al., 2012), weshalb es in dieser Studie ausgewählt wurde.[1]

Beim UQM (s. Abb. 1) werden den Teilnehmenden zwei Ja-/Nein-Fragen präsentiert: Eine „heikle Frage" nach dem zu erforschenden sensiblen Merkmal (z. B.: „Ich habe schon einmal Kokain konsumiert") und eine „neutrale Frage" nach einem neutralen Merkmal mit einer bekannten oder schätzbaren Auftretenswahrscheinlichkeit π_n (z. B.: „Ich habe in der ersten Jahreshälfte Geburtstag"). Zunächst wird durch die Teilnehmenden im Geheimen ein Zufallsexperiment durchgeführt – etwa ein Münzwurf. Dabei muss es sich um ein Bernoulli-Experiment handeln, also um ein Experiment mit zwei möglichen Ergebnissen (im Falle des Münzwurfs „Kopf" oder „Zahl"). Die Wahrscheinlichkeit p, dass das erste Ereignis eintritt, muss bekannt sein. Abhängig vom Ergebnis des Zufallsexperiments, welches nur den Teilnehmenden selbst bekannt sein darf, wird nun entweder auf die heikle (mit der Wahrscheinlichkeit p) oder auf die neutrale Ja-/Nein-Frage (mit der Gegenwahrscheinlichkeit $1 - p$) verwiesen. Die zugewiesene Frage soll im Anschluss ehrlich mit „Ja" oder „Nein" beantwortet werden – da nur die Befragten selbst wissen, was bei dem Zufallsexperiment resultierte, wissen auch nur sie selbst, ob sich ihre Antwort auf die neutrale oder auf die heikle Frage bezieht. Die Prävalenz π_s des

[1] Eine neuere und sehr effiziente indirekte Fragemethode ist das *Crosswise-Model* (CWM; Yu, Tian & Tang, 2008), eines der sog. *Nonrandomized Response Models*, bei denen die Verschleierung der Antwort nicht durch ein Zufallsexperiment erreicht wird. Für die Rolle von Verständnis im CWM s. Meisters, Hoffmann und Musch (2020).

sensiblen Merkmals in der untersuchten Stichprobe kann dann unter Einbeziehung der beobachteten relativen Häufigkeit einer Ja-Antwort und der per Design festgelegten Wahrscheinlichkeiten p und π_n geschätzt werden.
Die Wahrscheinlichkeit λ einer Ja-Antwort berechnet sich wie folgt:

$$\lambda = \pi_s p + \pi_n(1-p)$$

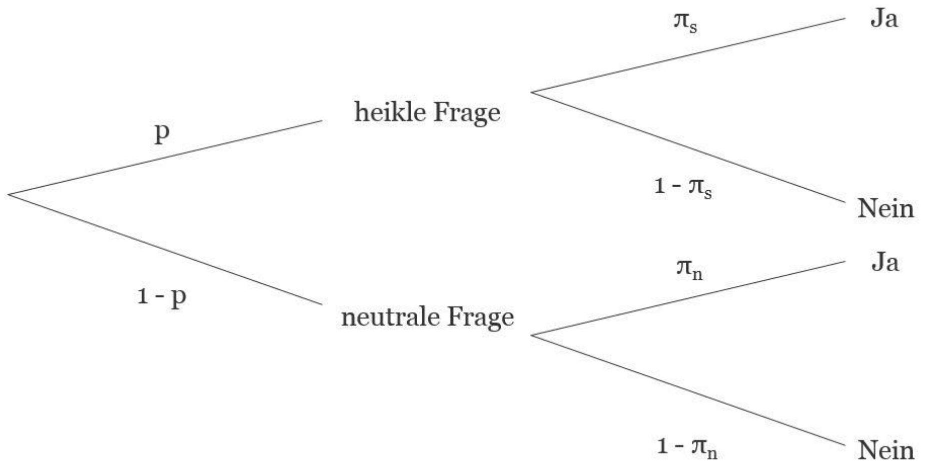
λ kann durch die beobachtbaren relativen Häufigkeiten der Antworten geschätzt werden:

$$\hat{\lambda} = \frac{N_{ja}}{N_{ja} + N_{nein}}$$

Eine Schätzung für π_s ergibt sich dann durch:

$$\hat{\pi}_s = \frac{\hat{\lambda} - (1-p) \cdot \pi_n}{p}$$

Abbildung 1. Das Unrelated Question Model nach Greenberg, Abul-Ela, Simmons und Horvitz (1969)



In Online-Befragungen wird als Zufallsmechanismus anstelle eines Münzwurfs oft der Geburtstag der Befragten genutzt. Da jeder mögliche Geburtstag ungefähr gleich häufig vorkommt (Ulrich et al., 2012), kann z. B. die Frage „Haben Sie in der ersten Jahreshälfte Geburtstag (vor dem 1. Juli)?" als Zufallsexperiment fungieren (mit der geschätzten Wahrscheinlichkeit p = 0.5 einer Ja-Antwort). Derartige Fragen nach dem Geburtsdatum können auch als neutrale Frage verwendet werden.

1.2.2 Vor- und Nachteile von Randomized Response Models

Die Anwendung von RRMs geht stets mit einigen Nachteilen einher. So steigt erstens durch das künstlich zugefügte Zufallsexperiment die Varianz der Messung bedeutend an (Ulrich et al., 2012). Das führt dazu, dass für belastbare inferenzstatistische Aussagen im Vergleich zu direkten Fragemethoden (*Direct Questioning*, DQ) große Stichproben vonnöten sind (ebd.).

Daneben ist die Flexibilität des Einsatzes von RRMs durch die strengen Vorgaben bei der Frageformulierung eingeschränkt (z. B. die Beschränkung auf Ja-/Nein-Fragen im klassischen UQM).

Der entscheidende Vorteil von RRMs gegenüber DQ besteht darin, dass bei sensiblen Fragestellungen validere Prävalenzen resultieren (Lensvelt-Mulders et al., 2005). Tendenziell gilt hierbei: Je sensibler die erforschte Thematik, desto größer der Vorteil der RRMs (ebd.). Ehemals wurde die Validität von RRMs oft anhand der sogenannten „more-is-better"-Annahme beurteilt, die besagt, dass bei heiklen Fragen höhere Prävalenzen auch die valideren sind (Buchman & Tracy, 1982; Lensvelt-Mulders et al., 2005; Umesh & Peterson, 1991). Die Gültigkeit dieser Annahme wird mittlerweile angezweifelt (Höglinger & Jann, 2018). Eine bessere Beurteilung der Validität wird durch sog. „starke Validierungsstudien" ermöglicht, durch die die Anteile falsch-positiver und falsch-negativer Antworten ermittelt werden können (ebd.).

Ein wichtiger Befund in einigen dieser starken Validierungsstudien ist, dass die oft höheren und valideren Prävalenzschätzungen bei RRMs in Teilen auch auf falsch-positive Antworten zurückzuführen sind (Höglinger & Diekmann, 2017; Höglinger & Jann, 2018; Meisters, Hoffmann & Musch, 2020). Einschlägig gibt es mehrere Hinweise darauf, dass das Verständnis von RRMs den Versuchspersonen nicht immer leicht fällt (z. B. De Schrijver, 2012; Hoffmann & Musch, 2016; Landsheer, Van Der Heijden & Van Gils, 1999; Lensvelt-Mulders & Boeije, 2007; Wolter, 2012). Das Verständnis der Proband*innen, wie oder dass ihre Anonymität in einem solchen Setting besser geschützt wird als bei DQ, ist jedoch eine wichtige Grundannahme bei RRMs. Denn wenn die Befragten sich bei RRMs nicht anonymer fühlen, warum sollten sie dann ehrlicher antworten? Es könnte gar argumentiert werden, dass eine Verletzung dieser Annahme die Daseinsberechtigung von RRMs in Frage stellt. Eine Forschungsfrage von höchster Priorität in Hinblick auf die Anwendbarkeit von RRMs ist also, welche Rolle dem Verständnis der Fragemethode zuteilwird.

Dieser Forschungsfrage widmete sich jüngst ein Forschungsteam der Universität Düsseldorf (Meisters, Hoffmann & Musch, 2020). Untersucht wurde in einem starken Validierungsdesign, inwiefern sich die Häufigkeit falsch-positiver und -negativer Antworten bei dem sog. *Crosswise-Model* (CWM; Yu, Tian & Tang, 2008) durch Verständnishilfen reduzieren lassen. Als Verständnishilfen wurde den Proband*innen die Befragungsmethode und der (stochastische) Hintergrund der Frage nach ihrem Geburtstag erklärt. Anschließend wurden ihnen vier Verständnisfragen gestellt, um ihr Instruktionsverständnis zu fördern und zu prüfen. Dabei sollten sie sich in Personen hineinversetzen, die an bestimmten Tagen Geburtstag hatten und entweder das sensitive Attribut trugen oder nicht. Aus deren Sicht sollte dann die CWM-Befragung beantwortet werden. In den Ergebnissen zeigte sich, dass die Verständnishilfen dazu beitragen konnten, falsch-positive Antworten zu verringern. Allerdings konnte dieser Effekt nur bei Versuchspersonen mit hohem Bildungsniveau beobachtet werden. Die Verständnishilfen hatten zudem keinen Einfluss auf die Häufigkeit falsch-negativer Antworten.

Dass das Verständnis von RRMs mit dem Bildungsstand zusammenhängt, zeigten zuvor auch andere Untersuchungen (Böckenholt & van der Heijden, 2007; Hoffmann & Musch, 2016; Landsheer, Van Der Heijden & Van Gils, 1999; Wolter, 2012). Die Studie von Meisters, Hoffmann und Musch (2020) liefert darüber hinaus erstmals Befunde zur Wirksamkeit von Verständnishilfen bei RRMs und zu einer möglichen Wechselwirkung mit dem Bildungsniveau der Befragten. Bis dato fehlt es indes gänzlich an vergleichbaren Erkenntnissen für das UQM. In der vorliegenden Untersuchung wird sich dieser Forschungslücke gewidmet.

1.3 Die vorliegende Studie

Aus den bisherigen Ausführungen lassen sich die zentralen Forschungsfragen der Untersuchung ableiten.

- Erstens: Wie hoch ist die Lebenszeitprävalenz für die bewusste Inkaufnahme von Alkoholverstößen im Straßenverkehr bei Studierenden?[2]
- Zweitens: Welchen Effekt haben ausführliche Verständnishilfen bei der Anwendung des UQM auf die ermittelte Prävalenz? Wie fallen im Vergleich die Prävalenzen unter Anwendung von DQ und von UQM ohne Verständnishilfen aus?

Folgende Forschungsergebnisse wurden vor Durchführung der Erhebung erwartet:

- Hypothese 1: Es wird davon ausgegangen, dass eine hohe Lebenszeitprävalenz der bewussten Inkaufnahme von Alkoholverstößen resultiert. Eine Schätzung der Prävalenz fällt a priori anhand mangelnder vergleichbarer Untersuchungen schwer.
- Hypothese 2: Es werden die Annahmen getroffen, dass die Frage nach Alkohol am Steuer als eine sensible Thematik wahrgenommen wird und dass höheres Verständnis zu ehrlicherem Antwortverhalten führt. Davon ausgehend wird erwartet, dass bei Anwendung des UQM höhere Prävalenzen als bei DQ resultieren. Da bei einer studentischen Stichprobe von einem hohen Bildungsniveau ausgegangen werden kann, wird eine Wirksamkeit der Verständnishilfen erwartet (vgl. Meisters, Hoffmann & Musch, 2020), wodurch die Prävalenzen beim UQM gesteigert werden dürften.

2. Methoden

2.1 Stichprobe

Die Stichprobe bestand aus 999 Studierenden der Eberhard Karls Universität Tübingen, die über den E-Mail-Verteiler der Universität kontaktiert wurden. Die Teilnehmenden gaben an, volljährig zu sein, einen Führerschein der Bundesrepublik Deutschland zu besitzen, als Studierende eingeschrieben zu sein und fließend Deutsch zu sprechen. Es wurden keinerlei demographische Daten erhoben.

60 Teilnehmende wurden ausgeschlossen, da sie die Umfrage vor der Beantwortung der zentralen Frage nach „Alkohol am Steuer“ abbrachen. 46 weitere Teilnehmende wurden ausgeschlossen, da sie die Umfrage zu schnell bearbeiteten. Als entsprechendes Ausschlusskriterium wurde ein Relative Speed Index (RSI; Leiner, 2019a) von über 2,0 ausgewählt.[3] Die Daten der

[2] „Die bewusste Inkaufnahme von Alkoholverstößen im Straßenverkehr“ wird hier wie folgt definiert: Das Steuern eines Kraftfahrzeuges (Auto, Motorrad, Motorroller etc.) trotz bestehender Unsicherheit, ob die gesetzlichen Vorschriften zur erlaubten Blutalkoholkonzentration eingehalten werden.

[3] Hierfür wird für jede Umfrageseite der Median der Bearbeitungszeit in der jeweiligen Gruppe durch die individuelle Bearbeitungszeit jeder Person dividiert. Ein RSI von 2,0 bedeutet, dass die Umfrage doppelt so schnell bearbeitet wurde als die mittlere Bearbeitungsdauer. Um den Einfluss von Ausreißern einzuschränken, wird das berechnete Verhältnis pro Umfrageseite auf maximal 3,0 begrenzt. Bei der Festlegung des Grenzwertes auf 2,0 wurde sich der Empfehlung Leiners angeschlossen.

übrigen 893 Teilnehmenden wurden für die Auswertung berücksichtigt. Im Laufe der Umfrage wurde die Stichprobe in die Gruppen „UQM" (*N* = 305), „UQM-V" (UQM mit Verständnishilfe, *N* = 278) und „DQ" (*N* = 310) unterteilt.

2.2 Material

Die Umfrage wurde als Online-Fragebogen über das Umfragetool SoSci Survey (Leiner, 2019b) erstellt und von den Teilnehmenden auf digitalen Endgeräten ihrer Wahl bearbeitet.

In einem Textblock wurde auf die möglichen Folgen von Alkohol am Steuer hingewiesen, um die empfindliche Natur des Themas zu unterstreichen (s. Tabelle 1). Der UQM-V-Gruppe wurden Verständnishilfen präsentiert (s. Tabelle 2). Zusätzlich kamen Übungsfragen zum Einsatz, bei denen die Studierenden aufgefordert wurden, UQM-Fragen aus der Sicht fiktiver Personen zu beantworten (s. Tabelle 3). Die Frage nach Alkohol am Steuer, welche den UQM- und UQM-V-Gruppen präsentiert wurde, ist in Tabelle 4 dargestellt, diejenige, die der DQ-Gruppe gestellt wurde, in Tabelle 5.

Allen Gruppen wurden die folgenden Fragen über den Eindruck der Methode gestellt: „Wie gut haben Sie verstanden, was Sie auf der vorherigen Seite dieser Umfrage tun sollten?", „Haben Sie das Gefühl, dass Ihre Anonymität in dieser Umfrage gewährleistet ist?", „Wie unangenehm wäre es für Sie, wenn Sie in einem persönlichen Gespräch nach Alkohol am Steuer gefragt werden würden?" (jeweils fünfstufige Likertskalen als Antwortmöglichkeiten). Den Gruppen UQM und UQM-V wurde außerdem noch die anschließende Frage gestellt: „Wie gut haben Sie verstanden, auf welche Weise Ihre Anonymität in dieser Umfrage geschützt wird?" (fünfstufige Likertskala).

Die Fragen nach den Geburtstagen wurden gewählt, sodass gilt:

$$p = \frac{245.25}{365.25} = 0.671$$

$$\pi_n = \frac{181.25}{365.25} = 0.496$$

2.3 Ablauf

Am 14.10.2020 wurden sämtliche Studierende der Universität Tübingen per E-Mail um die Teilnahme an der Studie gebeten. In der Nachricht wurde der Untersuchungsgegenstand grob umrissen und der Link zur Umfrage zur Verfügung gestellt.

Zur Überprüfung des RSI als sinnvolles Kriterium wurden darüber hinaus die mittleren Bearbeitungszeiten und Lesegeschwindigkeiten der ausgeschlossenen Personen und der Reststichprobe verglichen. Die ausgeschlossenen Personen bearbeiteten die Umfrage in durchschnittlich lediglich 84,9 Sekunden, während die Reststichprobe im Mittel 221,1 Sekunden Zeit benötigte. Die Lesegeschwindigkeit der Personen mit RSI über 2,0 betrug für den gesamten Umfragetext durchschnittlich 616,57 Wörter pro Minute, die der Reststichprobe 258,59. Dies stützt die Anwendung des RSI als Ausschlusskriterium: Eine Lesegeschwindigkeit von mehr als 600 Wörtern pro Minute gilt als zu schnell, um den Inhalt des gelesenen Textes verlässlich aufnehmen zu können (Brysbaert, 2019; Carver, 1992; Musch & Rösler, 2011).

Tabelle 1. „Sensitivitätstext“ zu den möglichen Folgen von Alkohol am Steuer

Alkohol am Steuer
Verstöße gegen die Regelungen zu Alkohol am Steuer werden in Deutschland mit hohen Bußgeldern, Punkten in Flensburg, Fahrverboten (auch dauerhaft) oder sogar mit Freiheitsstrafe bestraft. Alkohol am Steuer ist in Deutschland leider nicht selten und führt immer wieder zu Unfällen im Straßenverkehr. Nach Angaben des statistischen Bundesamtes ist *jede/r 13. Verkehrstote* auf Unfälle unter dem Einfluss von Alkohol zurückzuführen. Die sogenannten „Alkoholunfälle“ sind oft besonders schwer: Tote und Schwerverletzte im Straßenverkehr sind bei Alkoholunfällen fast *doppelt so häufig* wie bei Unfällen allgemein. Verheerend ist dabei, dass Verkehrsteilnehmer/-innen den Einfluss von Alkohol unterschätzen. Schon ab 0,3 Promille sind Wahrnehmung und Reaktionsvermögen deutlich beeinträchtigt. Ab 1,0 Promille ist die Unfallgefahr *zehnmal höher* als im nüchternen Zustand. Wer sich unsicher über den eigenen Alkoholpegel ist, besonders bei Restalkohol, sollte also besser die Finger vom Steuer lassen, um seine Mitmenschen und sich selbst nicht zu gefährden.

Tabelle 2. Verständnishilfe zu den Fragen nach dem Geburtsdatum

Geburtstage als Zufallsgenerator
Als „*Zufallsgenerator*“ zur Zuordnung der Frage A oder B benutzt man bei dieser Methode häufig Würfel oder das Ziehen von Karten. Da der Einsatz von Würfeln oder Karten online jedoch schwierig ist, benutzen wir als Zufallsmechanismus *Geburtstage*. Geburtstage sind ungefähr gleich verteilt. Das bedeutet, dass es *ungefähr gleich wahrscheinlich* ist, an jedem der 365 Tage im Jahr geboren worden zu sein.
Beispiel 1: *Die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person im Januar, Februar oder März geboren wurde, beträgt ungefähr 25%* (3 Monate geteilt durch 12 Monate = ¼ = 25%)
Beispiel 2: *Die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person zwischen dem 1. und 15. Tag eines Monats Geburtstag hat, beträgt ungefähr 50%* (15 Tage geteilt durch 30 Tage = ½ = 50%)

Tabelle 3. Übungsfrage zum Einsatz als zusätzliche Verständnishilfe

Denken Sie bitte an Ihr Geburtsdatum.
Liegt Ihr Geburtstag zwischen dem 1. und 10. Tag des entsprechenden Monats? Dann beantworten Sie bitte Frage A wahrheitsgemäß. Liegt Ihr Geburtstag zwischen dem 11. und 31. Tag des entsprechenden Monats? Dann beantworten Sie bitte Frage B wahrheitsgemäß.
Frage A: Liegt Ihr Geburtstag in der ersten Jahreshälfte, also vor dem 1. Juli eines Jahres?
Frage B: Haben Sie jemals ein Kraftfahrzeug (Auto, Motorrad, Motorroller etc.) gesteuert, obwohl Sie sich nicht sicher waren, ob Sie den gesetzlichen Vorschriften zur erlaubten Blutalkoholkonzentration nachkamen (max. 0,0 (Probezeit/unter 21) bzw. 0,5 Promille)?
Sven ist noch nie Auto gefahren, wenn er davor Alkohol getrunken oder einen Restalkoholpegel hatte. Er hat am 06.08 Geburtstag. Wie müsste Sven antworten? • Ja • Nein

Tabelle 4. UQM-Frage nach Alkohol am Steuer

Denken Sie bitte an Ihr Geburtsdatum.
Liegt Ihr Geburtstag zwischen dem 1. und 10. Tag des entsprechenden Monats? Dann beantworten Sie bitte Frage A wahrheitsgemäß. Liegt Ihr Geburtstag zwischen dem 11. und 31. Tag des entsprechenden Monats? Dann beantworten Sie bitte Frage B wahrheitsgemäß.
Frage A: Liegt Ihr Geburtstag in der ersten Jahreshälfte, also vor dem 1. Juli eines Jahres?
Frage B: Haben Sie jemals ein Kraftfahrzeug (Auto, Motorrad, Motorroller etc.) gesteuert, obwohl Sie sich nicht sicher waren, ob Sie den gesetzlichen Vorschriften zur erlaubten Blutalkoholkonzentration nachkamen (max. 0,0 (Probezeit/unter 21) bzw. 0,5 Promille)?
Ihre Antwort auf Frage A oder auf Frage B (nur Sie wissen, worauf Sie hier antworten) lautet: • Ja • Nein

Tabelle 5. DQ-Frage nach Alkohol am Steuer

Haben Sie jemals ein Kraftfahrzeug (Auto, Motorrad, Motorroller etc.) gesteuert, obwohl Sie sich nicht sicher waren, ob Sie den gesetzlichen Vorschriften zur erlaubten Blutalkoholkonzentration nachkamen (max. 0,0 (Probezeit/unter 21) bzw. 0,5 Promille)?
• Ja • Nein

Vor der eigentlichen Umfrage wurden die Studierenden auf der ersten Seite auf die Freiwilligkeit und die Voraussetzungen der Teilnahme hingewiesen. Es wurde betont, dass keinerlei personenbezogene Daten erhoben werden. Auf der zweiten Seite wurde das Ziel der Studie vorgestellt und „Alkohol am Steuer" anhand von Beispielen als die bewusste Inkaufnahme von Alkoholverstößen im Straßenverkehr definiert. Auf der dritten Seite folgte der kurze Infotext über die möglichen (rechtlichen und Unfall-)Folgen alkoholisierter Verkehrsteilnahme.

Im Anschluss erfolgten die je nach Gruppe unterschiedlich ausführlichen Erläuterungen und Verständnishilfen zur jeweils genutzten Fragemethode. Der UQM-V-Gruppe wurden die o.g. Verständnishilfen und drei Übungsfragen präsentiert.[4] Der UQM-Gruppe wurde kurz erläutert, dass es sich um eine Befragungsmethode handelt, bei der durch die zufällige Zuordnung einer von zwei Ja-/Nein-Fragen vollständige Anonymität gewährleistet wird. Der DQ-Gruppe wurde die direkte Frage angekündigt und Anonymität zugesichert.

Daraufhin wurde die zentrale Frage gestellt. Auf der nächsten Seite wurden die Fragen zum Eindruck über die Befragung gestellt. Die Umfrage endete mit einer Gutscheinverlosung und einer Danksagung.

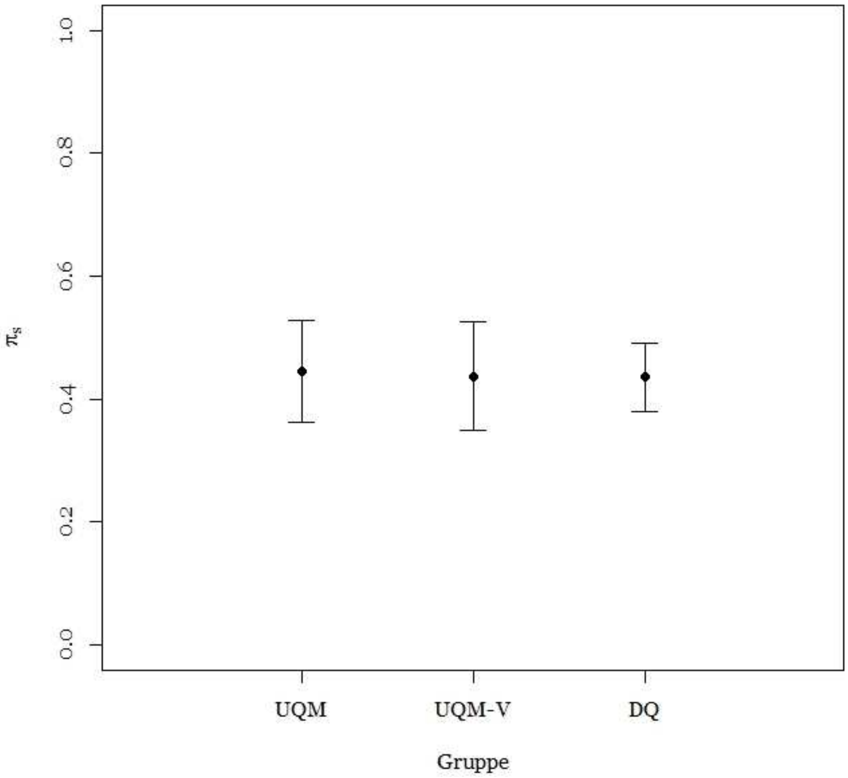
Am 20.11.2020 wurde die Umfrage offiziell beendet. Es waren keine weiteren Teilnahmen möglich. Ungefähr drei Monate nach Abschluss der Befragung erhielten die ausgelosten Gewinner*innen ihre Gutscheine.

[4] Dabei wurden aus fünf verschiedenen Übungsfragen zufällig drei gezogen, jedoch mit der Einschränkung, dass stets eine „inkongruente" Frage-Antwort-Konstellation enthalten war. Damit ist gemeint, dass die fiktive Person auf die neutrale Frage antworten müsste und diese Antwort nicht mit der Antwort auf die heikle Frage identisch ist.

3. Ergebnisse

Die Schätzungen der Lebenszeitprävalenz für die bewusste Inkaufnahme von Alkoholverstößen im Straßenverkehr bei Studierenden liegen über alle Gruppen gemittelt bei 44 %. Wie aus den Konfidenzintervallen in Abbildung 7 hervorgeht, unterscheiden sich die geschätzten Prävalenzen der UQM-Gruppe ($\pi_s = 0.45$, $N = 305$, 0,95-KI [0.36; 0.53]), der UQM-V-Gruppe ($\pi_s = 0.44$, $N = 278$, 0,95-KI [0.35; 0.52]) und der DQ-Gruppe ($\pi_s = 0.44$, $N = 310$, 0,95-KI[0.38; 049]) nicht signifikant voneinander.

Abbildung 2. Prävalenzschätzer und Konfidenzintervalle für die Gruppen UQM, UQM-V und DQ



Der Großteil der UQM-V-Gruppe beantwortete alle Übungsfragen richtig (68 %). 24 % beantworteten zwei der drei Fragen korrekt, 7 % eine Frage und 1 % der Befragten gaben drei falsche Antworten. Die Prävalenzschätzung für die Personen, die alle drei Fragen richtig beantworteten, beträgt $\pi_s = 0.44$ ($N = 190$, 0,95-KI [0.33; 0.54]). Für die Personen, die eine, zwei oder drei Fragen falsch beantworteten, ist $\pi_s = 0.43$ ($N = 88$, 0,95-KI [0.28; 0.59]). Die Prävalenzschätzer der beiden Untergruppen sind jeweils in den Konfidenzintervallen der anderen Untergruppe eingeschlossen und unterscheiden sich daher nicht signifikant.

Die Eindrücke der Personen von der Umfrage sind in Tabelle 6 (mittlere Werte) und Tabelle 7 (Häufigkeitsverteilung) aufgeschlüsselt. Die Unterschiede zwischen den Gruppen sind bei jeder Frage gering. Für das Instruktionsverständnis und das Anonymitätsgefühl geben Versuchspersonen in allen Gruppen durchschnittlich hohe Werte an, ebenso wie für das Verständnis des Schutzes der Anonymität (Frage nicht an DQ-Gruppe gestellt). Für die Frage, die auf die Sensitivität des Themas abzielen soll, resultieren in allen Gruppen im Schnitt sehr niedrige Werte.

Tabelle 6. Arithmetische Mittelwerte und Standardabweichungen der Eindrücke zu Instruktionsverständnis, Anonymitätsgefühl, Anonymitätsverständnis und Sensitivität

Frage	**UQM-V**	**UQM**	**DQ**
Instruktionsverständnis: „Wie gut haben Sie verstanden, was Sie auf der vorherigen Seite dieser Umfrage tun sollten?“ (1 = gar nicht, 5 = sehr gut)	4.29 (0.93)	4.40 (0.97)	4.83 (0.51)
Anonymitätsgefühl: „Haben Sie das Gefühl, dass Ihre Anonymität in dieser Umfrage gewährleistet ist?“ (1 = gar nicht, 5 = sehr)	4.55 (0.81)	4.41 (0.91)	4.48 (0.78)
Anonymitätsverständnis: „Wie gut haben Sie verstanden, auf welche Weise Ihre Anonymität in dieser Umfrage geschützt wird?“ (1 = gar nicht, 5 = sehr gut)	4.34 (0.95)	4.18 (1.05)	-
Sensitivität: „Wie unangenehm wäre es für Sie, wenn Sie in einem persönlichen Gespräch nach Alkohol am Steuer gefragt werden würden?“ (1 = gar nicht, 5 = sehr)	1.66 (0.93)	1.69 (1.02)	1.89 (1.21)

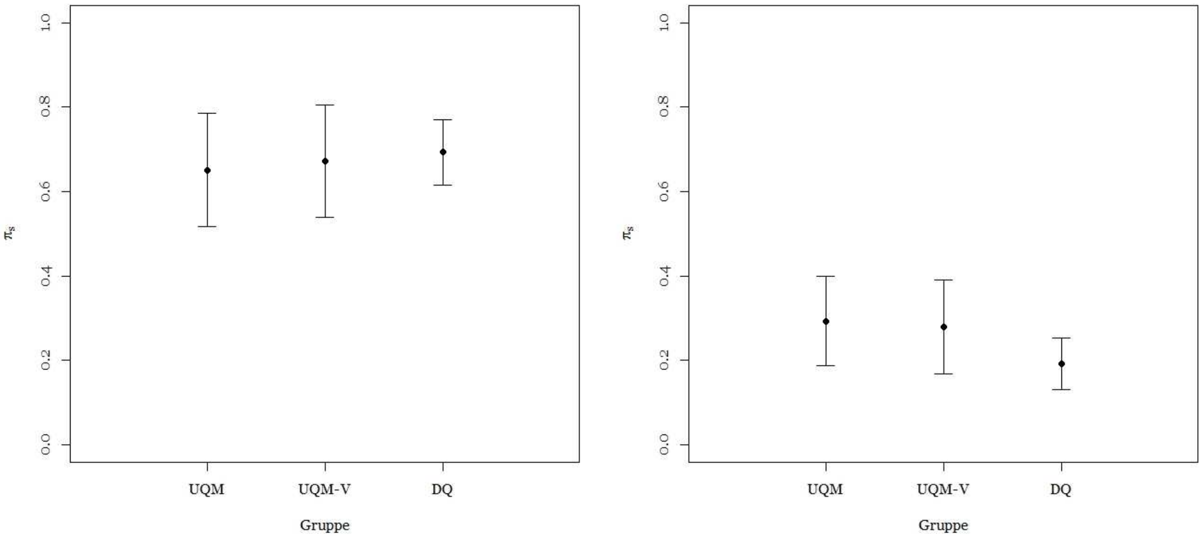
Tabelle 7. Verteilungen der absoluten Häufigkeiten und prozentualen Anteile (pro Gruppe) der Eindrücke zu Instruktionsverständnis, Anonymitätsgefühl, Anonymitätsverständnis und Sensitivität

Frage	**Gruppe**	**Antwortausprägung (1 = gar nicht, 5 = sehr)**				
		1	**2**	**3**	**4**	**5**
Instruktionsverständnis	UQM-V	0/0,0 %	20/7,4 %	30/11,1 %	72/26,6 %	149/55,0 %
	UQM	7/2,4 %	13/4,5 %	21/7,2 %	64/22,1 %	185/63,8 %
	DQ	1/0,3 %	2/0,7 %	6/2,0 %	29/9,7 %	260/87,2 %
Anonymitätsgefühl	UQM-V	4/1,5 %	3/1,1 %	21/7,7 %	56/20,7 %	187/69,0 %
	UQM	6/2,1 %	6/2,1 %	30/10,3 %	69/23,8 %	179/61,7 %
	DQ	1/0,3 %	6/2,0 %	29/9,7 %	76/25,5 %	186/62,4 %
Anonymitätsverständnis	UQM-V	3/1,1 %	15/5,5 %	29/10,7 %	65/24,0 %	159/58,7 %
	UQM	8/2,8 %	17/5,9 %	38/13,1 %	78/26,9 %	149/51,4 %
	DQ	-	-	-	-	-
Sensitivität	UQM-V	157/57,9 %	65/24,0 %	36/13,3 %	9/3,3 %	4/1,5 %
	UQM	175/60,3 %	59/20,3 %	32/11,0 %	19/6,6 %	5/1,7 %
	DQ	161/54,0 %	61/20,5 %	43/14,4 %	13/4,4 %	20/6,7 %

Ob sich die Antwortmuster der Eindrucksfragen zwischen den Gruppen unterscheiden, wurde mit Hilfe von χ^2-Unabhängigkeitstests geprüft. Bei den Fragen nach dem Instruktionsverständnis und dem Anonymitätsgefühl mussten jeweils die Zellen mit Antwortausprägungen 1 und 2 zusammengruppiert werden, um die Voraussetzungen des statistischen Tests zu erfüllen. Es bestehen signifikante Unterschiede in der Verteilung der Antworten bei den Fragen nach Instruktionsverständnis ($\chi^2(6) = 78.43$, $p < .01$) und Sensitivität des Themas ($\chi^2(8) = 21.78$, $p < .01$). Keine Unterschiede liegen in den Antwortverteilungen für die Fragen nach dem Anonymitätsgefühl ($\chi^2(6) = 5.76$, $p = .45$) und dem Anonymitätsverständnis ($\chi^2(4) = 4.47$, $p = .35$) vor.

Außerdem wurde geprüft, ob das Antwortverhalten bei der Frage nach der Sensitivität des Themas („Wie unangenehm wäre es für Sie, wenn Sie in einem persönlichen Gespräch nach Alkohol am Steuer gefragt werden würden?") einen Einfluss auf die Prävalenzschätzungen für die bewusste Inkaufnahme von Alkoholverstößen hat. Dafür wurde aus den Teilnehmenden, welche die Frage mit den Antwortstufen 2 bis 5 beantworteten, eine Gruppe gebildet ($N = 366$). Die restlichen Befragten, die allesamt mit „gar nicht" geantwortet hatten, bildeten die andere Gruppe ($N = 493$). Die Prävalenzen für die erste Gruppe betragen $\pi_{s/UQM} = 0.65$ ($N = 115$, 0,95-KI [0.52; 0.78]), $\pi_{s/UQMV} = 0.67$ ($N = 114$, 0,95-KI [0.54; 0.81]) und $\pi_{s/DQ} = 0.69$ ($N = 137$, 0,95-KI [0.62; 0.77]), die für die zweite Gruppe $\pi_{s/UQM} = 0.29$ ($N = 175$, 0,95-KI [0.19; 0.40]), $\pi_{s/UQMV} = 0.28$ ($N = 157$, 0,95-KI [0.17; 0.39]) und $\pi_{s/DQ} = 0.19$ ($N = 161$, 0,95-KI [0.13; 0.25]). Aus den in Abbildung 3 dargestellten Werten der Konfidenzintervalle geht hervor, dass sich die Prävalenzschätzer innerhalb der Sensitivitätsgruppen nicht signifikant voneinander unterscheiden, während zwischen den Sensitivitätsgruppen jeweils signifikante Unterschiede vorliegen: Die Prävalenzen für die bewusste Inkaufnahme von Alkoholverstößen sind in der Gruppe, die auf die Sensitivitätsfrage mit „gar nicht" antworteten, deutlich niedriger als in der Gruppe derjenigen Teilnehmenden, die die Antwortstufen 2 bis 5 auswählten.

Abbildung 3. Prävalenzschätzer und Konfidenzintervalle für die Gruppen UQM, UQM-V und DQ getrennt nach dem Antwortverhalten bei der Sensitivitätsfrage: „Sensitivität > 1" (links) und „Sensitivität = 1" (rechts)



4. Diskussion

Erwartungsgemäß resultiert in der vorliegenden Studie eine hohe Lebenszeitprävalenz von 44 % für die bewusste Inkaufnahme von Alkoholverstößen im Straßenverkehr bei Studierenden. Fast jede/r Zweite der befragten Studierenden steuerte also schon einmal ein Fahrzeug, obwohl er/sie nach vorherigem Alkoholkonsum nicht sicher war, die entsprechenden Regeln noch einzuhalten. Entgegen der Hypothese 2 unterscheiden sich die Prävalenzschätzer der Gruppen UQM, UQM-V und DQ nicht voneinander. Der Großteil der Befragten aus der UQM-V-Gruppe beantwortete die Übungsfragen korrekt, die Anzahl der richtig beantworteten Übungsfragen scheint indes keinen Einfluss auf die geschätzte Prävalenz zu haben. Während sich das Verständnis der Instruktionen zwischen den Gruppen signifikant unterscheidet, liegt überraschenderweise keine Abweichung in der wahrgenommenen Anonymität und dem Verständnis der Wirkweise der Anonymität zwischen den Gruppen vor. Die Sensitivität des Themas wird anscheinend als gering wahrgenommen. Für die Befragten, die dem Thema eine gewisse Sensitivität zugestehen, fallen die Prävalenzschätzungen deutlich höher aus als für die übrigen Befragten.

4.1 Interpretation der Ergebnisse

Zunächst ist festzuhalten, dass aufgrund der sehr ähnlichen Ergebnisse in allen drei Gruppen eine methodisch bedingte drastische Unter- oder Überschätzung der zu messenden Prävalenz wohl eher unwahrscheinlich ist. Demgegenüber können aber Selektionseffekte nicht ausgeschlossen werden. Da bereits in der E-Mail zur Rekrutierung der Inhalt der Befragung angekündigt wurde, könnten weniger Studierende teilgenommen haben, die aufgrund ihrer Vorerfahrungen Alkohol im Straßenverkehr für ein besonders heikles Thema halten. So könnte eine systematische Unterschätzung der Prävalenz entstanden sein.

Selbstverständlich ist eine studentische Stichprobe nicht repräsentativ für die Gesamtbevölkerung. Dennoch lassen sich ausgehend der vorliegenden Ergebnisse diesbezüglich Vermutungen anstellen. Vieles spricht dafür, dass eine studentische Stichprobe allein aufgrund ihres Durchschnittsalters eine geringere Prävalenz für die bewusste Inkaufnahme von Alkoholverstößen aufweisen sollte: Sowohl Walter (2012) als auch Krüger und Vollrath (1998) berichten tendenziell höhere Prävalenzen mit steigendem Alter (im DRS wurde dieser Befund nur in Unterfranken beobachtet, nicht hingegen in Thüringen, wo kein Alterseffekt auftrat). Auch bei den Alkoholunfällen liegt die Altersgruppe der 18- bis 24-Jährigen, die sicher den Großteil der studentischen Stichprobe ausmacht, unter sechs verschiedenen Altersgruppen nur an vierter Stelle (Statistisches Bundesamt, 2020b). Zudem dürften Studierende aufgrund ihrer allgemeinen Lebenssituation (Wohnort, Einkommen, Mobilität) im Mittel seltener Kraftfahrzeuge führen als der/die durchschnittliche Deutsche, weshalb weniger Gelegenheiten für Alkoholfahrten bestehen. Allerdings können insbesondere männliche Studierende als Angehörige einer „Risikopopulation" betrachtet werden, die überdurchschnittlich häufig leichtfertige Verhaltensweisen zeigen (Hilsenbeck & Löbmann, 1998). Der generelle Alkoholkonsum, welcher bei Studierenden hoch ist (z. B. Bailer et al., 2009; Ganz et al., 2017; Jacobs et al., 2021), gilt außerdem als besserer Prädiktor für Alkoholfahrten als die Häufigkeit der Verkehrsteilnahme (Löbmann et al., 1998). Diejenigen Studierenden, die regelmäßig Kraftfahrzeuge führen, könnten also sogar häufiger mit Alkohol am Steuer sitzen als durchschnittliche Verkehrsteilnehmende. Unabhängig von den angenommenen Unterschieden im erforschten Verhalten impliziert jedoch alleine die Messung einer *Lebenszeit*prävalenz bereits, dass bei einer Durchführung der Befragung in der durchschnittlich älteren Gesamtbevölkerung höhere Ergebnisse zu erwarten sind. Folglich muss bei einer Übertragung der vorliegenden Zahlen auf die deutsche Gesamtbevölkerung auch von einer (wahrscheinlich deutlich) höheren Prävalenz als 44 % für die bewusste Inkaufnahme von Alkoholverstößen ausgegangen werden.

Hinsichtlich der verschiedenen Gruppen tritt ein weder erwartetes noch triviales Ergebnis auf. Eine plausible Interpretation hinsichtlich der DQ-Gruppe ist, dass durch die als gering wahrgenommene Sensitivität höchstens ein geringer Effekt der sozialen Erwünschtheit auftritt. Somit würde nur dann eine Abweichung zu den UQM-Gruppen erwartet werden, wenn in diesen maßgebliche Antwortverzerrungen vorlägen. Bei der UQM-V-Gruppe kann man aus der hohen Anzahl der korrekten Antworten bei den Übungsfragen schließen, dass das Verständnis der Instruktionen generell gut war und daher keine systematischen Antwortverzerrungen existieren.

Schwieriger wird die Interpretation bei der UQM-Gruppe, bzw. bei der Erklärung, warum zwischen UQM- und UQM-V-Gruppe keine Unterschiede resultieren. Folgende Deutungen erscheinen hier möglich: Erstens könnte das Verständnis der Instruktionen in der UQM-Gruppe ähnlich gut sein wie in der UQM-V-Gruppe. Daraus würde sich ableiten, dass die Verständnishilfen in der UQM-V-Gruppe keinen Zweck erfüllen und dass von deren Anwendung aus ökonomischen Gesichtspunkten abgeraten werden muss. Die Beforschung einer studentischen Stichprobe könnte solch ein Ergebnismuster begünstigen, da Studierende möglicherweise besonders motiviert bei der Mitarbeit an Forschungsprojekten sind und über eine hohe kognitive Leistungsstärke verfügen. Dadurch könnten Verständnishilfen für Studierende müßig sein, da sie Zeit kosten, ohne einen Mehrwert zu erbringen. Hierfür spricht auch der hohe Anteil an Befragten der UQM-V-Gruppe, die alle drei Übungsfragen korrekt beantworteten. Dies widerspräche aber den Erkenntnissen aus der Literatur, nach denen Verständnishilfen insbesondere bei Personen mit hohem Bildungsstand effektiv sind (Böckenholt & van der Heijden, 2007; Hoffmann & Musch, 2016; Landsheer, Van Der Heijden & Van Gils, 1999; Wolter, 2012).

Eine zweite Erklärung wäre, dass in der UQM-Gruppe eine beträchtliche Anzahl an falsch-positiven und falsch-negativen Antworten gegeben wurde, die ungefähr gleich häufig auftreten und sich daher „ausgleichen“[5], während in der UQM-V-Gruppe durch die Verständnishilfen validere Antworten gegeben wurden. Daraus würde folgen, dass die Verständnishilfen durchaus einen Zweck erfüllen und die Validität der Antworten erhöhen. Eine Überprüfung dieses Interpretationsansatzes ist mit dem hier angewandten Studiendesign allerdings nicht möglich.

Der Vergleich innerhalb der UQM-V-Gruppe (drei richtige vs. weniger als drei richtige Antworten bei den Übungsfragen) ist schwierig zu deuten und sollte aufgrund der kleinen Unterstichprobe nicht überinterpretiert werden. Man könnte hier darauf schließen, dass die Übungsfragen auch dann zum Verständnis der Instruktionen beitragen, wenn sie falsch beantwortet werden, da den Befragten in einem Feedback erklärt wurde, warum ihre Antwort falsch war.

Bemerkenswert ist das Ergebnis, nachdem sich UQM- und UQM-V-Gruppen sowohl in der wahrgenommenen Anonymität als auch in dem Verständnis der Wirkweise der Anonymisierung nur geringfügig (und nicht signifikant) voneinander unterscheiden. Dies spricht einerseits eher gegen den Zweck der Verständnishilfen. Andererseits fallen die Antworten auf diese Fragen insgesamt sehr positiv aus – ein nicht auftretender Unterschied könnte also Ausdruck einer Art „Deckeneffekt“ sein. Auch denkbar ist, dass durch die Frageformulierung eine Konfundierung vorliegt: Mit den beiden „Anonymitätsfragen“ könnte unter anderem eine Art Vertrauen in die Versuchsleitung und/oder in den rein technischen Schutz der Anonymität durch SoSci Survey (dieses Tool war den meisten Teilnehmenden mutmaßlich bereits bekannt) gemessen werden. Bei der Frage nach dem Verständnis der Instruktionen ist es wenig überraschend, dass die DQ-Gruppe die höchsten Werte erzielt. Diese Beobachtung reiht sich in mehrere Befunde ein (z. B. Coutts & Jann, 2011; Hoffmann & Musch, 2016; Landsheer, Van Der Heijden & Van Gils, 1999), nach denen DQ-Methoden verständlicher sind als RRMs.

[5] Dass solche Ergebnismuster bei RRMs grundsätzlich auftreten können, zeigten bereits Höglinger und Diekmann (2017).

Ein zentraler Punkt der Studie ist die Frage nach der Sensitivität. Zunächst erscheint es recht eindeutig, dass das Thema als eher weniger heikel wahrgenommen wird – auch angesichts der hohen Prävalenzschätzer. Die Formulierung des Items könnte indessen zu einer Konfundierung geführt haben: „Wie unangenehm wäre es für Sie, wenn Sie in einem persönlichen Gespräch nach Alkohol am Steuer gefragt werden würden?“. Diese Frage beinhaltet möglicherweise eine „indirekte direkte Frage“ nach dem eigenen Verhalten – sie dürfte Personen eher weniger unangenehm sein, die mit Alkohol am Steuer bisher nichts zu tun hatten. Die Aufteilung der Befragten nach Antwortstufen (1, also die niedrigste Stufe, vs. 2-5) bestätigt diese Vermutung. Demnach sind über alle drei Gruppen hinweg deutliche Unterschiede in den Prävalenzschätzern nach Antwort auf die „Sensitivitätsfrage“ vorzufinden. Die naheliegende Erklärung für diesen Befund lautet also: Diejenigen Personen, die noch nie oder sehr selten mit Alkohol am Steuer saßen, antworten häufiger mit der Antwortstufe 1, während Personen, die mit einer höheren Wahrscheinlichkeit das Verhalten schon einmal gezeigt haben, eher mit 2-5 antworten.

4.2 Limitationen

In der vorliegenden Arbeit wurde mit der bewussten Inkaufnahme von Alkoholverstößen ein neues Konstrukt erforscht. Im Großteil der Arbeiten und Statistiken, die das Thema „Alkohol im Straßenverkehr“ zum Gegenstand haben, wird „Alkohol am Steuer“ objektiver bzw. eindeutiger definiert als in dieser Studie: Meist als die Überschreitung des zulässigen Grenzwertes, deren Vorliegen durch eine Messung der BAK beurteilt wird. Das hier untersuchte Verhalten ist von einer objektiven Regelverletzung klar abzugrenzen: Das Steuern eines Fahrzeuges trotz bestehender Unsicherheit, ob die BAK noch unter dem Grenzwert liegt, ist nicht automatisch ein Regelverstoß – dieser wird jedoch, ebenso wie ein erhöhtes Unfallrisiko, bei dem hier untersuchten Verhalten billigend in Kauf genommen. Der Untersuchungsgegenstand deckt somit ein breiteres Spektrum problematischen Verhaltens im Straßenverkehr ab, wobei die Überschreitung des BAK-Grenzwertes als Teilmenge beinhaltet ist.

Diese Ausweitung der Begrifflichkeit geht zum einen selbstverständlich mit einer größeren Unschärfe in Bezug auf das gemessene Konstrukt einher. Das zeigt sich bereits daran, dass sich keine Aussage darüber treffen lässt, wie viele der Studierenden, welche angaben, Alkoholverstöße bereits bewusst in Kauf genommen zu haben, auch tatsächlich schon einmal über dem zulässigen Grenzwert lagen – hier bilden die ermittelten 44 % lediglich eine Obergrenze ab. Überdies könnte eine Konfundierung mit dem Trinkverhalten entstanden sein, da Menschen die öfter trinken auch eher glauben, nach dem Alkoholkonsum noch fahrtüchtig zu sein (Löbmann et al., 1998) – die Prävalenz von „Vieltrinkenden“ könnte dadurch unterschätzt, die von „Wenigtrinkenden“ überschätzt worden sein. Gegebenenfalls ist dieser Zusammenhang bei Studierenden mit wenig Fahrerfahrung sogar noch verstärkt. Darüber hinaus könnte die gemessene niedrige Sensitivität mit der breiten Begriffsauffassung zusammenhängen, was der Anwendbarkeit von RRMs abträglich wäre und so die Erforschung des Einflusses von Verständnis im UQM erschweren würde. Unter Umständen wurde die Frage zur Sensitivität des Themas nicht optimal gewählt (eventuell Konfundierung mit der Eigenschaft als Merkmalsträger*in, s. o.). Wie heikel das Thema tatsächlich ist, kann nur schwer beurteilt werden. Die Frage nach der Sensitivität ist zentral – wenn der bewussten Inkaufnahme von Alkoholverstößen keine Sensitivität beigemessen wird, sind RRMs hier generell ungeeignet und DQ stets vorzuziehen. In nachfolgenden Studien sollte daher eine Umformulierung erwogen werden, etwa, indem nach einer Beurteilung der „gesellschaftlichen Verwerflichkeit“ des Verhaltens gefragt wird (vgl. Krüger et al., 1998).

Zum anderen kann durch das weitere Konstrukt jedoch das Dunkelfeld des riskanten Verhaltens, welches auch dem gängigen Alkoholverstoß zu Grunde liegt, besser beurteilt werden. Eine präzise und objektive Messung der BAK in einem Fragebogen ist indes kaum möglich. Da die meisten objektiven BAK-Messungen wohl im Rahmen von Verkehrskontrollen stattfinden, würden bei Fragen nach sicheren Alkoholverstößen neben der Erfassung eindeutiger Trunkenheitsfahrten nur die öffentlich zugänglichen Statistiken reproduziert. Für die gewählten Methoden ist das hier gewählte breitere Konstrukt vermutlich also sogar besser geeignet, um das Dunkelfeld zu Alkohol am Steuer aufzuhellen.

Eine Limitation der Untersuchung liegt per Design in der Unmöglichkeit, falsch-positive von falsch-negativen Antworten zu unterscheiden. Über die obigen Mutmaßungen und Erklärungsversuche hinaus können daher keine neuen Erkenntnisse über die Wirkweise von Verständnishilfen bei Anwendung des UQM gewonnen werden. Insoweit ermöglichte das Studiendesign keine hinreichende Beantwortung der Frage nach der Rolle von Verständnis im UQM. Zugunsten der Bestimmung praxisbezogener Prävalenzen wurde diese Limitation allerdings in Kauf genommen. Des Weiteren ist mit dem Befund, dass zwischen den Gruppen keine *offensichtlichen* Unterschiede in den Prävalenzschätzern zu beobachten sind, durchaus ein Erkenntnisgewinn verbunden: Allein durch Verständnishilfen im UQM per se können keine Veränderungen in dem Antwortverhalten der Befragten erwartet werden; um die Rolle von Verständnis zu klären, müssen in zukünftigen Forschungsdesigns repräsentativere Stichproben untersucht und mehrere zusätzliche Variableneinflüsse geprüft werden (allen voran Bildungsniveau und Sensitivität der Fragestellung).

Bezüglich der Auswahl des UQM als indirekte Fragemethode gibt es Vor- und Nachteile. Insbesondere zur Überprüfung der Auswirkungen von Verständnisfragen existieren zumindest zwei möglicherweise besser geeignete RRTs: Das Cheater Detection Model (CDM; Clark & Desharnais, 1998) und eine jüngst vorgestellte „Cheating“-Erweiterung des UQM (UQMC; Reiber, Pope & Ulrich, 2020). In beiden Modellen ist die Schätzung eines „Cheating“-Parameters möglich, der den Anteil der Befragten repräsentiert, die stets mit einem selbstschützenden „Nein“ antworten. Diese selbstschützenden Antworten sollten mit größerem Verständnis der Fragemethode seltener sein, weshalb diese Modelle gut geeignet wären, um den Einfluss von Verständnishilfen zu prüfen. Das UQM ist jedoch deutlich simpler und dadurch – rein rechnerisch – auch das effizienteste der drei Modelle (Ulrich et al., 2012). Das CDM und das UQMC ermöglichen zwar eine realitätsnähere Beschreibung des Antwortverhaltens der Befragten, für eine Schätzung der zusätzlichen Cheating-Parameter sind aber deutlich größere Stichproben nötig (Reiber, Pope & Ulrich, 2020; Ulrich et al., 2012). Da in der vorliegenden Studie kein hinreichend großer Rücklauf garantiert werden konnte, wurde das UQM für das hier angewandte Design als am geeignetsten betrachtet. Weil das UQM eine weit verbreitete RRT-Anwendung ist, wurde es dennoch für wertvoll erachtet, die Rolle für das Verständnis in diesem Modell zu beforschen. Außerdem wurde – im Falle von eindeutigeren Ergebnissen – die Möglichkeit eines Erkenntnistransfers vom UQM auf die anderen Modelle für naheliegender eingeschätzt als umgekehrt.

In einer Folgestudie, die den Einfluss von Verständnishilfen beim CDM oder UQMC erforscht, sollten im Vorfeld Poweranalysen durchgeführt werden, um eine optimale Verteilung der Stichprobe auf die einzelnen Gruppen zu gewährleisten. Denn DQ-Gruppen müssen für eine gleiche Teststärke aufgrund ihrer größeren Effizienz nicht so groß sein wie RRM-Gruppen. Auch beim Vergleich von DQ und UQM ist die Aufteilung der Befragten in drei gleich große Gruppen daher nicht optimal. Aufgrund der a priori schwer einzuschätzenden Prävalenz (von der die Teststärke im UQM unter anderem abhängt; Ulrich et al., 2012) und dem nicht absehbaren Rücklauf wurde hier von einer Poweranalyse abgesehen. Die hier ermittelte Prävalenz kann als Grundlage für die Poweranalysen in nachfolgenden Forschungsprojekten dienen.

Wie bereits erwähnt, ist auch die Generalisierbarkeit durch die aus forschungsökonomischen Gründen gewählte selektive Stichprobe begrenzt. Die Lebenszeitprävalenz von 44 % für die bewusste Inkaufnahme von Alkoholverstößen im Straßenverkehr kann nur als Untergrenze für die Prävalenz in einer repräsentativen Stichprobe betrachtet werden (s. o.), es kann aber nur spekuliert werden, wie groß genau die Abweichung zwischen Studierenden und Gesamtbevölkerung ist. Daneben können, wie oben erwähnt, einige weitere Einflüsse durch die selektive Stichprobe nicht ausgeschlossen werden, beispielsweise eine erhöhte Motivation, sich mit den komplexen Instruktionen auseinanderzusetzen und ein sehr hohes Anonymitätsgefühl durch hohe Vertrautheit mit Online-Umfragen. Dadurch könnte die Auftretenswahrscheinlichkeit eines Effekts der sozialen Erwünschtheit minimiert worden sein, was wiederum die Vorteile der RRMs gegenüber DQ annulliert haben könnte. Um diese möglichen Einflüsse zu kontrollieren und den Effekt von Verständnishilfen besser überprüfen zu können, sollten Folgestudien auf diesem Gebiet also – wenn möglich – repräsentative Stichproben beforschen.

Von Relevanz ist zudem, dass bei Alkoholfahrten das Geschlecht eine äußerst wichtige Einflussgröße ist: Männer nehmen deutlich häufiger alkoholisiert am Straßenverkehr teil als Frauen und verursachen mehr Alkoholunfälle (z. B. Krüger & Vollrath, 1998; Statistisches Bundesamt 2020a, 2020b; Walter, 2012). In der vorliegenden Studie wurde bewusst von einer Erhebung demographischer Daten abgesehen, da deren Einflüsse nicht Bestandteil der untersuchten Fragestellung waren. Nichtsdestotrotz spielen Geschlechts- und möglicherweise auch Altersunterschiede eine beträchtliche Rolle in dem Forschungsfeld, weshalb in zukünftigen Forschungsarbeiten auf dem Gebiet die entsprechenden Variablen miterhoben werden sollten.

4.3 Zukünftige Forschung

Zwei weiterführende Forschungsdesigns ergeben sich nicht zuletzt auch aus den aufgeführten Limitationen dieser Studie: Um die Wirkweise der Verständnishilfen beim UQM besser zu verstehen, sollte eine starke Validierungsstudie mit einer repräsentativen Stichprobe durchgeführt werden.[6] Ein solches Design würde die Überprüfung der oben aufgestellten Erklärungsansätze für die ähnlichen Prävalenzen der UQM- und UQM-V-Gruppen ermöglichen. Eine weitere, praxisnähere Folgestudie könnte die Fragestellung zu Alkohol am Steuer beibehalten, jedoch sollte bei gleichzeitiger Überprüfung der Rolle von Verständnis im UQM die Wahl auf dessen Erweiterung, das UQMC (Reiber, Pope & Ulrich, 2020), fallen. Außerdem sollte erwogen werden, anstatt der Lebenszeitprävalenz aktuelleres Verhalten, z. B. in Form der Monatsprävalenz, zu erfragen. Für den kürzeren Zeitraum wäre auch die Erhebung relevanter Variablen wie der derzeitigen Trink- und Fahrgewohnheiten der Befragten denkbar. Die wichtige Frage nach der Sensitivität des Themas sollte verbessert bzw. angepasst werden (s. o.). Nach Möglichkeit sollte die Befragung einer repräsentativen Stichprobe erfolgen. Somit würden

[6] Die Anwendung des UQMC anstelle des UQM ist in einer starken Validierungsstudie nicht unbedingt nötig, da bereits per Design falsch-negative Antworten geschätzt werden können.

wichtige Schlüsse auf die Gesamtbevölkerung ermöglicht. Ebenso könnte der Bildungsstand bei der Frage nach den Unterschieden zwischen DQ, UQM(C) und UQM(C)-V in den Mittelpunkt rücken. Eine interessante Variable wären auch die Wohnorte der Befragten: Es steht zu vermuten, dass durch die höhere Verfügbarkeit öffentlicher Verkehrsmittel in urbanen Gebieten im Zweifelsfall eher auf eine Fahrt im Anschluss an den Alkoholkonsum verzichtet werden kann. Möglicherweise existiert dadurch ein erhebliches Stadt-Land-Gefälle bei der Prävalenz von Alkohol am Steuer. Zuletzt wäre im Rahmen einer solchen Folgestudie die Einbeziehung möglicher Geschlechtereffekte in das Forschungsdesign unabdingbar. Bei beiden skizzierten Folgestudien sollte im Voraus eine Poweranalyse auf Grundlage der hier ermittelten Prävalenz durchgeführt werden, sodass eine möglichst effiziente Aufteilung der Teilnehmenden auf die Gruppen und eine ausreichende Teststärke erreicht werden können.

Trotz der oben aufgezählten Limitationen stellt die hier präsentierte Studie wichtige und neue Erkenntnisse vor, sowohl in Bezug auf Alkohol im Straßenverkehr als auch bei der Erforschung der Rolle von Verständnis in RRMs. Aus forschungsökonomischen Gründen war es nach hiesiger Auffassung sinnvoll, sich zunächst mit einem simplen Design und einer studentischen Stichprobe an das Forschungsgebiet heranzuwagen, obgleich die Repräsentativität der Studie dadurch begrenzt ist. Für ein besseres Verständnis der Funktionsweisen des UQM und von RRMs allgemein sind weiterführende Forschungsarbeiten unabdingbar. Mit dem UQMC (Reiber, Pope & Ulrich, 2020) steht hierfür ein besser geeignetes methodisches Werkzeug zur Verfügung, sofern eine hinreichend große Stichprobe beforscht werden kann. Doch auch die praktischen Anwendungen von RRMs sollten nicht außer Acht gelassen werden. Bei der bewussten Inkaufnahme von Alkoholverstößen im Straßenverkehr handelt es sich um ein bedeutsames und potentiell heikles Thema von großem gesellschaftlichem Interesse. Der hier vorgestellte Befund zeigt, dass Alkohol am Steuer noch immer weit verbreitet ist – auch bei jungen Menschen. Die hier gefundene hohe Prävalenz stützt die anfangs geäußerte These, dass die leicht alkoholisierte Verkehrsteilnahme bisweilen als „Kavaliersdelikt“ betrachtet und verharmlost wird. Weitere Forschung auf dem Gebiet ist nötig, um Alkohol im Straßenverkehr besser verstehen und die Präventionsarbeit weiter verbessern zu können.

Literaturverzeichnis

Alkohol? Kenn dein Limit. (2021). *Alkohol am Steuer – Warum Alkohol im Straßenverkehr tabu ist*. www.kenn-dein-limit.de/alkoholverzicht/alkohol-am-steuer/ (2021, Juni 22).

Bailer, J., Stübinger, C., Dressing, H., Gass, P., Rist, F., & Kühner, C. (2009). Zur erhöhten Prävalenz des problematischen Alkoholkonsums bei Studierenden. *Psychotherapie Psychosomatik Medizinische Psychologie, 59*(09/10), 376-379. doi.org/10.1055/s-0029-1215596

Beasley, E. E., Beirness, D. J., & Beirness & Associates. (2012). *Alcohol and drug use among drivers following the introduction of immediate roadside prohibitions in British Columbia: findings from the 2012 roadside survey*. https://www2.gov.bc.ca/assets/gov/driving-and-transportation/driving/roadsafetybc/data/bc-roadside-report2012.pdf (2021, Juni 22).

Berning, A., Compton, R., & Wochinger, K. (2015). Results of the 2013-2014 national roadside survey of alcohol and drug use by drivers. *Journal of Drug Addiction, Education, and Eradication*, *11*(1), 47-57.

Böckenholt, U., & Van der Heijden, P. G. (2007). Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. *Psychometrika*, *72*(2), 245-262. doi.org/10.1007/s11336-005-1495-y

Buchman, T. A., & Tracy, J. A. (1982). Obtaining responses to sensitive questions: conventional questionnaire versus randomized response technique. *Journal of Accounting Research*, 263-271.

Bund gegen Alkohol und Drogen im Straßenverkehr e.V. (2011). *Alkohol und Drogen im Straßenverkehr.* https://www.bads.de/media/1293/2_b_alkohol_und_drogen_im_stra_enverkehr.pdf (2021, Juni 22).

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, 109. doi.org/10.1016/j.jml.2019.104047

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, *36*(2), 84-95.

Clark, S. J., & Desharnais, R. A. (1998). Honest Answers to Embarrassing Questions: Detecting Cheating in the Randomized Response Model. *Psychological Methods*, *3*(2), 160-168.

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research*, 40(1), 169-193. doi.org/10.1177/0049124110390768

DER SPIEGEL. (2008, 16. September). *Beckstein findet zwei Maß Bier vertretbar*. www.spiegel.de/auto/aktuell/alkohol-am-steuer-beckstein-findet-zwei-mass-bier-vertretbar-a-578446.html (2021, Juni 22).

De Schrijver, A. (2012). Sample survey on sensitive topics: Investigating respondents' understanding and trust in alternative versions of the randomized response technique. *Journal of Research Practice*, *8(*1), 1-17.

Ganz, T., Braun, M., Laging, M., & Heidenreich, T. (2017). Erfassung des riskanten Alkoholkonsums bei Studierenden deutscher Hochschulen. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *46*(3), 187-197. doi.org/10.1026/1616-3443/a000432

Greenberg, B. G., Abul-Ela, A. L. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, *64*(326), 520-539.

Hilsenbeck, T., & Löbmann, R. (1998). Das Risikoprofil alkoholauffälliger Verkehrsteilnehmer zwischen 18 und 30 Jahren. In H.-P. Krüger (Hrsg.), *Fahren unter Alkohol in Deutschland* (S. 87-106). Stuttgart: Gustav Fischer.

Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques: A Stochastic Lie Detector versus the Crosswise Model. *Behavior Research Methods*, *48*(3), 1032-1046. doi.org/10.3758/s13428-015-0628-6

Houwing, S., Hagenzieker, M., Mathijssen, R., Bernhoft, I. M., Hels, T., Janstrup, K., ... & Verstraete, A. (2011a). *Prevalence of alcohol and other psychoactive substances in drivers in general traffic, part I: General results*. biblio.ugent.be/publication/1988541/file/1988562 (2021, Juni 22).

Houwing, S., Hagenzieker, M., Mathijssen, R., Bernhoft, I. M., Hels, T., Janstrup, K., ... & Verstraete, A. (2011b). *Prevalence of alcohol and other psychoactive substances in drivers in general traffic, part II: country reports*. biblio.ugent.be/publication/1988588/file/1988629 (2021, Juni 22).

Höglinger, M., & Diekmann, A. (2017). Uncovering a blind spot in sensitive question research: false positives undermine the crosswise-model RRT. *Political Analysis*, *25*(1), 131-137. doi.org/10.1017/pan.2016.5

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *PloS one*, *13*(8): e0201770, 1-22. doi.org/10.1371/journal.pone.0201770

Ingsathit, A., Woratanarat, P., Anukarahanonta, T., Rattanasiri, S., Chatchaipun, P., Wattayakorn, K., Lim, S. &Suriyawongpaisal, P. (2009). Prevalence of psychoactive drug use among drivers in Thailand: a roadside survey. *Accident Analysis & Prevention*, *41*(3), 474-478. doi.org/10.1016/j.aap.2009.01.010

Jacobs, T., Linke, M., Richter, E. P., Drössler, S., Zimmermann, A., & Berth, H. (2021). Alkoholkonsum bei Medizinstudierenden. Eine Analyse über 7 Jahre (2011 bis 2017). *Prävention und Gesundheitsförderung*, 1-7. doi.org/10.1007/s11553-021-00877-2

John, U., & Hanke, M. (2018). Trends des Tabak-und Alkoholkonsums über 65 Jahre in Deutschland. *Das Gesundheitswesen*, *80*(2), 160-171. doi.org/10.1055/s-0043-110854

Kraftfahr-Bundesamt (2021). *Verkehrsauffälligkeiten – Zahlen im Überblick*. https://www.kba.de/DE/Statistik/Kraftfahrer/Verkehrsauffaelligkeiten/verkehrsauffaelligkeiten_node.html (2021, September 23).

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity*, *47*(4), 2025-2047. doi.org/10.1007/s11135-011-9640-9

Krüger, H.-P. (Hrsg.) (1998). *Fahren unter Alkohol in Deutschland*. Stuttgart: Gustav Fischer.

Krüger, H.-P., Schöch, H., Vollrath, M., & Löbmann, R. (1998). Die Auswirkung der Erhöhung der Promillegrenze: Quantitative Überprüfungen. In H.-P. Krüger (Hrsg.), *Fahren unter Alkohol in Deutschland* (S. 121-160). Stuttgart: Gustav Fischer.

Krüger, H.-P., & Vollrath, M. (1998). Fahren unter Alkohol in Deutschland. Die Ergebnisse des Deutschen Roadside Surveys. In H.-P. Krüger (Hrsg.), *Fahren unter Alkohol in Deutschland* (S. 33-57). Stuttgart: Gustav Fischer.

Kulemeier, R. (1991). *Fahrverbot (§ 44 StGB) und Entzug der Fahrerlaubnis (§§ 69 ff. StGB): ein Beitrag zum Verhältnis dieser Sanktionsformen und zum vikariierenden System von Strafen und Massregeln im Verkehrsstrafrecht*. Lübeck: Max Schmidt-Römhild.

Landsheer, J. A., Van Der Heijden, P., & Van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality and Quantity*, *33*(1), 1-12.

Leiner, D. J. (2019a). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, *13*(3), 229-248. doi.org/10.18148/srm/2019.v13i3.7403

Leiner, D. J. (2019b). *SoSci Survey (Version 3.1.06) [Computer software]*. München: SoSci Survey GmbH.

Lensvelt-Mulders, G. J., Hox, J. J., Van der Heijden, P. G., & Maas, C. J. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research*, *33*(3), 319-348. doi.org/10.1177/0049124104268664

Lensvelt-Mulders, G. J., & Boeije, H. R. (2007). Evaluating compliance with a computer assisted randomized response technique: a qualitative study into the origins of lying and cheating. *Computers in Human Behavior*, *23*(1), 591-608. doi.org/10.1016/j.chb.2004.11.001

Löbmann, R., Krüger, H.-P., Vollrath, M., & Schöch, H. (1998). Zur Phänomenologie der Alkoholfahrt. In H.-P. Krüger (Hrsg.), *Fahren unter Alkohol in Deutschland* (S. 59-86*)*. Stuttgart: Gustav Fischer.

Meisters, J., Hoffmann, A., & Musch, J. (2020). Can detailed instructions and comprehension checks increase the validity of crosswise model estimates?. *PloS one*, *15*(6): e0235403, 1-19. doi.org/10.1371/journal.pone.0235403

Musch, J., & Rösler, P. (2011). Schnell-Lesen: Was ist die Grenze der menschlichen Lesegeschwindigkeit?. In M. Dresler (Hrsg.), *Kognitive Leistungen* (S. 89-106). Heidelberg: Spektrum Akademischer Verlag.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263-280. doi.org/10.1002/ejsp.2420150303

Reiber, F., Pope, H., & Ulrich, R. (2020). Cheater detection using the unrelated question model. *Sociological Methods & Research*, 1-23. doi.org/10.1177/0049124120914919

Schöch, H. (2001). Alkohol im Straßenverkehr. *Neue Kriminalpolitik*, *13*(1), 28-31.

Statistisches Bundesamt. (2020a). Fachserie 8 Reihe 7 – Verkehrsunfälle 2019. www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/Publikationen/Downloads-Verkehrsunfaelle/verkehrsunfaelle-jahr-2080700197004.pdf?__blob=publicationFile (2021, Juni 22).

Statistisches Bundesamt. (2020b). *Verkehrsunfälle – Unfälle unter dem Einfluss berauschender Mittel im Straßenverkehr 2019*. www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/Publikationen/Downloads-Verkehrsunfaelle/unfaelle-alkohol-5462404197004.pdf?__blob=publicationFile (2021, Juni 22).

Süddeutsche Zeitung. (2008, September 17). *Darf Beckstein noch fahren?* www.sueddeutsche.de/bayern/nach-zwei-mass-bier-darf-beckstein-noch-fahren-1.703819 (2021, Juni 22).

Ulrich, R., Schroter, H., Striegel, H., & Simon, P. (2012). Asking sensitive questions: A statistical power analysis of randomized response models. *Psychological Methods*, *17*(4), 623-641. doi.org/10.1037/a0029314

Umesh, U. N., & Peterson, R. A. (1991). A critical evaluation of the randomized response method: Applications, validation, and research agenda. *Sociological Methods & Research*, *20*(1), 104-138.

Vanlaar, W. (2005). Drink driving in Belgium: results from the third and improved roadside survey. *Accident Analysis & Prevention*, *37*(3), 391-397. doi.org/10.1016/j.aap.2004.12.001

Walter, M. (2012). *A new methodological approach to assess drug driving – The German Smartphone Survey.* opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/6385/file/WalterMartinaDiss.pdf (2021, Juni 22).

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63-69.

Wolter, F. (2012). *Heikle Fragen in Interviews – Eine Validierung der Randomized Response-Technik*. Wiesbaden: Springer VS.

Yu, J. W., Tian, G. L., & Tang, M. L. (2008). Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika*, *67*(3), 251-263. doi.org/10.1007/s00184-007-0131-x

Kontakt | Contact

Benedikt Iberl | Universität Tübingen | Institut für Kriminologie | [redacted]