International Journal of Drug Policy 73 (2019) 228–234

Contents lists available at ScienceDirect

International Journal of Drug Policy

journal homepage: www.elsevier.com/locate/drugpo

Research Paper

Surveying drug consumption: Assessing reliability and validity of the European Web Survey on Drugs questionnaire

Kateřina Škařupová[a,*], Nicola Singleton[b], João Matias[b], Viktor Mravčík[c,d,e]

[a] *Institute for Research on Children, Youth and Family, Faculty of Social Studies, Masaryk University, Jostova 10, 60200 Brno, Czech Republic*
[b] *European Monitoring Centre for Drugs and Drug Addiction (EMCDDA), Praça Europa 1, Cais do Sodré, 1249-289 Lisbon, Portugal*
[c] *Department of Addictology, First Faculty of Medicine, Charles University and General University Hospital in Prague, Apolinarska 4, 128 00 Praha 2, Czech Republic*
[d] *National Monitoring Centre for Drugs and Addiction, Office of the Government, nábřeží Edvarda Beneše 4, 118 01 Prague 1, Czech Republic*
[e] *National Institute of Mental Health, Topolová 748, 250 67 Klecany, Czech Republic*

ARTICLE INFO

Keywords:
Web survey
Test/re-test reliability
Prevalence
Patterns of use

ABSTRACT

Background: The European Web Survey on Drugs aimed to obtain in-depth data on consumption of cannabis, ecstasy/MDMA, cocaine, and amphetamines in different populations of drug users in 16 European countries. This paper examines test-retest reliability, the consistency and the comprehensibility of the prevalence and frequency of use questions in the Czech part of the survey.

Methods: A baseline web survey was performed (N = 610) with follow-up data collection in a sub-sample of volunteers providing email addresses (N = 158). The baseline sample was self-selecting, responding to advertisements made available through multiple channels designed to attract diverse samples of drug users. Test-retest analysis was conducted for core questionnaire items.

Results: Respondents to the follow-up were predominantly socially integrated; 91% reported last year cannabis use, 42% used Ecstasy/MDMA, 23% amphetamines, and 27% reported cocaine use. Test-retest reliability was rated moderate to good (reliability coefficients between 0.55–0.87) for most prevalence items with sufficient sample sizes. Items assessing frequency of use were more reliable for most substances when asking about the exact number of days used, compared to categorical items that implicitly assume a regular pattern of use and were interpreted differently by different respondents.

Conclusions: Simplicity and unambiguity of questions increase the reliability of results. Tools measuring drug consumption need to take into consideration the irregularity of drug using patterns. Question testing is important to increase validity and support a correct interpretation of the data.

Introduction

General population surveys are one of the key methods used to monitor the drug situation and trends, and to inform policies. In many countries, representative surveys are conducted on a regular basis and form part of the minimum requirements for standard drug monitoring, in accordance to the framework of key epidemiological indicators developed by the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) for EU monitoring systems (see http://www.emcdda.europa.eu/activities/key-indicators). Data from surveys often complement other drug-related indicators (such as law-enforcement or criminological data) and feed into secondary estimates (such as market size estimates). However, some forms of drug use may be under-represented in samples of the general population, and the depth of the information collected is often limited to basic prevalence (Johnson 2014; Enghoff & Aldridge, 2019). For instance, estimates of the size of the retail market, due to lack of detailed data, must often use assumptions concerning the quantities of drugs consumed by individuals. These often need to be based on extrapolation of frequency of use in the past 30 days to the whole year from general population survey averages, but these assumptions may not reflect the complexity of drug using patterns and the differences between groups of users (Kilmer et al. 2011). Most intensive users, for example, are reported to consume the largest quantities and are thus responsible for the greatest share of the drug market (Korf et al. 2007, Trautmann et al. 2013, Caulkins et al. 2015).

Targeted and web surveys have the potential to fill this gap and provide data on various quantitative aspects of drug markets such as frequency of use, amounts of consumed drugs, street prices, and purchase behaviours from various drug using populations (Ritter 2006).

[*] Corresponding author.
E-mail address: skarupovakat@gmail.com (K. Škařupová).

https://doi.org/10.1016/j.drugpo.2019.03.005

Web surveys, while not aiming for representativeness, have been successfully implemented when targeting particular groups of drug users and other specific populations. The Global Drug Survey has become influential by attracting large self-nominated samples of drug users. However, little is known about the properties and development of its tools (Barratt et al. 2017). Online data collection can be used in many different ways and generally appears to have no or little impact on reliability, or on social desirability bias in responses (Denscombe, 2006; Khadjesari et al., 2009; Dodou and de Winter, 2014). Some authors warn that data quality may be jeopardized in web surveys, and it is an issue that needs to be borne in mind. However, researchers operating in the traditional face to face context also have to trust that their respondents provide true answers, and may need to use similar strategies to detect mischievous participants, such as cross-checking answers to repeated items, or the use of a non-existing substance (Miller and Sønderlund 2010). In research involving hidden populations of people who use drugs where sampling frames are not available and representativeness is not the key requirement of the sample, the advantages of web surveys may outweigh the disadvantages (Miller and Sønderlund 2010).

Nevertheless, the number of studies collecting detailed information on consumption and purchasing behaviours from targeted samples of drug users is relatively small. As is the case in other areas, such as alcohol consumption and diet, such surveys may suffer from validity and reliability issues related to the inability of individuals to judge their consumption habits with precision. The timeline follow-back method, developed for marketing and consumption research, has been used to assist respondents to recall past month consumption of cannabis and other drugs (Robinson et al., 2014). Assessing amounts of the substance used, however, remained a challenge as the method assumes regular doses and, incorrectly, extrapolates recent use, typically in the last week or last month, to broader time frames (Hoeppner et al., 2010; Trautmann et al. 2013).

Attempts have been made to assess the extent and nature of such problems and to develop methods to facilitate and guide people towards the provision of correct responses in surveys. In this respect, cannabis is the most widely researched substance. The unit 'standard joint' has been widely adopted to determine amounts of cannabis used, yet it assumes identical amounts of the drug across users (Zeisser et al. 2012). Alternative probing methods have been developed, including prompt cards displaying amounts of the drug to guide respondent's answers (Trautmann et al. 2013; Cuttler and Spradlin, 2017), or the weighing of surrogate substances, cannabis substitutes, or actual cannabis (Mariani et al. 2011, Norberg et al. 2012, van der Pol et al., 2013). Surrogate substances and other cannabis substitutes often fail to mimic the appearance of the common types of cannabis, while use of the actual drug may be legally problematic in many countries. Using prompt cards appears most suitable when surveying large (online) samples, as they do not involve weighing and manipulation of the substance. However, the reliability of this method has yet to be assessed (van der Pol et al., 2013).

The European Web Survey on Drugs (EWSD; see http://www.emcdda.europa.eu/activities/european-web-survey-on-drugs_en) was established initially to collect data from international samples of people who use drugs on their drug consumption and purchasing behaviours in order to introduce more precision into demand-side estimates of the size of European drug markets. The sub-study presented here, undertaken in the Czech Republic, involved a baseline and a follow-up data collection and aimed at assessing the test/re-test reliability of the EWSD questionnaire, including prompt cards, and cognitive interviews to explore how individuals understand and respond to particular questions, and how this might affect their reliability and validity. In this paper, we focus primarily on test-retest reliability, but also examine the consistency of responses to two different question formats for frequency of use in the last year, the consistency of reported patterns of use in the past month, and use the knowledge from the cognitive interviews to support the interpretation of results.

Methods

Data collection

The questionnaire used in a study coordinated by the Trimbos Institute (Trautmann et al., 2013) was adapted for this project and covered four substances: cannabis (herbal and resin), cocaine powder, ecstasy/MDMA, and amphetamines (amphetamine/speed or methamphetamine). Data were collected using LimeSurvey software hosted by LimeService. Respondents were recruited using multiple solicitation techniques to capture diverse drug-using subpopulations. People who use cannabis and those that use drugs in nightlife/recreational settings were addressed via online advertisements on Facebook, on websites associated with nightlife culture and cannabis legalisation initiatives, and in the closed online forums where users share their experiences (such as nyx.cz, which hosts the most prominent online rooms for Czech psychonauts). People who use amphetamines (mainly methamphetamine) in more problematic ways were approached via needle and syringe exchange programmes.

Within the baseline (test) condition, data were collected over a period of six weeks between 29 February and 12 April 2016. The final sample comprised 610 respondents, of which 231 (37.9%) agreed to participate in the follow-up and provided an e-mail address for this purpose. The re-test exposed respondents to the identical questionnaire, but the number of socio-demographic items was reduced. In total, 158 respondents (68.4% of those who agreed to follow-up; 25.9% of the baseline sample) participated in the follow-up. There were no statistically significant differences between those who responded to the follow-up and those who only participated at the baseline (N = 452) in terms of age, gender, or last year use of cannabis, cocaine, ecstasy, and amphetamines.

The follow-up (re-test) condition was scheduled for two weeks after the baseline data collection was concluded. However, the actual timelag between test and re-test ranged between 13 and 64 days (M = 41.8, SD = 13.2), so the 30-day recall periods overlapped to some extent for just over a fifth (22.2%) of the follow-up survey participants (i.e. those who responded to the follow-up between 13 and 30 days after completing the baseline survey).

Measures

The web survey questionnaire was organized into seven question groups: (1) information and consent, (2) age, gender, prevalence questions and opinions, (3) cannabis module, (4) cocaine module, (5) ecstasy/MDMA module, (6) amphetamines module, (7) socio-demographic questions. All questions, apart from informed consent, were voluntary. Modules 3–6 were displayed in a random order to respondents who reported having used the respective drug(s) in the last 12 months. At the end of each module, if they had used more than one drug, respondents were asked whether they wished to answer another drug module or finish the questionnaire. This procedure aimed to reduce the burden for users of multiple drugs and to reduce drop out due to time constraints and/or boredom.

Substance use

The prevalence questions asked respondents whether they had used a range of drugs in the past 30 days, past 12 months, more than 12 months ago, or never. An individual question format was used for each of the four main substances (cannabis, cocaine, amphetamines, and ecstasy/MDMA), while for other substances the prevalence items were displayed using a matrix question format. Last year cannabis users were also asked whether they used herbal cannabis, cannabis resin, and/or cannabis oil. Subsequent questions on herbal cannabis and cannabis resin were then asked separately. Due to a problem with the questionnaire programming, some herbal cannabis questions were not displayed to respondents in the test condition; for these items the reliability indicators could not be calculated. Last year users of amphetamines were asked whether they used amphetamine and/or methamphetamine, however, subsequent questions referred jointly to "amphetamines".

Frequency of use

Frequency of use in the past 12 months was assessed using two separate items: a categorical approach with options ranging from daily to less than once a month, and a question on the number of days of use. Frequency of use in the last 30 days was assessed only by number of days of use in the last 30 days. Following the methodology established by Trautmann et al. (2013), although with slightly modified cut-offs for the categories, users were categorised on the basis of their use in the last 12 months. Four categories were defined for cannabis users: (1) Infrequent users = people using cannabis on less than 12 days in the past year or "less than once a month"; (2) Occasional users = people using cannabis on 12–51 days or "less than once a week but at least once a month"; (3) Regular users = people using cannabis on 52–250 days or "once a week and 'more than once a week, but not daily or almost daily"; (4) Intensive users = people using cannabis on more than 250 days or "daily or almost daily". Three categories were defined for all other drugs, with infrequent and occasional users defined as for cannabis, but frequent users defined as people using the drug on 52 or more days, or "once a week or more".

Consumed amounts

A complex set of questions assessing consumption habits and usual amounts was asked for each substance. As herbal cannabis items (including those using prompt cards) were not displayed to respondents in the baseline questionnaire, the amounts could only be calculated for cannabis resin. Respondents were asked how they usually use the drug (in a joint, dry pipe or chillum, water pipe, food item, beverage, other), how many of each they use on a typical day, and how much resin they usually put in. Amounts of the drug consumed in food and beverages were not asked, as almost no respondents reported this route of administration. The ordinal questions on the amount of resin in the typical joint/pipe used a prompt cards showing crumbled and non-crumbled material in the quantity of 0.05 g, 0.1 g, 0.2 g, and 0.3 g alongside a credit card and a ruler for scale, and the response categories also included "in-between" options. Cocaine users were directly asked how many grams they use on a typical day they use cocaine. Amphetamines and ecstasy/MDMA users were asked whether they use the drug in the form of tablets or powder/crystal and, for each form, how many tablets/grams they use on a typical day they use the drug.

Statistical analysis

Analysis and data management were done using IBM Statistics (with an extension bundle for weighted Kappa – see Nichols 2015). The test and retest data were merged using a unique identifier. To assess test/retest reliability, the weighted kappa statistic (κ, linear weighting) was calculated for categorical and dichotomous variables (Cohen 1968, Landis and Koch 1977, Nichols 2015), while the interclass correlation coefficient (ICC) for single measurement, absolute agreement, two-way mixed effects was computed for ordinal and continuous items (Koo and Li 2016, Weir 2005, Landers 2015). Following the suggestion of Koo and Li (2016), reliability was rated as poor (less than 0.5), moderate (0.50 to 0.74), good (0.75 to 0.90) or excellent (greater than 0.90). The consistency between test and re-test was also assessed using Pearson correlation coefficient (r, continuous variables) and Spearman coefficient (rho, ordinal variables with longer scales), although these measures are considered weak in terms of systematic error detection (Berchtold 2016, Weir 2005). For prevalence questions, both Spearman coefficients and Kappa statistics were computed as these carry some information on order and may be ranked. The respective statistics were calculated for each item of the questionnaire and all constructed measures; only items with sample sizes of 20 or more, and relevant for consumption estimates, are reported (full results are available upon request from the corresponding author). Having an interval between test and re-test greater than 30 days for the majority of respondents provides an opportunity to consider the consistency of reported last month use over time and the potential implications of this for using it as a proxy for last year use. Consistency between test and retest was measured using the statistics shown above. In addition, frequency of use in the past 12 months was cross-tabulated with frequency of use in the last 30 days for each substance to assess whether last month frequency of use is a good representation of the yearly frequency of use.

Table 1
Sample characteristics at the baseline.

	Not followed	n	Follow-up	n	p*
Socio-demographics					
mean age (SD)	29.3 (7.3)	452	28.4 (6.7)	158	.192
% male	66.6	452	59.5	158	.218
% completed university	45.5	376	39.5	157	.458
% stable occupation or studies	92.0	388	95.6	158	.794
% referred from online source	85.3	387	85.4	158	.959
Use in the last 12 months:					
% used cannabis	86.3	452	91.1	158	.112
% used MDMA/ecstasy	41.9	451	41.8	158	.976
% used amphetamine	25.7	452	23.4	158	.575
% used cocaine	29.5	451	26.6	158	.487
% used crack	0.9	441	1.3	157	.692
% used heroin	1.8	443	0.6	157	.300
% used alcohol	95.5	447	95.6	158	.982
% used GHB	2.3	439	2.6	153	.814
% used ketamine	6.4	440	5.8	155	.805
% used LSD	23.3	442	23.7	156	.916
% used other hallucinogens	28.0	440	26.5	155	.719
% used cathinones	3.8	443	1.3	155	.120
% used synthetic cannabinoids	3.6	442	2.0	154	.309

*Note: Mean difference tested using t-test, proportions using chi-squared test.

Results

In total, 158 respondents, aged 18–58 (M_{age} = 28.4, SD_{age} = 6.7), 59.5% male, completed at least one drug module in both the test and retest questionnaire. There were no statistically significant differences between those who responded in the follow-up and those who did not – see Table 1 for details.

Test-retest reliability

Table 2 shows reliability indicators for prevalence and frequency items, and for usual amounts used daily. Prevalence questions showed high reliability, with the exception of alcohol, ketamine, and other hallucinogens questions that showed modest to moderate reliability. Regarding frequency of use, the question on number of days in the past 12 months was more reliable than the categorical item. This finding applies to all drug groups (incl. cannabis resin where days within the last 12 months represented the only significantly reliable measure, while the categorical approach had low reliability coefficient despite the test/re-test averages showing little difference). The same pattern was observed for amphetamines where, however, compared to other substances, days within past 12 months showed only moderate reliability. The categorical approach was the least consistent measure across all substances.

Table 2
Test/re-test reliability of self-reported drug use.

	Test % or Mean (SD)	Re-test % or Mean (SD)	Diff. (T2-T1)	N	r/rho	ICC or κ
COCAINE						
Cocaine use, last 12 months	24.6	28.6	4	154	.92**	**.86****
Last 12 months frequency, weekly+ (categorical)	7.5	6.7	−0,8	30	.52*	.45**
Days in the last 12 months,	7.80 (10.58)	8.7 (11.7)	0,9	30	.93**	**.93****
AMPHETAMINES						
Amphetamines use, last 12 months	23.2	25.8	2,6	155	.88**	**.84****
Last 12 months frequency, weekly+ (categorical)	31.3	17.2	−14,1	27	.53*	.46**
Days in the last 12 months	40.04 (69.10)	32.60 (46.75)	−7,44	27	.67**	.62**
CANNABIS						
Any cannabis use, last 12 months	90.9	90.9	0	154	.73**	.71**
Resin / Last 12 months frequency, weekly+ (categorical)	22.9	23.1	0,2	39	.33*	.11
Resin / Days in the last 12 months	18.84 (34.09)	18.44 (26.99)	−0,4	38	.71**	**.86****
Herbal / Last 12 months frequency, weekly+ (categorical)	55.2	55.9	0,7	134	.92**	**.85****
herbal / Days in the last 12 months	153.79 (145.96)	151.29 (145.24)	−2,5	123	.96**	**.96****
ECSTASY/MDMA						
Ecstasy/MDMA use, last 12 months	42.2	44.2	2	154	.87**	**.83****
Last 12 months frequency, weekly+ (categorical)	7.9	8.8	0,9	57	.75**	.55**
Days in the last 12 months	10.42 (11.78)	10.56 (12.35)	0,14	55	.75**	**.76****
OTHER SUBSTANCES						
Crack cocaine use, last 12 months	1.3	0.7	−0,6	153	.73**	**.75****
Heroin use, last 12 months	0.7	0	−0,7	153	.89**	**.87****
Alcohol use, last 12 months	95.4	97.4	2	153	.39**	.55**
GHB use, last 12 months	2.7	2.7	0	148	.82**	**.81****
Ketamine use, last 12 months	5.3	4.7	−0,6	150	.68**	.64**
LSD use, last 12 months	24.7	24.7	0	150	.90**	**.85****
Other hallucinogens use, last 12 months	24.8	24.8	0	149	.64**	.64**
Synthetic cathinone, last 12 months	1.3	2.7	1,4	150	.74**	.73**
Synthetic cannabinoids, last 12 months	2.0	2.7	0,7	149	.72**	.72**
USUAL AMOUNTS USED						
Cannabis resin in joint – grams (prompt)	0.17 (0.14)	0.16 (0.12)	−0,01	30	.56**	.56**
Cocaine use/day – grams	0.79 (0.48)	0.75 (0.48)	−0,04	26	.94**	**.94****
Amphetamines use/day – grams	0.46 (0.47)	0.42 (0.46)	−0,04	25	.68**	.69**
Ecstasy/MDMA use/day – grams	0.61 (0.88)	0.49 (0.36)	−0,12	32	.44**	.31*
Ecstasy/MDMA use/day – tablets	1.43 (1.00)	1.61 (1.30)	0,18	35	.76**	.73**
DAYS IN THE LAST 30 DAYS						
Cocaine	0.90 (1.81)	1.16 (2.21)	0,26	30	.54*	.53**
Amphetamines	2.81 (3.91)	3.07 (4.31)	0,26	27	.56*	.56**
Cannabis resin	2.08 (4.17)	2.58 (4.95)	0,5	38	.52**	.00
Herbal cannabis	12.86 (12.21)	12.51 (12.20)	−0,35	120	.92**	**.92****
Ecstasy	1.33 (1.79)	1.18 (1.93)	−0,15	55	.64**	.64**

Note: *p < 0.05; **p < 0.001. Good (0.75 − 0.90) and excellent (> 0.90) reliability indicators in bold.

Items using prompt cards had small samples. Only the amount users usually put in joints could be assessed and this showed moderate reliability. Amounts used in a dry pipe and water pipe, and amounts usually purchased were reported by only 13, 4, and 11 respondents respectively. For cocaine, amphetamine, and ecstasy, only the amounts usually consumed on a typical day could be assessed due to the small sample sizes of the items asking for amounts purchased. The best reliability was observed for cocaine and ecstasy in the form of tablets. The measure of amphetamines daily amounts used in grams was moderately reliable. There were no users of amphetamines in the form of tablets in the sample.

With respect to the typologies of users based on frequency of use, Table 3 summarizes the assessment of the consistency between the categorical and numerical approach and the stability between baseline and follow-up. For herbal cannabis and amphetamine, the types calculated on the basis of the two different items were consistent within the baseline survey; the categorisation of people using cannabis resin, cocaine and ecstasy showed only poor consistency between the two question types. When comparing baseline and follow-up, the typologies based on the number of days within past 12 months produce more consistent results. Typologies of cocaine users perform only moderately well regardless the type of items used to construct them.

Consistency of reports on the last 30 days

In contrast to last 12 months use, the reports on frequency of use in the last 30 days were less consistent between baseline and follow-up (the reliability indicators were rated as poor to moderate). Herbal cannabis being an exception with ICC = 0.92 – see Table 2.

The types of users based on frequency of use in the past 12 months were cross-tabulated with the frequency of use in the last 30 days for each substance to assess whether last month frequency of use is a good representation of the yearly consumption – see Table 4. The comparison revealed that, for most substances, there are some users for whom it appears that use in the last 30 days is not representative of the whole year. For instance, for each substance there is a proportion of last month non-users who reported having used the substance on more than 11 days in the past year; for amphetamines, some of these non-users qualify as intensive users when the whole year is considered. Overall, the proportion of users misclassified in this way was 11.4% for herbal cannabis, 15.2% for cannabis resin, 5.3% for cocaine, 12.0% for amphetamines, and 5.6% for ecstasy/MDMA.

Table 3
Assessment of stability of typologies of users: (a) comparing typologies based on different indicators of frequency of use in the past year within baseline; and (b) for the different typologies between baseline and follow-up data collection.

	rho	κ	SE	N
(a) Agreement between categorical and numerical item (baseline)				
Herbal cannabis	.919***	.846***	.015	476
Cannabis resin	.581***	.473***	.052	183
Cocaine	.618***	.490***	.077	152
Amphetamines	.817***	.746***	.050	124
Ecstasy/MDMA	.602***	.497***	.054	124
(b) Test/re-test reliability				
Herbal cannabis				
Categorical item	.921***	.848***	.031	124
Number of days	. 930***	.850***	.029	123
Cannabis resin				
Categorical item	.335*	.167	.104	39
Number of days	.851***	.736**	.088	38
Cocaine				
Categorical item	.516**	.450**	.173	30
Number of days	.671**	.667**	.177	30
Amphetamines				
Categorical item	.520**	.447**	.143	27
Number of days	.837***	.769***	.096	27
Ecstasy/MDMA				
Categorical item	.749***	.585***	.083	57
Number of days	.754***	.706***	.090	55

Note: *p < 0.05; **p < 0.01; ***p < 0.001.

Table 4
Comparison of the user types based on frequency of use in the past 12 months with the reported frequency of use in the past 30 days, percentages calculated from totals (baseline data only).

Days in the last 12 months	Days in the last 30 days				
Herbal cannabis (N = 471)	0	1-2	3-10	11-20	> 20
0-11	**15.3%**	**9.1%**	1.7%	0.0%	0.0%
12-51	1.5%	**7.0%**	**7.9%**	0.0%	0.4%
52-250	0.6%	0.2%	**10.4%**	**10.0%**	1.5%
251-365	0.0%	0.4%	0.6%	4.5%	**28.9%**
Cannabis resin (N = 182)	0	1-2	3-10	11-20	> 20
0-11	**27.5%**	**23.6%**	3.8%	0%	0%
12-51	8.2%	**13.7%**	**11.5%**	0%	0.5%
52-250	0%	1.6%	**6.0%**	**0%**	1.1%
251-365	0%	0%	0%	0%	**2.2%**
Cocaine (N = 151)	0	1-2	3-10	11-20	> 20
0-11	**53.0%**	**31.8%**	1.3%	0%	0%
12-51	4.0%	**5.3%**	**4.6%**	0%	0%
52-365	0%	0%	**0%**	**0%**	**0%**
Amphetamines (N = 124)	0	1-2	3-10	11-20	> 20
0-11	**37.1%**	**13.7%**	3.2%	0%	0%
12-51	3.2%	**9.7%**	**12.1%**	0.8%	0%
52-365	1.6%	3.2%	**6.5%**	**4.8%**	**4.0%**
Ecstasy/MDMA (N = 215)	0	1-2	3-10	11-20	> 20
0-11	**43.3%**	**31.6%**	2.3%	0%	0%
12-51	3.3%	**10.7%**	**8.8%**	0%	0%
52-365	0%	0%	**0%**	**0%**	**0%**

*Note: Bold values mark overlapping categories. The remaining fields have no overlap and their sum represents cases for whom frequency of use in the past 30 days does not match frequency of use in the past 12 months. For herbal cannabis this represents 11.4% of cases, for cannabis resin 15.2%, cocaine 5.3%, amphetamines 12.0%, and for ecstasy/MDMA 5.6% of cases.

Discussion

Our study, conducted in the Czech Republic in the context of European Web Survey on Drugs, aimed to assess the test-retest reliability of items measuring prevalence and frequency of use of a number of substances and the usual amounts consumed. Cognitive interviews helped to explain why some of the items could be more problematic than others. The reliability of most items was sufficient (rated moderate to excellent), with variations marked by recall period, question format, or type of substance.

The categorical questions on frequency of use showed only moderate reliability for most drugs, except herbal cannabis. They prompted additional questions from respondents in cognitive interviews and seemed too rigid for the irregular drug using patterns associated with some substances. For instance, the lowest category ("less than once a month") implies a certain regularity or repetition and was not selected by those who used the drug only once or twice. The exception of herbal cannabis is perhaps related to the high proportion of regular users in the sample that, on average, smoked cannabis on 13 days in the past 30 days.

Drug use is a complex behaviour, often regulated by subcultural norms that differ for different groups of users; it involves the illegal activities of purchasing and sharing of drugs and as such it responds to market fluctuations and is prone to irregularity and instability even for more frequent users (Golub et al., 2005, Järvinen and Ravn, 2011). The cognitive interviews highlighted the question of irregular patterns of use and suggested that frequency of use questions asking for numerical responses are likely to be more accurate than those presenting categorical options, as they accommodate unusual periods (for instance related to sickness or attempt to quit) and inspire more effort in responding than the categorical approach that assumed patterns to be more regular than the typical respondents' use. Some authors suggest that fuzzy questioning, which allows respondents to provide a range rather than a point estimate, may improve accuracy of self-reporting of drug consumption because it reflects better the nature of such behaviours (Matt et al., 2003). In our study point estimates of days of use within certain periods showed high reliability, their accuracy however still needs to be assessed.

Question layout might also have an impact on the reliability of the items, which was the same for the main four drugs (cannabis, cocaine, ecstasy/MDMA, and amphetamines were each addressed in a separate question). There was more variability in the reliability of prevalence questions for substances where these questions were displayed in matrix format. In particular, items on the prevalence of ketamine, other hallucinogens and alcohol were only moderately reliable. Some variation may be explained by the low experience and familiarity of respondents with some terms. This is particularly so in the case of new psychoactive substances that may be difficult to distinguish and categorise. Ketamine is a less common drug, which may have influenced the reliability. On the other hand, inconsistencies between test and re-test in the alcohol question could be linked to the fact that the period of data collection covered a period in which there are various abstinence-orientated campaigns (such as Dry March) when people avoid alcohol - these activities are very popular among Czechs. In the Czech context, other hallucinogens often refer to psilocybin (mushrooms) that are seasonally collected and consumed fresh, or dehydrated for later use (Mravčík et al. 2017). Lower Kappa values were observed in substances for which irregular using patterns are more likely, or that are less available and therefore used only a few times a year but often for more days on these occasions – cannabis resin is the typical example within the population of Czech cannabis users (as well as cocaine, at the time of data collection).

The prompt cards used to display amounts for cannabis appeared to be a promising way of helping respondents to answer questions concerning amounts used or purchased and were evaluated as helpful during the cognitive interviews, although these items generally had insufficient sample sizes for reliability testing. Estimates of the amount of cannabis resin in a joint were only moderately reliable and cognitive interviews offered several possible explanations. For cannabis users of both herbal cannabis and resin, the typical amount of drug they put into the pipe differs in relation to the size of the instrument that may range from small "one-shot" pipes (typically used throughout the day by daily users) to pipes with large containers (such as "chillum" that is often used in the evening or shared with friends). Similarly, the size and potency of a joint depends on who prepares it, on the time of the day and the number of people sharing it. Some participants also commented that the pictured resin is not representative of all forms in which it appears on the market. In this respect, it should also be noted that the web survey questionnaire can be completed on different devices that display the prompt cards in varying sizes and resolutions which may be a source of unexpected bias. The cognitive interviews showed that the ruler and credit card that were included in the pictures to guide the estimation of size of the drug are not always immediately seen or regarded by respondents.

The groupings of drug users calculated on the basis of frequency of use questions asked in different ways (numerical or categorical format) showed satisfactory consistency, that may indicate that respondents are in each case referring to the true value when making their answer or aim to be consistent when the same information is requested more than once (the explanation cognitive psychology would provide – Tourangeau et al. 2000). On the other hand, the test-retest consistency of frequency of use types was related to the reliability of the individual items from which they were constructed and could have been affected by the time-lag between the measurements.

In our study the test-retest interval for many individuals was greater than one month; for the majority of participants (77.8%) there was no overlap between the last 30 day periods that they were reporting on, responses were therefore susceptible to correctly reported changes in last month use. The reliability coefficients for items covering the period of past month therefore should be interpreted primarily as indicators of stability of drug using patterns from month to month in the sample, rather than the reliability of the respective items. Self-reported last month drug use is believed to be easier to recall and therefore likely to be more accurate than self-reported last year drug use, and hence it is sometimes used to impute values for last year use (Kilmer et al., 2011). However, the reports of frequency of use in the last 30 days showed quite low consistency between the baseline and follow-up suggesting that, for most drugs, use may be variable from month to month. The exception of herbal cannabis may be related to the high proportion of regular cannabis users in the sample, who reported on average 13 days of use in the past 30 days.

Asking just one question on frequency of use (in Europe this is typically concerning use in last 30 days) is a common strategy to reduce survey costs. Previously reported findings (Trautmann et al. 2013) suggest that frequency of use in the past 30 days is not a good indicator of the yearly frequency of use and should not be simply extrapolated. The comparison between the reported last month and last year frequency of use in our study provides support for this view. Drug using patterns tend to be irregular and even people who use frequently or intensively experience short periods of abstinence or binges. This factor may have impacted on the differences between test and retest reports of past month use in our study and may cause bias when estimating annual consumption by extrapolation from past month use.

The literature also suggests that people who use drugs tend to provide more reliable answers for longer stretches of time, as shown in the comparison to weekly and monthly follow-back recalls of cocaine use (Hersh et al., 1999). This can be a result of two contradictory processes: on one hand, longer periods may better accommodate day-to-day or week-to-week variability leading to more precision; on the other hand, the responses may be more automatic and stereotypical, than well thought through. More in-depth understanding on how respondents craft answers to frequency questions and their validation against other methods, such as daily reports, would be useful to address this uncertainty.

Our data neither confirm, nor refute concerns about data quality from self-selecting samples in web surveys. Answering questions on past substance use behaviours in a precise manner requires complex cognitive processing and decision-making (Schwarz 2007), yet our cognitive interviews suggest that often efforts are not made to answer questions in a precise manner even when the questionnaire is completed with an interviewer present. This is particularly so when a description of typical or average behaviour is requested about activities that do not follow regular patterns. Different drug using habits were highlighted in the cognitive interviews by different subgroups of drug users or within the same subgroup for different substances, which had an impact on how they responded to the questionnaire or why some questions were difficult to answer. Some interviewees that previously filled the questionnaire online reported making less effort to recall certain events and estimates when completing the questionnaire online, compared to the face-to-face interview.

It should be noted also that the web survey reached mainly stable, socially integrated drug users with internet access and enough time to dedicate to filling out the questionnaire. The composition of the sample also mirrors the means of recruitment and shows that online solicitation was the most successful strategy. Socially integrated users with ready access to internet and social media seem to be the ideal target group for web surveys. People who use drugs in a problematic way only represented a minority in the sample and, despite offline recruitment strategies, only a small number of survey respondents were reached through needle and syringe exchange programmes. Moreover, people with problematic drug use faced more difficulties in understanding the questionnaire and made the least effort to provide precise and accurate answers during the cognitive testing. This suggests that other methods would be the better suited to survey this population.

It should be stressed, however, that the willingness of respondents to provide contact email address at baseline and the response rate to the follow-up survey were relatively high, suggesting the questionnaire was of reasonable length and acceptable burden for respondents. Nevertheless, the sample size in the follow-up data collection was rather low, and particularly so for cocaine and cannabis resin. The randomisation of which drug modules were displayed to respondents also reduced the sample sizes available for comparison as it could have happened that respondents were not exposed to the same set of questions in the test and re-test condition. Some items on herbal cannabis were not displayed to respondents in the base-line data collection due to a technical problem that remains unresolved.

Despite these limitations, the reliability of most indicators was rated as high and statistically significant. The response to the questionnaire was largely positive and willingness to participate in the follow-up and cognitive interviews was high. The findings from the reliability and cognitive testing resulted in some modifications in the subsequent data collection in other countries.

Conclusions

Pre-testing the questionnaire with members of the target population remains an important pre-requisite of successful implementation of web surveys. Our study shows the importance of including items that do not assume regular behaviours, especially when asking about drug use that is occurring in illegal or semi-legal contexts and therefore more prone to disruptions and changes. Both last 30 days and last 12 months frequency of use might be considered for inclusion into the questionnaires of larger (general population) surveys so that they can provide more valid drug consumption estimates. Undertaking further reliability and validity studies in different populations will facilitate the development of robust questions that can be included in different web surveys thus enhancing comparability across studies. Cross-national studies such as the European Web Survey on Drugs may provide an opportunity for testing the new versions of the questions with bigger samples and validating alternative methods for gathering information on amounts consumed and testing the questionnaire in different cultural and linguistic environments. Contextual factors related to language, national, local and regional practices and habits of various subgroups of drug users impact on interpreting and answering the questionnaire. Any study seeking cross-cultural comparability should take this into account in its design.

Author contribution

KS was involved in study design and data collection, performed the analysis and prepared the first draft of the paper. NS, JM, and VM supervised the research process and contributed to the draft of the paper.

Funding

The study was conducted within projects supported by the EMCDDA [CT.15.EPI.0112.1.0 and CT.16.EPI.0121.1.0]. Authors would also like to acknowledge support provided by Charles University, Programme Progres No. Q06/LF1 and Project No. LO1611, with financial support from the Ministry of Education, Youth, and Sports of the Czech Republic as part of the NPU I programme.

Ethical review

The study did not require ethical review. Only participants aged 18+ who provided informed consent could participate in the study. Anonymity and confidentiality were ensured for all participants.

Availability of data and materials

Data available upon request.

Conflict of interest

Authors declare no conflict of interest.

Acknowledgements

Authors would like to thank to all respondents of the survey and to all participants of the cognitive interview testing for their time and willingness to disclose sensitive information for research purposes. We also thank to our reviewers for valuable comments and suggestions.

References

Barratt, M. J., Ferris, J. A., Zahnow, R., Palamar, J. J., Maier, L. J., & Winstock, A. R. (2017). Moving on from representativeness: Testing the utility of the Global Drug Survey. *Substance Abuse Research and Treatment, 11*, 1–17.

Berchtold, A. (2016). Test–Retest: Agreement or reliability? *Methodological Innovations*. https://doi.org/10.1177/2059799116672875.

Caulkins, J. P., Kilmer, B., Reuter, P. H., & Midgette, G. (2015). Cocaine's fall and marijuana's rise: Questions and insights based on new estimates of consumption and expenditures in US drug markets. *Addiction, 110*, 728–736. https://doi.org/10.1111/add.12628.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Cuttler, C., & Spradlin, A. (2017). Measuring cannabis consumption: Psychometric properties of the daily sessions, frequency, age of onset, and quantity of cannabis use inventory (DFAQ-CU). *PloS One, 12*(5), e0178194.

Denscombe, M. (2006). Web-based questionnaires and the mode effect: An evaluation based on completion rates and data contents of near-identical questionnaires delivered in different modes. *Social Science Computer Review, 24*(2), 246–254.

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.

Enghoff, O., & Aldridge, J. (2019). The value of unsolicited online data in drug policy research. *The International Journal of Drug Policy*. https://doi.org/10.1016/j.drugpo.2019.01.023.

Golub, A., Johnson, B. D., & Dunlap, E. (2005). Subcultural evolution and illicit drug use. *Addiction Research & Theory, 13*(3), 217–229.

Hersh, D., Mulgrew, C. L., Van Kirk, J., & Kranzler, H. R. (1999). The validity of self-reported cocaine use in two groups of cocaine abusers. *Journal of Consulting and Clinical Psychology, 67*(1), 37.

Hoeppner, B. B., Stout, R. L., Jackson, K. M., & Barnett, N. P. (2010). How good is fine-grained Timeline Follow-back data? Comparing 30-day TLFB and repeated 7-day TLFB alcohol consumption reports on the person and daily level. *Addictive Behaviors, 35*(12), 1138–1143.

Järvinen, M., & Ravn, S. (2011). From recreational to regular drug use: Qualitative interviews with young clubbers. *Sociology of Health & Illness, 33*(4), 554–569.

Johnson, T. P. (2014). Sources of error in substance use prevalence surveys. *International Scholarly Research Notices, 2014*. https://doi.org/10.1155/2014/923290.

Khadjesari, Z., Murray, E., Kalaitzaki, E., White, I. R., McCambridge, J., Godfrey, C., & Wallace, P. (2009). Test–retest reliability of an online measure of past week alcohol consumption (the TOT-AL), and comparison with face-to-face interview. *Addictive Behaviors, 34*(4), 337–342.

Kilmer, B., Caulkins, J. P., Pacula, R. L., & Reuter, P. H. (2011). Bringing perspective to illicit markets: Estimating the size of the U.S. Marijuana market. *Drug and Alcohol Dependence, 119*(1), 153–160. https://doi.org/10.1016/j.drugalcdep.2011.08.008.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Korf, D. J., Benschop, A., & Wouters, M. (2007). Differential responses to cannabis potency: A typology of users based on self-reported consumption behaviour. *The International Journal of Drug Policy, 18*(3), 168–176.

Landers, R. N. (2015). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. *The Winnower, 5*, e143518.81744. https://doi.org/10.15200/winn.143518.81744.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374.

Mariani, J. J., Brooks, D., Haney, M., & Levin, F. R. (2011). Quantification and comparison of marijuana smoking practices: Blunts, joints, and pipes. *Drug and Alcohol Dependence, 113*(2-3), 249–251. https://doi.org/10.1016/j.drugalcdep.2010.08.008.

Matt, G. E., Turingan, M. R., Dinh, Q. T., Felsch, J. A., Hovell, M. F., & Gehrman, C. (2003). Improving self-reports of drug-use: Numeric estimates as fuzzy sets. *Addiction, 98*(9), 1239–1247.

Miller, P. G., & Sønderlund, A. L. (2010). Using the internet to research hidden populations of illicit drug users: A review. *Addiction, 105*(9), 1557–1567.

Mravčík, V., Chomynová, P., Grohmannová, K., Janíková, B., Tion Leštinová, Z., Rous, Z., Kiššová, L., Kozák, J., Nechanská, B., Vlach, T., Černíková, T., Fidesová, H., & Vopravil, J. (2017). In V. Mravčík (Ed.). *Výroční zpráva o stavu ve věcech drog v České republice v roce 2016 [annual report on drug situation 2016 – Czech Republic]*. Praha: Úřad vlády České republiky.

Nichols, D. (2015). *Stats Weighted Kappa, v1.2.1*. Accessed on 20 September fromhttps://www.ibm.com/developerworks/community/files/app#/file/9c07d417-3f28-4087-9306-b73fdd72047a.

Norberg, M. M., Mackenzie, J., & Copeland, J. (2012). Quantifying cannabis use with the Timeline Followback approach: A psychometric evaluation. *Drug and Alcohol Dependence, 121*(3), 247–252. https://doi.org/10.1016/j.drugalcdep.2011.09.007 PubMed PMID: 71908363.

Robinson, S. M., Sobell, L. C., Sobell, M. B., & Leo, G. I. (2014). Reliability of the Timeline Followback for cocaine, cannabis, and cigarette use. *Psychology of Addictive Behaviors, 28*(1), 154–162. https://doi.org/10.1037/a0030992 PubMed PMID: 2012-34893-001.

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 21*(2), 277–287.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Trautmann, F., Kilmer, B., & Turnbull, P. (2013). *Further insights into aspects of the EU illicit drugs market Editors. Report of consortium of Trimbos Institute, RAND Europe and ICPR for European Commission- Directorate-General for Justice*Luxembourg: Publications Office of the European Union.

van der Pol, P., Liebregts, N., de Graaf, R., Korf, D. J., van den Brink, W., & van Laar, M. (2013). Validation of self-reported cannabis dose and potency: an ecological study. *Addiction (Abingdon, England), 108*(10), 1801–1808. https://doi.org/10.1111/add.12226 PubMed PMID: 23627816.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research, 19*(1), 231–240.

Zeisser, C., Thompson, K., Stockwell, T., Duff, C., Chow, C., Vallance, K., et al. (2012). A' standard joint'? The role of quantity in predicting cannabis-related problems. *Addiction Research & Theory, 20*(1), 82–92. https://doi.org/10.3109/16066359.2011.569101 PubMed PMID: 69733427.