International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 21(2): 158–168 (2012)
Published online 23 February 2012 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1353

Concordance and consistency of answers to the self-delivered ESPAD questionnaire on use of psychoactive substances

SABRINA MOLINARO, VALERIA SICILIANO, OLIVIA CURZIO, FRANCESCA DENOTH & FABIO MARIANI

Institute of Clinical Physiology, Epidemiology Section, Italian National Research Council, Pisa, Italy

Key words
self-administrated questionnaire, psychoactive substance use, school survey, reliability, ESPAD

Correspondence
Sabrina Molinaro, Clinical Physiology Institute, CNR, Via Moruzzi 1, 56124 Pisa, Italy.
Telephone (+39) 050-3152094
Fax (+39) 050-3152095
Email: molinaro@ifc.cnr.it

Received 5 February 2010;
revised 8 September 2010;
accepted 17 December 2010

Abstract

Considering the prevalence of drug use in Italy, it is crucial to develop a reproducible screening test. Test–retest reliability and internal consistency are important indicators of a measurement's temporal stability and are a necessary condition for validity. The aim of the study was to assess the consistency and concordance of the European School Survey Project on Alcohol and Other Drugs (ESPAD) questionnaire; participating students completed the questionnaire twice, with a three-week interval. To verify the concordance for variables relating to use of alcohol, cigarettes and illicit drugs, the original ordinal variables as well as the same dichotomically recodified variables were used. Data analysis was done using Kappa and weighted Kappa. The method proposed by Lipsitz was used to evaluate the influence of gender and age on concordance. Questions about drug use, examined in ordinal form, show a good test–retest concordance and an excellent concordance for answers relating to the use of cigarettes, alcohol and cannabis. Regarding the effect of age adjusted for gender, 15-year-old subjects showed a lower concordance than 19-year-olds. ESPAD is a tool with a good reproducibility. Results focus on the effect of gender and age covariates on the concordance of answers regarding drug use and suggest the importance of examining the concordance in relation to the covariate levels.

Introduction

Studies on the estimation of the pattern of substance use play a primary role in our understanding of consumption behaviour. The findings of such studies help focus public health campaigns and develop prevention measures and treatments. Survey research can provide a thorough profile of drug use and abuse among a broader cross-section of the population, and it can also provide a wide range of information for use in designing intervention strategies (Harrison, 1995).

The findings of the sample-based European School Survey Project on Alcohol and Other Drugs (ESPAD) survey initiated by the Council of Europe (Pompidou Group) and conducted in Italy since 1995 by the National Research Council, reveal widespread use of psychotropic drugs.

One-third of the Italian students who answered the questionnaire report the use of illicit drugs at least once in their lifetime and 28% report use of illicit drugs during the last 12 months (European Monitoring Centre for Drugs and Drug Addiction, 2003). Prevalence rates are higher among boys than among girls (33% of boys and 23% of girls report the use of illicit drugs during the last 12 months) and increase as users get older (12% of users among 15-year-olds and 39% of users among 19-year-olds). If the focus is shifted to licit drugs such as alcohol and tobacco, we have a larger

number of users (48% of boys and 37% of girls report drunkenness at least once during the previous 12 months; 65% of boys and 69% of girls tried smoking cigarettes at least once in their lifetime, 40% of boys and 42% of girls smoked at least one cigarette during the last 30 days). In addition to the use of self-reported questionnaires, biological tests can be used to estimate the spread of patterns of drug use (Hawks and Chiang, 1986; Wolff *et al.*, 1999).

Self-reporting questionnaires are less expensive, less invasive and can provide information about long-term consumption. Nevertheless, there is widespread debate about their validity and reproducibility. Some authors support the use of self-administered questionnaires (Hibell *et al.*, 2009; Hibell *et al.*, 2004; Brown *et al.*, 1993; Sherman and Bigelow, 1992) while others tend to be skeptical (Harrison *et al.*, 1993). Kokkevi *et al.* (2007a) points out that the ESPAD study has some limitations. One is that the cross-sectional design does not allow etiological inferences and another is its reliance on self-reports, although various studies have demonstrated the validity of assessing substance use in this way (Shillington *et al.*, 1995).

Test–retest reliability had been examined by Hibell *et al.* (2000) and mean results were highly reliable. Several methods have been proposed to identify one or more factors which can be predictive of concordance (Donner and Koval, 1980; Graham, 1995; Williamson and Manatunga, 1997; Gonin *et al.*, 2000; Lipsitz *et al.*, 1994; Agresti, 1990).

The test–retest reproducibility of the standard ESPAD questionnaire matching individuals has not yet been evaluated.

Preventive interventions in schools, addressing the use of the most popular legal and illegal drugs, environmental risk factors, and deviant behaviour, might be a successful approach to curbing adolescents' further involvement in a life-style of problem behaviour (Kokkevi *et al.*, 2007b).

The current study aims to examine the reproducibility of the ESPAD questionnaire (with a three-week gap between administrations) and to assess whether concordance between repeated measurements within the same subject can depend on covariates, namely on the peculiarities of the adolescents included in the sample. This was done using the regression model for the Kappa-index which was proposed by Lipsitz *et al.* (2001).

Methods

Sampling, data collection and tools

The sampling and data collection procedures are summarized in this paragraph; full details can be found in the 2003 ESPAD Report (Hibell *et al.*, 2004). The target population comprised Italian high school students aged 15–19 years. The survey takes place every year in March. The ESPAD questionnaire consists of core questions about licit and illicit drugs in terms of prevalence of use (lifetime, last year, last month), and additional questions about leisure activities, relationships at school, attitude concerning drug use (approval or perceived risk), satisfaction with relationships with parents or friends, social and cultural status.

Unlike the European questionnaire, the Italian questionnaire administered in 2004 contained structured questions about frequency of use of all substances in lifetime, the last year and the last month.

A test–retest methodology was used for a subsample from the ESPAD-Italia®2004 database. Concordance was evaluated by administering the questionnaire again after a 20–25 day interval. A sample of schools (stratified by four school types) distributed throughout the country, was representative of different types of Italian schools. The total sample size in our study comprised 910 students. Of these, 37.7% of respondents attended "liceo" (high school), 28.1% technical schools, 18% professional schools and 16.2% art schools. Of the sample, 51.8% were girls, 48.2% boys. The target population results were composed as follows: 20.3% ≤ 15 years old; 17.8% were 16 years old; 21.1% were 17 years old; 22.3% were 18 years old; 18.5% ≥ 19 years old.

A total of 788 students completed the test questionnaire, while 753 students completed the retest questionnaire. Nobody refused to answer. On the day of the test 13% of students were absent during the test, 17% absent during the retest. A total of 650 students completed both questionnaires, as reported in the class register (71% of the entire sample). For 499 of the students (77%) it was possible to match test and retest answers using a special subject code, although for 151 students (23%) it was not possible because the code was not completed or was wrongly completed. After deleting the cases of incongruent answers (11 students) the sample consisted of 488 students (75% of students who answered both questionnaires). More girls than boys completed the personal code correctly (52% versus 37 %), with a significant gender difference ($chi^2 = 27.9$, $p < 0.01$).

An analysis of internal and logical consistency was conducted on the 788 students completing the test-questionnaire.

Missing answers were considered as missing data. Three measures were discussed about internal consistency. One is the concordance between two sets of questions measuring the lifetime prevalence for different drugs. The questionnaire contained questions about age at first use of different drugs, such as "When (if ever) did you

FIRST do each of the following things?". These questions included the alternative "never", which makes it possible to differentiate the "users" from those who answered that they never used the drug (Hibell *et al.*, 2004).

Second is a quotient between the proportion of students who on the "honesty question" answered that they had already said they had used cannabis, and the proportion who actually gave this answer (Hibell *et al.*, 2004). The students were asked "If you had ever used marijuana or hashish (or heroin), do you think you would have said so in this questionnaire?" One of the response alternatives was "I already said I have used it", and this proportion that reported has been compared with the proportion that reported cannabis use on the lifetime prevalence question.

The other concerns the answers reporting use (taking into account the number of occasions) in the last year higher than in lifetime, and cases in which use in the last month was higher than in the last year and/or in lifetime. These answers were considered incongruent. Questionnaires in which at least two answers about the same substance were incongruent (i.e. last year use higher than lifetime use) were rejected.

In order to verify the test–retest concordance for the variables relating to lifetime use of alcohol, cigarettes and illicit drugs, the original ordinal variables as well as the same dichotomically recodified variables were used.

Statistical analysis

For all categorical variables, Cohen's k coefficient was calculated (1960). Weighted k was calculated for polytomous or ordinal data (Cohen, 1968).

In this way, some discordant answers were considered as concordant because they were logically possible regarding lifetime use. For example, having answered "never" to the test and "1–2 times" to the retest could be considered a plausible answer, since the subject could have begun using cannabis within the interval between the two administrations. Although a theoretical shift off more than one level is possible we considered concordant those answers which diverged by not more than one level in the reported frequency of use.

In the case of test–retest with an appropriate weights set, weighted k is asymptotically equal to the intraclass correlation coefficient (Donner and Koval, 1996). Landis and Koch (1977) described intervals for the values of k and associated different empirical concordance levels with these values.

Empirical concordance levels according to Landis and Koch are: $k < 0$ low; $0.00 \le k \le 0.20$ weak; $0.21 \le k \le 0.40$ sufficient; $0.41 \le k \le 0.60$ good; $0.61 \le k \le 0.80$ excellent; $0.81 \le k \le 1.00$ almost perfect. It can be reduced to three categories: $k \le 0.40$ low; $0.40 < k < 0.75$ good; $k \ge 0.75$ excellent.

In order to evaluate the influence of gender and age on the concordance, the method proposed by Lipsitz *et al.* (2001) was used.

In addition to the earlier mentioned analyses both the Spearman rank correlation for the original ordinal variables and Cramer's phi for the same dichotomically recodified were performed to test the concordance between answers.

The analyses were conducted using the STATA statistical package, Version 8.2.

Results

Table 1 shows the prevalence of drug use as reported by students who completed the first ($n = 788$), the second ($n = 753$) or both questionnaires ($n = 650$). The last group

Table 1 Prevalence of drug use reported by students who completed only the test or only the retest and by students who completed both questionnaires

Drugs examined by the ESPAD-Italia® 2004 questionnaire	Students present at the time of the first questionnaire (test)	Students present at the time of the second questionnaire (retest)	Students present at the time of both questionnaires		
	$P \times 100$	$P \times 100$	$P \times 100$	chi^2	$Prob > chi^2$
Tobacco	73.3	73.8	67.5	4.86	0.08
Alcohol	93.3	88.2	90.9	5.04	0.08
Drunkenness	67.9	68.2	56.8	15.05	0.001
Cannabis	47.3	52.5	35.3	25.48	0.000
Other illicit drugs	19.7	24.3	12.4	19.50	0.000

Int. J. Methods Psychiatr. Res. 21(2): 158–168 (2012). DOI: 10.1002/mpr

(respondents to both questionnaires) shows a significantly lower prevalence of drunkenness (prevalence = 56.8%; $chi^2 = 15.05$; $p = 0.01$), cannabis use (prevalence = 35.3 % ; $chi^2 = 25.48$; $p < 0.000$) or other illicit drugs (prevalence = 12.4 % ; $chi^2 = 19.50$; $p < 0.000$) use if compared with the other two categories.

Table 2 shows the results of internal consistency analyses.

The concordance of students' answers reporting drug use in the two questions, using Cohen's *k* (1960) coefficient and Cramer's phi for dichotomous variables, show an excellent concordance ($k \geq 0.75$) for answers relating to the use of cigarettes, drunkenness, cannabis, cocaine and ecstasy and good concordance ($0.40 < k < 0.75$) in the case of tranquillizers and/or sedatives and heroin.

Table 2 included the quotient between two proportions, with the "honesty answer" as the numerator and the "lifetime answer" as the denominator. A value of 1.0 means that the values are identical for both measures. The quotient is above 1.0 if more students answered that they had already said they had used the drug, than actually answered this to the direct question. The quotient is 0.82 for marijuana or hashish and 0.93 for heroin.

The percentage figures of missing answers in the test and in the retest questionnaire vary between 0.5% and 2.8%, while figures of incongruent answers vary between 0.5% and 4.3% (Table 3).

Tables 4 and 5 show results obtained by analysing the variables in their original form as well as in their recodified dichotomous form. The percentage of incongruent answers per prevalence is constant for each substance except for cannabis (Pearson $chi^2 = 3.9540$; Pr = 0.047) and alcohol mixed with marijuana/hashish (Pearson $chi^2 = 4.9932$; Pr = 0.025).

Good concordance was observed for prevalence rates of other illicit drugs considered separately.

Based on the classification elaborated by Landis and Koch (1977) the findings reported in Table 4, where answers to the questions about drug use were examined in ordinal form, generally show a good test–retest concordance and even an excellent concordance for the answers relating to the use of cigarettes, alcohol and cannabis. Except in the case of tranquillizers and sedatives, concordance increases when answers about drug use are examined in dichotomous form (Table 5). This codification allows us to

Table 2 Estimation of Cohen's *k* (1960) and Cramer's phi for the answer to the ESPAD-Italia®2004 questionnaire about lifetime use of cigarettes, drunkenness, cannabis, cocaine, hallucinogens, tranquillizers or sedatives, ecstasy, heroin (dichotomous answers) and quotient between two questions about cannabis and heroin use

Questions	*N*		Value[a]	Standard error	Z	Prob > Z	Quotient between two questions[b]
Cigarette use	774	Cohen's *k*	0.888	0.017	25.29	0.000	
		Cramer's phi	0.890			0.000	
Drunkenness	773	Cohen's *k*	0.862	0.018	24.40	0.000	
		Cramer's phi	0.862			0.000	
Cannabis	761	Cohen's *k*	0.862	0.018	24.53	0.000	0.82
		Cramer's phi	0.867			0.000	
Cocaine	761	Cohen's *k*	0.857	0.034	24.26	0.000	
		Cramer's phi	0.857			0.000	
Hallucinogens	762	Cohen's *k*	0.708	0.058	20.07	0.000	
		Cramer's phi	0.709			0.000	
Tranquillizers or sedatives	761	Cohen's *k*	0.654	0.059	18.61	0.000	
		Cramer's phi	0.657			0.000	
Ecstasy	761	Cohen's *k*	0.802	0.047	22.73	0.000	
		Cramer's phi	0.802			0.000	
Heroin	762	Cohen's *k*	0.689	0.069	19.68	0.000	0.93
		Cramer's phi	0.695		0.000		

[a]The first question is the self-reported lifetime prevalence question for the drug (a), while the second is a later one about the age at first use of the drug (b).

[b]Quotient a/b between the proportion of students answering "I already said that I have used it" to the question "If you ever used marijuana or hashish, do you think that you would have said so in this questionnaire?" (a) and the proportion of those reported that they ever used it (b).

Table 3 Analysis of respondents to the questions about drug use and of incongruent answers in the test–retest study

	Respondents		Incongruent answers	
Drugs examined by the ESPAD-Italia® 2004 questionnaire	Test (percentage of "missing" in the total number of students present, $n=788$)	Retest (percentage of "missing" in the total number of students present, $n=753$)	Test (percentage of the total number of respondents)	Retest (percentage of the total number of respondents)
Tobacco	784 (0.5)	748 (0.7)	26 (3.3)	26 (3.5)
Alcohol	782 (0.8)	746 (0.9)	10 (1.3)	12 (1.6)
Drunkenness	779 (1.1)	742 (1.5)	13 (1.7)	15 (2.2)
Cannabis	777 (1.4)	738 (2.0)	5 (0.6)	8 (1.1)
Other illicit drugs	772 (2.0)	732 (2.8)	26 (3.4)	31 (4.3)

Table 4 Estimation of Cohen's *k* (1960) and Spearman's rho for the answer to the ESPAD-Italia® 2004 questionnaire about lifetime use of cigarettes, alcohol, cannabis, tranquillizers or sedatives, hallucinogens, cocaine, ecstasy, alcohol mixed with pills, alcohol mixed with marijuana/hashish (ordinal answers) in test–retest study

Questions	*N*		Value	Standard error	Approximate *T*	Prob > *T*
Cigarette use	481	*k* unweighted	0.7249	0.0230	31.52	0.0000
		k weighted	0.8847	0.0358	24.68	0.0000
		Spearman' rho	0.9446	0.0082	62.97	0.0000
Drunkenness	479	*k* unweighted	0.5525	0.0254	21.77	0.0000
		k weighted	0.7032	0.0304	23.17	0.0000
		Spearman' rho	0.8048	0.0241	29.61	0.0000
Cannabis	470	*k* unweighted	0.6168	0.0261	23.62	0.0000
		k weighted	0.8009	00366	21.90	0.0000
		Spearman' rho	0.8547	0.0230	35.62	0.0000
Tranquillizers or sedatives	474	*k* unweighted	0.5960	0.0315	18.93	0.0000
		k weighted	0.4724	0.0315	15.01	0.0000
		Spearman' rho	0.2453	0.0825	5.503	0.0000
Hallucinogens	473	*k* unweighted	0.5320	0.0290	18.38	0.0000
		k weighted	0.5436	0.0349	15.58	0.0000
		Spearman' rho	0.6889	0.0667	20.63	0.0000
Cocaine	475	*k* unweighted	0.4121	0.0287	14.37	0.0000
		k weighted	0.4174	0.0341	12.23	0.0000
		Spearman' rho	0.5297	0.1065	13.58	0.0000
Ecstasy	474	*k* unweighted	0.3480	0.0309	11.26	0.0000
		k weighted	0.3421	0.0315	10.87	0.0000
		Spearman' rho	0.4680	0.0975	11.50	0.0000
Alcohol with pills	473	*k* unweighted	0.3994	0.0281	14.23	0.0000
		k weighted	0.4626	0.0326	14.18	0.0000
		Spearman' rho	0.5733	0.0932	15.18	0.0000
Alcohol mixed with marijuana/hashish	475	*k* unweighted	0.5184	0.0266	19.46	0.0000
		k weighted	0.6952	0.0357	19.47	0.0000
		Spearman' rho	0.7222	0.0376	22.71	0.0000

evaluate two further variables: every form of heroin consumption, in any way it is consumed, and use of illicit drugs, except for cannabis mixed with alcohol considered as a whole.

Figure 1 shows the values of the concordance in reported lifetime cannabis use stratified by age. In general there is a greater concordance with Spearman's rho and lower concordance with the unweighted Cohen's *k*.

Int. J. Methods Psychiatr. Res. 21(2): 158–168 (2012). DOI: 10.1002/mpr

Table 5 Answers to the ESPAD-Italia® 2004 questionnaire about cigarette smoking in lifetime, drunkenness, use of cannabis, tranquillizers or sedatives, hallucinogens, cocaine, heroin, ecstasy, alcohol mixed with pills, alcohol mixed with marijuana/hashish, all illicit drugs except for cannabis mixed with alcohol; and estimation of Cohen's *k* (1960) and Cramer's phi in test–retest study

Questions	*N*		Value	Standard error	*Z*	Prob > *Z*
Cigarette use	481	Cohen's *k*	0.8694	0.0455	19.11	0.0000
		Cramer's Phi	0.8710			0.0000
Drunkenness	479	Cohen's *k*	0.7369	0.0311	16.13	0.0000
		Cramer's phi	0.7371			0.0000
Cannabis	470	Cohen's *k*	0.8094	0.0461	17.56	0.0000
		Cramer's phi	0.8098			0.0000
Cocaine	473	Cohen's *k*	0.6948	0.0460	15.11	0.0000
		Cramer's phi	0.6949			0.0000
Hallucinogens	474	Cohen's *k*	0.6348	0.0455	13.95	0.0000
		Cramer's phi	0.6408			0.0000
Tranquil. or sedatives	475	Cohen's *k*	0.2539	0.0458	5.540	0.0000
		Cramer's phi	0.2541			0.0000
Ecstasy	474	Cohen's *k*	0.4654	0.0452	10.30	0.0000
		Cramer's phi	0.4729			0.0000
Heroin	488	Cohen's *k*	0.6934	0.0451	15.38	0.0000
		Cramer's phi	0.6940			0.0000
Alcohol with pills	477	Cohen's *k*	0.5669	0.0456	12.43	0.0000
		Cramer's phi	0.5715			0.0000
Alcohol mixed with marijuana/hashish	475	Cohen's *k*	0.6886	0.0458	15.03	0.0000
		Cramer's phi	0.6897			0.0000
All substances except for cannabis mixed with alcohol	467	Cohen's *k*	0.6110	0.0461	13.25	0.0000
		Cramer's phi	0.6133			0.0000

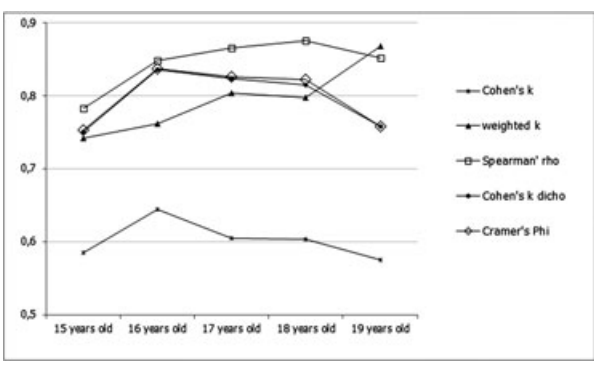

Figure 1 Answers to the ESPAD-Italia® 2004 questionnaire about cannabis use in lifetime, by age; and estimation of Cohen's *k* unweighted and weighted, Spearman's rho, Cohen's *k* for dichotomous and Cramer's phi.

Regarding students aged 17–19 years, we observed a higher concordance in the weighted *k* and Spearman's rho.

Table 6 shows the results of analysis conducted using the model proposed by Lipsitz (2001). Values of γ relate to the covariates of gender and age obtained within a reduced model, which does not include interactions between gender and age classes. Interaction between the covariates were not included in the model because they were never statistically significant. Concordance observed for girls was used as a reference for concordance observed for boys, while concordance for the class age 19 years was used as a reference for concordance in other age classes.

The model was always applied by examining dependent variables in both their ordinal and dichotomous forms. Some drugs are not included in Table 6 since due to the low prevalence rate the sample size was too small to use in this model. The effect of gender adjusted for age was statistically significant in questions about drunkenness ($\gamma = -0.17$; SE = 0.07; $p < 0.01$) and when examining all answers about illegal substances as a whole except for cannabis mixed with alcohol ($\gamma = 0.29$; SE = 0.1; $p < 0.00$). Boys showed a lower concordance for the drunkenness variable and a higher concordance for the variable resulting from the set of answers about illicit substances, except for cannabis mixed with alcohol.

Table 6 Estimation of γ relating to the covariates of gender and age class used for the multiple regression applied to answers about lifetime use of cigarettes, drunkenness, cannabis, tranquillizers or sedatives, all illicit substances except for cannabis mixed with alcohol

		Males			Females	Age 15			Age 16			Age 17			Age 18			Age 19
Questions	Variable	γ	SE	$p <$	γ	γ	SE	$p <$	γ	SE	$p <$	γ	SE	$p <$	γ	SE	$p <$	γ
Use of cigarettes	Dichotomous	0.05	0.05	0.29	REF	−0.02	0.07	0.82	−0.34	0.10	0.00	−0.36	0.10	0.00	−0.14	0.08	0.09	REF
	Ordinall	0.04	0.05	0.40	REF	−0.18	0.08	0.03	−0.14	0.08	0.08	−0.15	0.09	0.22	−0.11	0.08	0.22	REF
Drunkenness	Dichotomous	−0.17	0.07	0.01	REF	−0.03	0.10	0.76	0.01	0.11	0.95	−0.03	0.12	0.79	0.01	0.12	0.97	REF
	Ordinal	−0.07	0.06	0.24	REF	−0.04	0.12	0.75	0.08	0.09	0.39	0.11	0.10	0.25	0.07	0.10	0.46	REF
Cannabis	Dichotomous	−0.03	0.06	0.55	REF	−0.32	0.13	0.01	0.14	0.09	0.12	0.15	0.09	0.10	0.13	0.09	0.17	REF
	Ordinal	0.06	0.06	0.33	REF	−0.37	0.17	0.03	0.13	0.10	0.19	0.15	0.09	0.12	0.11	0.10	0.31	REF
Tranquillizers or sedatives	Dichotomous	0.15	0.20	0.45	REF	0.16	0.39	0.67	0.38	0.27	0.15	0.33	0.55	0.55	0.46	0.40	0.24	REF
All illicit drugs except cannabis mixed with alcohol	Dichotomous	0.29	0.10	0.00	REF	−0.97	0.27	0.00	−0.08	0.16	0.62	−0.13	0.17	0.45	−0.04	0.17	0.80	REF

Int. J. Methods Psychiatr. Res. 21(2): 158–168 (2012). DOI: 10.1002/mpr

As to the effect of age adjusted for gender, if compared with 19-year-old subjects, 15-year-old subjects showed a significantly lower concordance for cigarette smoking ($\gamma = -0.18$; SE = 0.8; $p < 0.03$), cannabis use (dichotomous $\gamma = -0.32$; SE = 0.13; $p < 0.01$ and ordinal $\gamma = -0.37$; SE = 0.17; $p < 0.03$) and for the variable resulting from the set of answers about illicit substances, except for cannabis mixed with alcohol ($\gamma = 0.97$; SE= 0.27; $p < 0.00$). The 16-year-old ($\gamma = -0.34$; SE = 0.1; $p < 0.00$) and 17-year-old subjects ($\gamma = -0.36$; SE = 0.1; $p < 0.00$) showed a significantly lower concordance for cigarette smoking. When the two covariates in the model were statistically significant, it is more relevant to focus on specific values of k for the covariate levels, than on the comprehensive value of k.

Discussion

The study shows a high internal consistency and a high test–retest reproducibility of the ESPAD questionnaire after a three-week re-administration interval to gather information about tobacco, alcohol and cannabis use and consumption of other illicit substances among the Italian school population.

Questionnaires in which at least two answers regarding the same substance were incongruent (i.e. last-year use higher than lifetime use) were rejected. Test–retest reliability without incongruent cases leads to higher reliability than exists "*in vivo*". Because these cases usually contribute to the unreliability of the test and at the same time to values of prevalence rates, it was found that by analysing all questionnaires, the cases eliminated ($n = 11$), did not significantly affect the k of Cohen values (weighted and unweighted), Spearman's rho and Cramer's phi reported in Tables 4 and 5. Thus we chose to use the model that eliminates incongruent cases in the same way as the one generally used for prevalence analysis in the ESPAD study (see Methods section).

The stratified analysis shows a difference between the frequency of use reported by subjects whose answers allowed matching of the questionnaires, and subjects who were not matching because they were not at school, did not complete the subject code, or completed it in the wrong way.

This difference maybe mostly explained by subjects who did not complete the subject code or who completed it erroneously, as it is reasonable to assume that there is no difference between subjects who took part only in the first measurement and not in the second because they were not at school on that day. This result suggests that subjects who report a higher consumption tend not to complete the subject code, or it could also be explained by the fact that those who take drugs are rather absent and not as conscientious as the others. The sensitivity of collecting data on drug use has always made reliability an important issue. Survey research on drugs, where questions are asked about socially disapproved and illegal behaviours, may generate inaccurate reporting and bias in survey estimates. Survey researchers recognize the need to design methods that elicit accurate and truthful reporting of drug use experience. Little research has been conducted on the factors that improve an individual's reporting of sensitive information regarding questions about potentially embarrassing or self-incriminating behaviour (Harrison, 1995).

The incongruence analysis suggests the relevance of collecting information about time intervals (use during lifetime, during the last 12 months and the last 30 days) for each question about substance use, not only in order to better describe the data on use experience, but also to verify the consistency of the given answers and thus of the exclusion or the attribution of incongruent subjects with respect to the consumer group.

Evaluation of subject prevalence referring to the use of a particular substance might be corrected by taking into account the weight of other variables among those collected, associated with substance use prevalence (such as gender, age, way of substance use, and so on).

In our study, where missing and incongruent answer frequency is low, this correction would bring about slight changes. The situation under study where frequency of missing or incongruent answers was high, might be very different.

Another point to consider is that this correction can be easily employed through the analysis of statistical procedure, saving the estimated values of the logistic regression function (Bishop *et al.*, 1977). Incongruent answers to the various questions taken into account may be used in the logistic regression model as indicators of the probability of the dichotomic dependent variable related to substance use or disuse (Van Buuren and Van Ruckevorsel, 1992). Moreover the analysis of missing and incongruent answers can help the researcher to better define the survey tool to implement in the study, trying to improve the formulation of questions leading to a higher incongruence frequency or a higher missing answer frequency.

The method proposed by Lipsitz *et al.* (2001) using the linear multiple regression model in order to estimate the k coefficient was useful because it is easy to conceptualize and implement using common software. The results focus on the effect of the gender and age covariates on the concordance of answers about drug use, and suggest the importance of examining the concordance relating to the

covariate levels. Analysis of the effect of the covariates pointed out situations in which no subject reports use of substances. In this case, one or more cells of the test–retest matrix, stratified according the values of the covariate or according their combination (if considering the effect of interaction as well) may remain empty. As a consequence, a value of γ outside the existence field $-1 \leq \gamma \leq 1$ can be obtained.

A further observation regarding γ, which expresses the variation of the value of k at the varying of 1-unity of the value of covariate, is that it represents the difference of concordance between the same levels of the covariate. To prevent biases in the estimate of k, these methods should overfit rather than underfit the models, but overfitting the model (i.e. including non-significant terms in the model) leads to very little increase in the estimated standard error of γ (Lipsitz *et al.*, 2001). The presence of the effect of the covariates in determining the significance of γ (that is, in the case under study, the decrease of k) with respect to different substance typologies (intoxications for males), cannabis use (15-year-old subjects), cigarette smoke (15-,16- and 17-year old subjects), alcohol use associated with other illicit drugs other than cannabis (males and 15-year-old subjects), may be attributed to the high dynamism of first-use behaviours. In fact, the estimate is made on lifetime use, and behaviour incidence may vary with respect to covariates in the period between the first and second questionnaire administration (three weeks).

One of the problems with Cohen's k or weighted k, useful when codes are ordered (Bakeman and Gottman, 1997), is that it does not always produce the same results in the cases with equal agreement between (test and retest) but different prevalence rates for substances (Gwet, 2002). All agreement statistics depend on the magnitude of the trait prevalence rate.

In case of ordinal variables (Table 4) both Cohen's k and weighted k always show values lower than Spearman's rho, apart from the considered substance, and therefore also from the prevalence variable. Moreover weighted k, generally higher than Cohen's k, tends to become similar to unweighted k for prevalence values lower than 10%. In the case of dichotomic variables (Table 5), Cohen's k and Cramer's phi provide similar values. In the future, due to its easy calculation features, for the conservative valence with respect to the estimate of significant concordance between test and retest and for good operation also with low prevalence values (in case of examined substances lower than 10%), it seems to be correct to use Cohen's k both for ordinal and for dichotomic variables.

Comparing the different methods it is possible to observe that the weighted k methodology provides coefficients higher among 19-year-old students that is the subgroup with the major drug consumption; these students may represent the subgroup that more likely could increase their frequency of use within three weeks. In this case the weighted k is similar to Spearman's rho (Figure 1). Analyses of the longitudinal follow-ups of the Monitoring the Future data have also shown relationships among variables to persist over time. Drug use in the years following high school is highly consistent with and predictable from senior year drug use. Moreover, marijuana use was more reliably measured than the use of other illicit drugs (O'Malley *et al.*, 1984). There are undoubtedly multiple influences on respondents in terms of their ability and desire to provide a response. These factors include setting, real or perceived consequences of reporting use, literacy, clarity of questions, and memory. In general the more stigmatized the drug, the more prevalence rates are suppressed (Harrison, 1995).

Limitations of the study

As to the results obtained from the analysis of the reliability of the ESPAD questionnaire in terms of reproducibility evaluated using the test-retest study, and concerning questions about drug use, we observe that the questionnaire is good for measuring the use of substances such as tobacco, alcohol and cannabis. Measurement resulted slightly more difficult in the case of substances with lower prevalence rates.

Some types of errors present in the surveys (Groves, 1987) and in particular missing answers and measurement errors may influence the results. In the present study this occurred for the absent students who did not complete the questionnaire on the day of its administration. The magnitude of this kind of error depends on both the share of the sample from which data are obtained (answer rate), and on the difference between responding and non-responding in the presence or not of the feature to be measured (for instance, substance use). The measurement error, described as the difference between the real value of the feature belonging to the subject participating in the study and the data obtained through the survey, derives from different sources such as the way of carrying out the survey, the way in which the questions are formulated, the context and behaviours during the survey, discretion of the questions, coding errors, the cognitive process underlying the answer and the desire of the interviewed subject to answer truthfully (Sudman *et al*, 1996).

In this case, further studies with larger samples are needed in order to estimate reproducibility of measures

Int. J. Methods Psychiatr. Res. 21(2): 158–168 (2012). DOI: 10.1002/mpr

concerning consumption of drugs with a lower expected prevalence.

Acknowledgements

The author would like to thank Dr Giuseppe Rossi, Dr Patricia Iozzo, Dr Annibale Biggeri for helpful comments and suggestions and Dr Alison Frank for the proof-reading. Funding for this study was provided by Ministry of Welfare and Ministry of Education; the founder had no further role in study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication.

Sabrina Molinaro and Fabio Mariani conceived the study and the design and carried out the data analysis. Valeria Siciliano conducted the statistical analysis, helped identify background information and participated in writing the manuscript. Olivia Curzio and Francesca Denoth assisted with statistical analysis, helped identify background information and participated in writing the manuscript. All authors read and approved the final manuscript.

Declaration of interest statement

The authors have no competing interests.

References

Agresti A. (1990) *Categorical Data Analysis*. Chichester: Wiley.

Bakeman R., Gottman J.M. (1997) *Observing interaction: An introduction to sequential analysis*. Cambridge: Cambridge University Press, DOI: 10.2277/0521574277

Bishop Y.M.M., Fiemberg S.E., Holland P.W. (1977) *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MA, MIT Press, DOI: 10.1177/014662167700100218

Brown J., Kranzler H.R., Del Boca F.K. (1993) Self-reports by alcohol and drug abuse inpatients: factors affecting reliability and validity. *British Journal of Addiction*, **87**(7), 1013–1024, DOI: 10.1111/j.1360-0443.1992.tb03118.x

Cohen J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46, DOI: 10.1177/001316446002000104

Cohen J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**(4), 213–220.

Donner A., Koval J. (1980) The estimation of intraclass correlation in the analysis of family data. *Biometrics*, **36**(1), 19–25, DOI: 10.2307/2530491 DOI:dx.doi.org

Donner A., Koval J. (1996) The statistical analysis of kappa statistics in multiple samples. *Journal of Clinical Epidemiology*, **49**(9), 1053–1058, DOI: 10.1037/h0026256

European Monitoring Centre for Drugs and Drug Addiction (2003) *Annual Report 2003: The State of the Drugs Problem in the Acceding and Candidate Countries to the European Union*, Luxembourg, Office for Official Publications of the European Communities.

Gonin R., Lipsitz S.R., Fitzmaurice G.M., Molenberghs G. (2000) Regression modeling of weighted k by using generalized estimating equations. *Journal of Applied Statistical Science*, **49**(1), 1–18.

Graham P. (1995) Modelling covariate effects in observer agreement studies: the case of nominal scale agreement. *Statistics in Medicine*, **14**(3), 299–310, DOI: 10.1002/sim.4780140308

Groves R.M. (1987) Research on survey data quality. *Public Opinion Quarterly*, **51**(2), 156–172.

Gwet K. (2002) Inter-rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity. Statistical Methods for Inter-rater Reliability Assessment, **2**. http://www.stataxis.com/files/articles/inter_rater_reliability_dependency.pdf [7 June 2010].

Harrison L.D. (1995) Validity of self-reported data on drug use. *Journal of Drug Issues*, **25**(1), 91–111.

Harrison E.R., Haaga J., Richards T. (1993) Self-reported drug use data: what do they reveal? *The American Journal of Drug and Alcohol Abuse*, **19**(4), 423–441.

Hawks R.L., Chiang C.N. (1986) Urine testing for drugs of abuse. *NIDA Research Monograph*, **73**, 84–112.

Hibell B., Andersson B., Ahlström S., Balakireva O., Bjarnason T., Kokkevi A., Morgan M. (2000) The 1999 ESPAD Report. Alcohol and Other Drug Use Among Students in 30 European Countries, The Swedish Council for Information on Alcohol and Other Drugs (CAN) and The Pompidou Group at the Council of Europe.

Hibell B., Andersson B., Bjarnason T., Ahlström S., Balakireva O., Kokkevi A., Morgan M. (2004) The ESPAD Report 2003. Alcohol and Other Drug Use Among Students in 35 European Countries, Stokholm, the Swedish Council for Information on Alcohol and Other Drugs (CAN) and the Pompidou Group at the Council of Europe.

Hibell B., Guttormsson U., Ahlström S., Balakireva O., Bjarnason T., Kokkevi A., Kraus L. (2009) The 2007 ESPAD Report – Substance Use Among Students in 35 European Countries, Stockholm, the Swedish Council for Information on Alcohol and Other Drugs (CAN).

Kokkevi A., Arapaki A.A., Richardson C., Florescu S., Kuzman M., Stergar E. (2007a) Further investigation of psychological and environmental correlates of substance use in adolescence in six European countries. *Drug and Alcohol Dependence*, **88**(2–3), 308–312, DOI: 10.1016/j.drugalcdep.2006.10.004

Kokkevi A., Richardson C., Florescu S., Kuzman M., Stergar E. (2007b) Psychosocial correlates of substance use in adolescence: a cross-national study in six European countries. *Drug and Alcohol Dependence*, **86**(1), 67–74, DOI: 10.1016/j.drugalcdep.2006.05.018

Landis J.R., Koch G.C. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.

Lipsitz S.R., Kim K., Zhao L. (1994) Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine*, **13**(11), 1149–1163, DOI: 10.1002/sim.4780131106

Lipsitz S.R., Williamson J., Klar N., Ibrahim J., Parzen M. (2001) A simple method for estimating a regression model for K between a pair of rates. *Journal of the Royal Statistical Society*, **164**(3), 449–465.

O'Malley P.M., Bachman J.G., Johnston L.D. (1984) Reliability and consistency in self-reports of drug use. *International Journal of the Addictions*, **18**(6), 805–824.

Sherman M.F., Bigelow G.E. (1992) Validity of patients' self-reported drug use as a function of treatment status. *Drug and Alcohol Dependence*, **30**(1), 1–11, DOI: 10.1016/0376-8716(92)90030-G

Shillington A.M., Cottler L.B., Mager D.E., Compton W.M. 3rd (1995) Self-report stability for substance use over 10 years: data from the St Louis Epidemiologic Catchment Study. *Drug and Alcohol Dependence*, **40**(2), 103–109, DOI: 10.1016/0376-8716(95)01176-5

Sudman S., Bradburn N.M., Schwartz N. (1996) *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass Publishers.

Van Buuren S., Van Ruckevorsel J.Z.A. (1992) Imputation of missing categorical data by maximizing internal consistency. *Psychometrika*, **57**(4), 567–580.

Williamson J.M., Manatunga A.K. (1997) Assessing interrater agreement from dependent data. *Biometrics*, **53**(2), 707–714.

Wolff K., Farrell M., Marsden J., Monteiro M., Ali R., Welch S., Strang J. (1999) A review of biological indicators of illicit drug use, practical considerations and clinical usefulness. *Addiction*, **94**(9), 1279–1298, DOI: 10.1046/j.1360-0443.1999.94912792.x