UNITED NATIONS INTERNATIONAL DRUG CONTROL PROGRAMME
Vienna

BULLETIN ON NARCOTICS

Volume LII, Nos. 1 and 2, 2000

Economic and social costs of substance abuse

UNITED NATIONS
New York, 2001

UNITED NATIONS PUBLICATION
ISSN 0007-523X

Preface

The *Bulletin on Narcotics* is designed to provide information on developments in drug control at the local, national, regional and international levels that would benefit the international community. It is a United Nations publication that is available in Arabic, Chinese, English, French, Russian and Spanish.

Individuals and organizations are invited by the Editor to contribute articles to the *Bulletin* dealing with policies, approaches, measures and developments (theoretical and/or practical) relating to various aspects of the drug control effort. Of particular interest are the results of research, studies and practical experience that would provide useful information for policy makers, practitioners and experts, as well as the public at large.

The present issue of the *Bulletin* contains papers that were originally prepared for the Third International Symposium on the Economic and Social Costs of Substance Abuse, organized by the Canadian Centre on Substance Abuse in Banff, Alberta, Canada, from 31 May to 3 June 2000.

Editorial policy and guidelines for publication

All manuscripts submitted for publication in the *Bulletin* should constitute original and scholarly work that has not been published elsewhere or is not being submitted simultaneously for publication elsewhere. The work should be of relatively high professional calibre in order to meet the requirements of a United Nations technical publication. Contributors are kindly asked to exercise discretion in the content of manuscripts so as to exclude any critical judgement of a particular national or regional situation.

The preferred mode of transmission of manuscripts is in Word format. Each submitted manuscript should consist of an original hard copy and a 3.5” diskette, in Word for the text and Excel or Lotus for charts and tables, in any of the six official languages of the United Nations. The manuscript should be accompanied by an abstract of approximately 200 words and by a complete set of references numbered in the order of their appearance in the text. The manuscript should be between 10 and 20 double-spaced typewritten pages, including tables, figures and references. Tables should be self-explanatory and should supplement, not duplicate, information provided in the text.

Manuscripts, together with brief curricula vitae of their authors, should be addressed to the Editor, *Bulletin on Narcotics*, United Nations International Drug Control Programme, P.O. Box 500, A-1400 Vienna, Austria. A transmittal letter should designate one author as correspondent and include his or her complete address, telephone number and, if available, facsimile number and e-mail address. Unpublished manuscripts will be returned to the authors; however, the United Nations cannot be held responsible for loss.

Views expressed in signed articles published in the *Bulletin* are those of the authors and do not necessarily reflect those of the United Nations Secretariat. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat concerning the legal status of any country, territory, city or area, or its authorities, or concerning the delimitation of any frontiers or boundaries.

Material published in the *Bulletin* is the property of the United Nations and enjoys copyright protection in accordance with the provisions of Protocol 2 annexed to the Universal Copyright Convention concerning the application of that Convention to the works of certain international organizations.

Reprints, purchases and subscriptions

All issues of the *Bulletin* (from vol. I, No. 1 (1949), to the present issue) are available on the home page of the United Nations International Drug Control Programme, at http://www.undcp.org.

The following special issues of the *Bulletin* are also available as United Nations publications:

1989

Drug abuse assessment: double issue (vol. XLI, Nos. 1 and 2)

1990

Emerging directions and trends in drug abuse control (vol. XLII, No. 1)

1991

Involvement of intergovernmental and non-governmental organizations in matters of drug abuse control (vol. XLIII, No. 1)

1992

The role of law enforcement agencies in drug abuse control (vol. XLIV, No. 1)

The environmental impact of drug abuse (vol. XLIV, No. 2)

1993

Policy issues relating to drug abuse and the human immunodeficiency virus (HIV) (vol. XLV, No. 1)

Drug testing in the workplace (vol. XLV, No. 2)

1994

The family and drug abuse (vol. XLVI, No. 1)

General issue on drug abuse (vol. XLVI, No. 2)

1995

Special issue on gender and drug abuse (vol. XLVII, Nos. 1 and 2)

1996

Special issue on rapid assessment of drug abuse (vol. XLVIII, Nos. 1 and 2)

1997 and 1998

Double issue on cannabis: recent developments (vol. XLIX, Nos. 1 and 2, and vol. L, Nos. 1 and 2)

1999

Occasional papers (vol. LI, Nos. 1 and 2)

Requests for permission to reprint signed material should be addressed to the Secretary of the Publications Board, United Nations, New York, New York 10017.

Correspondence regarding the purchase of copies of and subscriptions to the *Bulletin* should be addressed as follows:

For Asia, North America, Oceania and South America:

The Chief
Sales and Marketing Office in New York
United Nations Publications
United Nations Headquarters
New York, New York 10017
United States of America

For Africa, Europe and the Middle East:

The Chief
Sales and Marketing Office in Geneva
United Nations Publications
United Nations Office at Geneva
Palais des Nations
CH-1211 Geneva 10
Switzerland

CONTENTS

Offprint from Bulletin on Narcotics, vol. LII, Nos. 1 and 2, 2000

Introduction: Improving economic data to inform decisions in drug control

D. COLLINS

Adjunct Professor in Economics, Macquarie University, Sydney, Australia

H. LAPSLEY

Senior Lecturer in Health Economics, University of New South Wales, Sydney, Australia

J. LECAVALIER

Senior Associate, Canadian Centre on Substance Abuse, Ottawa, Canada

E. SINGLE

Research Associate, Canadian Centre on Substance Abuse, and Professor of Public Health Sciences, University of Toronto, Canada

Background

There is enormous variation between countries and regions of the world with regard to illicit drug use and the problems associated with illicit drugs [1]. In South America and south-east and south-west Asia, where a large share of the world's cocaine and heroin is produced, the illicit drug trade has created substantial underground economies with problems for law enforcement and economic control. In the drug-consuming countries in Europe and North America, the problems are more related to the adverse health, social and economic consequences of illicit drug use.

Although international treaties provide a common framework for drug policy and a certain degree of uniformity in social responses to the problems caused by illicit drugs, there is also wide variation in national drug policies. Responses to illicit drug problems range from strict enforcement of punitive drug laws to benign neglect. In parts of Australia, Europe and North America, harm reduction policies have been implemented to reduce the adverse consequences of illicit drug use for users who cannot be expected to cease their drug use at the present time [1]. Even within countries, there are often cycles of panic over emerging drug problems, followed by periods of indifference when other pressing issues push the problem of illicit drugs to a relatively low place on the national policy agenda [2].

Thus, there are substantial differences in the nature and magnitude of illicit drug problems as well as the social responses to such problems, over time, both between countries and regions of the world and within countries. Nonetheless, there is a strong consensus among people involved in addressing addiction issues that success requires long-term commitment and investment. As for vaccination programmes, progress can be made only over extended periods of time as consistent efforts are applied to each new generation. There is, therefore, a strong need for more rigorous and comprehensive economic data on substance abuse to promote evidence-based decision-making and a more consistent response to substance abuse. The economic ramifications of illicit drug use are not well understood, either in the producing countries or the consuming countries and regions.

The purpose of the series of articles appearing in the present issue of the *Bulletin on Narcotics* is to bring together related approaches of several authors who examine various methodological issues and data requirements involved in developing more complete, reliable and comparable data on economic aspects of substance abuse. The articles are based on papers presented at the Third International Symposium on the Economic and Social Costs of Substance Abuse, organized by the Canadian Centre on Substance Abuse in Banff, Canada from 31 May to 3 June 2000. The aim of the symposium was to extend the scope of the cost estimation of substance abuse and to facilitate improvements in cost-estimation methodology.

In the article entitled “Economic evaluation of policies and programmes: further uses of estimates of the social costs of substance abuse”, Collins and Lapsley review areas of cost estimation that still need to be addressed or that are in need of further development. It proceeds to consider the extension of the use of the data derived from those studies into the more policy-oriented areas of drug programmes and project evaluation. The other articles deal with particular issues referred to in general terms in the Collins and Lapsley article.

In the article entitled “The cost to employers of employee alcohol abuse: a review of the United States literature”, Harwood and Reichman review literature from the United States of America concerning the costs borne by employers as a result of alcohol abuse by employees. The authors’ analysis of the types of impact of alcohol abuse is also relevant to the workplace effects of smoking and illicit drug use.

One of the most intractable issues in the cost estimation of substance abuse has been how to identify the extent of crime that is attributable to alcohol and drug use. Crimes may be associated with drug use but such an association does not necessarily imply a causal relationship. In the article entitled “Attributable fractions for alcohol and drugs in relation to crime in Canada: conceptualization, methods and internal consistency of estimates”, Pernanen and Brochu examine the problems involved in developing crime-attributable fractions in Canada.

There are difficulties in identifying the size of public expenditure that arises as a result of substance abuse. Drug-related expenditures can be spread across the budgets of many individual government departments and the issue of the development of attributable fractions arises once again, in areas such as health and justice where similar types of expenditures can be attributable to a range of disparate causes. In the article entitled “Estimating the costs of substance abuse to state budgets in the United States of America”, Foster and Modi present the analysis underlying a study that attempts to estimate the costs of substance abuse across the complete range of state budgetary units of the United States.

Most of the existing research into the social costs of substance abuse has been undertaken in and for Western developed economies. Those results are likely to be of limited applicability to developing economies that may have radically different institutional structures and may face different types of drug problems. In the article entitled “Estimating the economic costs of drug abuse in Colombia”, Pérez and

Wilson, drawing on the Colombian case, illustrate the types of issues that may limit the applicability of existing research results to developing countries.

The trade in illicit drugs leads to extensive but largely unquantified underground economic activities. Such activities involve various forms of crime, including drug-dealing, violence, tax evasion and smuggling. The issues involved in estimating the size of the shadow economy are reviewed in the article by Schneider entitled “Illegal activities and the generation of value added: size, causes and measurement of shadow economies”.

In the present introduction, the key issues that policy makers must address in dealing with illicit drug problems are first discussed. A conceptual framework for estimating the economic costs of illicit drug use is then presented. Focus is placed on the data requirements, methodological issues and prospects for future development of cost estimation studies in more countries and regions of the world. The problems involved in estimating the avoidable costs of illicit drug abuse and the cost-effectiveness of interventions are then addressed. A summary of the present situation regarding the current understanding of economic aspects of illicit drugs and a discussion of the implications for data systems and research on drug problems conclude the introduction.

Key substance abuse issues from a policy-making perspective

Four key issues need to be addressed to help policy makers make well-informed decisions on drugs:

(a) The cost of drug abuse to society;

(b) The portion of those costs that are realistically avoidable;

(c) What investments policy makers should make to avoid such costs and where they should make them;

(d) How well such investments are performing over time.

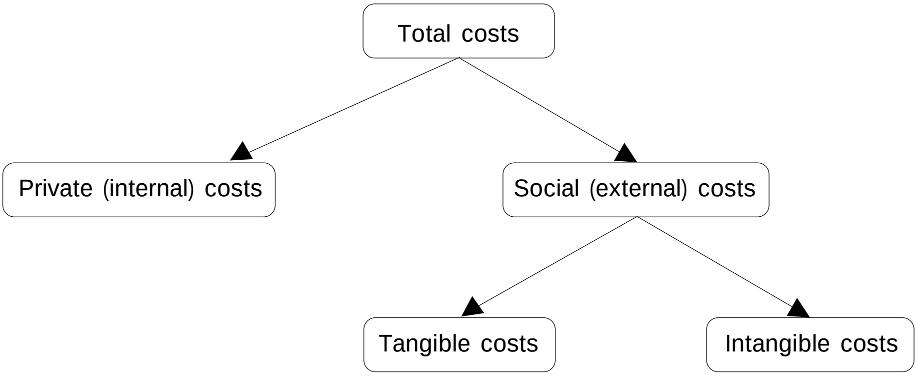
With regard to those issues, researchers need to pay greater attention to economic aspects of drug abuse and interventions. Policy makers are rarely concerned with the details regarding the data requirements and methodological issues involved in addressing those issues. Nevertheless, those aspects need to be addressed at least in part, if policy makers are to make well-informed, empirically-based decisions on drug policy.

Substantial attention has been paid by researchers to the costs that society bears as a result of the use and abuse of alcohol and tobacco. Very little attention, on the other hand, has been paid to the costs borne by society as a result of the abuse of illicit drugs. There is no doubt that such a deficiency can be attributed largely to the serious data problems inherent in any attempt to quantify the social costs of illicit drug use. It is difficult to quantify the production, consumption, import, export or price of illicit drugs. In addition, although significant information is available on the causal links between drug abuse and health, the causal links in other areas, in particular crime, are extremely difficult to quantify. For those reasons, little quantitative information exists on the social costs of illicit drug abuse.

Although it is inevitable that economic cost data in that area will be deficient, it is possible to provide much more comprehensive information than is currently available. Reasonable estimated values can be placed on some drug abuse costs, narrowing the range of costs on which policy makers will be required to make qualitative judgements. That is not an unusual circumstance. In most social benefit-cost analyses, some benefits or costs cannot be valued and so must be judged qualitatively against those which can be quantified. Assigning values where it is possible to do so narrows the range of uncertainty in the decision-making process.

The composition of the total community costs of drug abuse is illustrated in figure I below.

Figure I. Community costs associated with drug abuse



Private costs are those borne by the drug abusers themselves, having made a rational decision to consume in the full knowledge of the effects upon them of drug abuse. If drug abusers do not bear the full costs, for example, if their health costs are subsidized from the public purse, then those costs become the concern of public policy. If the actions of drug abusers are determined by perceived costs that are less than actual costs, the difference between the two can be viewed as the social cost because the abusers have not adjusted their behaviour to reflect those higher costs and so the latter are not accounted for [3]. The three conditions to be satisfied simultaneously if costs are to be classified as private costs are as follows: *(a)* costs fully borne by abuser; *(b)* full knowledge of the effects of drug abuse; and *(c)* rationality in decision to abuse drugs. That is a stringent set of requirements. In practice it means that a high proportion of the total costs of drug abuse is likely to be external rather than internal.

The social (external) costs of drug abuse can be subdivided into tangible and intangible costs. Tangible costs are those that, if reduced, would release resources to the rest of the community for alternative consumption or investment purposes. For example, a reduction in public health-care costs will release resources for government expenditure in other areas. A market exists for the resources used in those areas, therefore, it is possible to attach a price to them.

Intangible costs (for example, loss of life, pain and suffering), when reduced, do not release resources for other uses. The reduction of such costs is very important but does not yield benefits that can be redistributed to other people. No market exists in such benefits, thus it is difficult to assign a value to them.

The difficulty of valuing intangibles should not lead to the avoidance of including those costs in policy analysis. It is possible that a reduction in drug abuse might lead to an increase in certain tangible costs. If the inevitable reduction in intangible costs is not set against the increase in tangible costs, the erroneous conclusion might be drawn that drug abuse was in the public interest.

The example of health-care costs attributable to drug abuse illustrates the point. It is possible that the lifetime health-care costs of abusers are less than those of non-abusers, because the former have lower life expectancies and so draw upon the health-care resources of the community for a much shorter period. There is some evidence that that might be the case for smokers; however, the necessary research has not been undertaken in relation to drug abusers. If it were assumed, for the purposes of argument, that drug abusers imposed lower lifetime health costs than non-abusers, the nonsensical conclusion might be drawn that drug abuse was, in some sense, in the public interest. Such a conclusion could only be drawn if the high intangible costs of abuse (such as loss of life, pain, suffering and bereavement) were ignored, and they clearly should not be ignored.

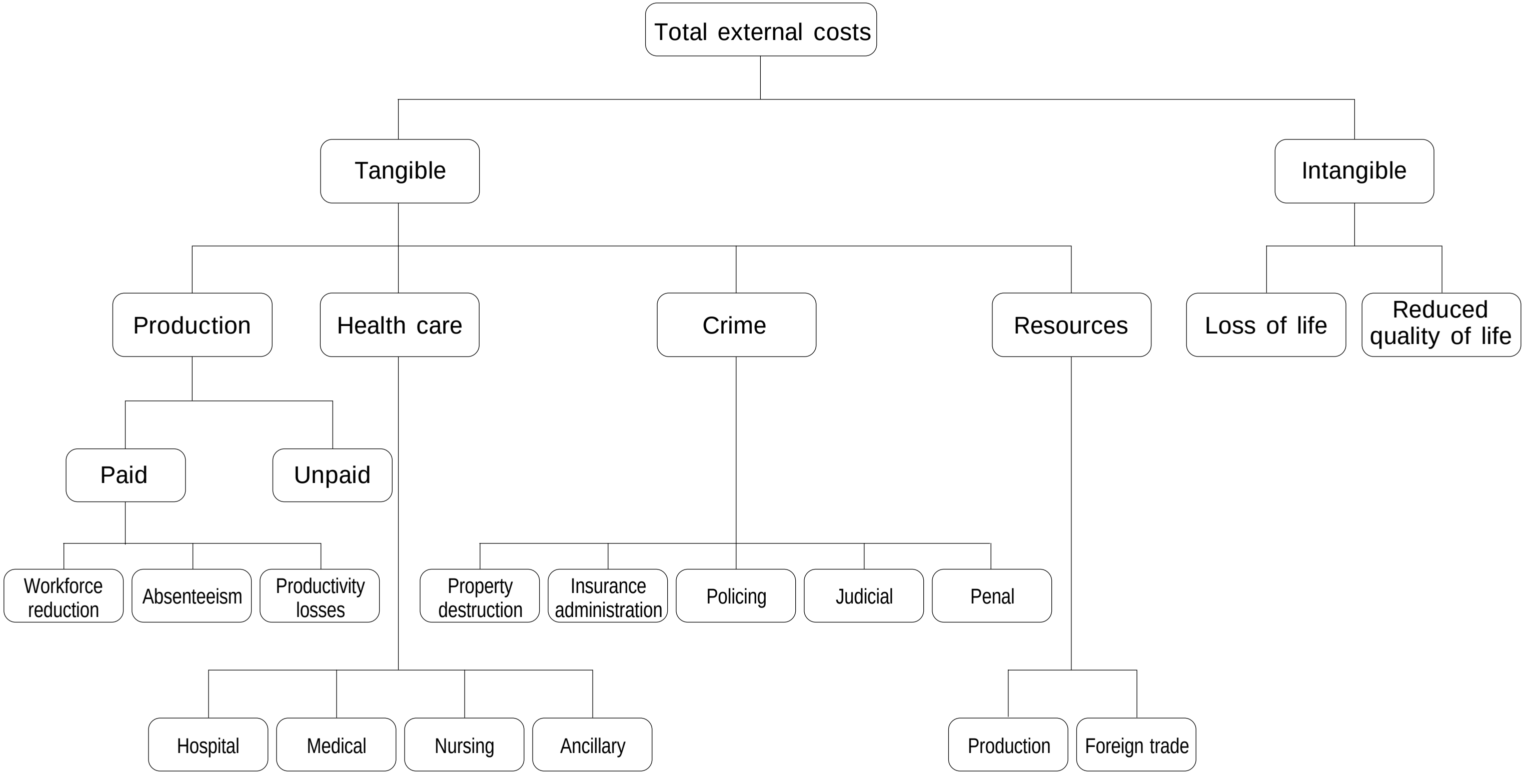
A summary of the types of external costs of drug abuse is presented in figure II. Production losses can occur in the paid workforce but they can also represent the loss of unpaid work (for example, household work and volunteer and community work). Unpaid work, though productive, is not counted in conventional national accounts statistics. It should, however, be taken into account in the assessment of production losses attributable to drug abuse.

Drug-attributable mortality can reduce the size of the workforce. Its impact on the level of employment may be alleviated by the existence of a pool of unemployed individuals who are ready to fill the employment gap, although the skills available among the unemployed may not match the skills lost as a result of drug-attributable mortality. Drug abuse may also result in reduced production, as a result of increased absenteeism or reduced on-the-job productivity arising from drug-attributable morbidity.

Health-care costs result from a range of medical conditions attributable to drug abuse. As indicated above, health costs in any given period of time will also be affected by the previous premature mortality of drug abusers who otherwise would still have been alive and imposing costs on the health-care system. Given that the health-care costs of drug abuse come earlier in the typical life cycle than the health-care benefits (that arise from the foregone use of health-care services by those who die prematurely), a drug abuse epidemic would lead to a substantial increase in current health-care costs for a considerable period of time, even if the lifetime costs of drug abusers were lower than those of non-abusers. The health-care costs would exceed the health-care benefits.

Crime costs relate to the policing, judicial and penal expenditures directly and indirectly attributable to drug-related criminal activity. That is one of the most difficult areas of quantification of drug-attributable costs. Although it is widely believed that a significant proportion of burglaries and physical assaults is drug-related, crime statistics are usually inadequate for the purposes of identifying such a relationship. Most of the property stolen by drug abusers is redistributed rather than destroyed and thus, from a social point of view, cannot be treated as a cost of illicit drug use. Only the value of property that is destroyed and the insurance administration associated with losses by drug-related theft should be incorporated into that component of drug-attributable costs.

Figure II. External costs



The production and consumption of illicit drugs involve the use of resources that would otherwise be available for alternative production or consumption purposes. If the drugs are produced domestically, the production resources can be considered to be a cost of illicit drug use. If the drugs are imported, they are drawing on foreign exchange that would otherwise be available for the purchase of other goods or services.

It is debatable whether public expenditures on education and research related to drug abuse should be counted as a drug-attributable cost. On the one hand, those expenditures can be considered to be discretionary policy responses to drug abuse rather than directly attributable to that abuse. On the other hand, it could be argued that such programmes should yield benefit-cost ratios of at least unity, so that in the absence of the programmes, social costs would have been higher by at least the cost of those programmes. In the absence of a resolution of that issue, a compromise is to distinguish the social cost estimates from the costs of research and education programmes related to drug abuse.

Intangible costs are additional costs borne by society over and above the tangible resource costs. For example, the premature death of a drug abuser of working age will cause an actual or potential loss of production, but that is not the total loss to society. Such deaths will cause suffering and bereavement to others; the abusers themselves, like most members of society, are likely to value their lives more highly than simply the value of their production. Many societies place a higher value on life as shown by the substantial resources devoted to extending lifespans and improving the quality of life of citizens who are over the age of retirement.

The cost estimates referred to here are aggregate cost estimates of illicit drug abuse. They attempt to compare the actual situation of drug abuse with a hypothetical counterfactual state that would have existed had there been no drug abuse. Such a comparison yields the total costs attributable to drug abuse. There is no suggestion that the counterfactual situation is achievable and, for that reason, no suggestion that there exists a set of feasible policies capable of reducing the drug abuse level to zero. One implication of this is that there is no conceivable set of public policies capable of eliminating the costs of drug abuse. A significant proportion of current costs arises as a result of past abuse. In addition, there can be no prospect of reducing drug abuse to zero. The component of abuse costs that is susceptible to elimination as a result of public policies can be identified as avoidable costs. The extent of avoidable costs indicates the potential benefits available to public expenditures on drug policies. The major public controversies about the most appropriate approaches to drug policies, for example, zero tolerance or harm minimization, make it difficult to estimate what proportion of the total costs of drug abuse is avoidable.

The most efficient drug policies will be those that yield the highest social rate of return. Without the information underlying social cost analysis, it is not possible to undertake policy and programme evaluation. In spite of the high cost of many public policies and programmes, surpassingly little formal evaluation of them is undertaken. In general, there is little information available about how well those investments are performing over time.

A conceptual framework for estimating the costs of substance abuse

Estimates of the economic costs of drug abuse serve several purposes. Firstly, economic cost estimates are frequently used to argue that policies on drugs should be given a high priority on the public policy agenda. Without such a standard for assessing cost estimates, there is a tendency by the advocates for each social problem to overbid, adding in items to make their concern a suitably high (even exaggerated) number. Secondly, cost estimates are helpful in targeting specific problems and policies. It is important to know which aspects of drug abuse involve the greatest economic costs. The specific types of cost may also indicate specific areas where public attention is needed or where specific measures may be effective. Thirdly, economic cost studies are helpful in identifying information gaps, research needs and desirable refinements to national statistical reporting systems. There is probably no better way to set a national research agenda than to conduct a cost estimation study and use it to map out key areas and topics where information is lacking.

The development of improved estimates of the costs of drug abuse also offers the potential to provide baseline measures for determining which policies and programmes are the most effective. International comparisons of reliable cost estimates could provide important indicators of the effectiveness of national policies. Such comparisons could indicate, for example, whether the costs of drug abuse are lower or higher in less restrictive societies, whether the social costs of cannabis are greater in countries where it has been decriminalized or whether there is less drug abuse in countries where a greater proportion of the costs are borne by the individual, other things being equal. Ultimately, cost estimates could be used to construct social cost functions for optimal tax policy and national target-setting. Perhaps most immediately promising is the prospect for cost estimates to be extended to more comprehensive cost-benefit analyses of specific drug policies and programmes.

The present section includes a brief discussion of cost estimation studies to date and a description of a new set of international guidelines for estimating the economic costs of substance abuse. The data requirements for carrying out cost estimation are detailed and areas for further development are discussed.

Cost estimation studies

As noted in a recent review [4], the majority of studies estimating the economic costs of drug abuse have been conducted in developed countries, in particular in Australia [5, 6], Canada [7, 8], Switzerland [9], the United Kingdom of Great Britain and Northern Ireland [10] and the United States [11-13]. Those studies generally use a prevalence-based approach, measuring the costs in a given year associated with the prevalence of substance-related morbidity and mortality attributable to past and present substance use.

The vast majority of cost estimation studies use a cost-of-illness approach, in which the impact of drug abuse on the material welfare of a given society is estimated by examining the direct costs of resources expended for treatment, prevention, research and law enforcement, plus losses of production due to increased morbidity and mortality, relative to a counterfactual scenario in which there is no drug abuse [14, 15]. The focus of cost-of-illness studies is on the tangible social costs of substance abuse, which equal the sum of the private and external costs after adjusting for transfers within society. An alternative model is the "externality" approach, which strictly limits estimates to external costs [12].

Within the cost-of-illness framework, there are two major variations regarding the valuation of productivity losses due to premature mortality attributable to drug use: the more commonly adopted human capital approach (see, for example, Rice and others [11]) and the more recent demographic approach pioneered by Collins and Lapsley [5]. In the human capital approach, the lost value of a deceased worker's production is estimated by present earnings plus a discounted rate of future earnings. The demographic approach compares the actual population size and structure to that of an "otherwise healthy" population, that is, an alternative population in which there were no drug-related deaths. Those two approaches are complementary rather than contradictory. The demographic approach is based on the supposition that there had never been any substance abuse or problems associated with the use of psychoactive substances. The human capital approach is based on the supposition that all substance abuse and problems associated with the use of psychoactive substances were to end immediately. The human capital approach generates an estimate of the present and future costs attributable to drug-related mortality in the current year, while the demographic approach estimates the present costs of drug-related mortality in past and present years.

The results of cost estimation studies show sizeable costs attributable to drugs (see table). There is enormous variation in the results. Cost estimates in the 1980s ranged from 349 United States dollars per capita in Canada in 1984 [7] to US$ 51 in Switzerland in 1988 [9]. It is difficult to compare the results of those studies, however, because of differences in methodology and study design. For example, the estimate for Canada included the costs of licit drugs while the estimate for Switzerland was limited to illicit drugs. In all of the studies, the largest economic cost is the productivity losses from drug-related morbidity and premature mortality [4].

Comparison of selected estimates of the economic costs of illicit drug abuse in various countries

Author of the study (and year of publication)	*Country*	*Year in which the data were gathered*	*Original total cost[a] estimate (millions of local currency units)*	*Cost (millions of US dollars)*	*Cost per capita (US dollars)*
Studies undertaken prior to the development of the international guidelines for estimating the costs of substance abuse					
Adrian and others (1989)[b, c]	Canada	1984	Can$ 11 840	8 960	349
Rice and others (1990)[b]	United States	1985	US$ 44 050	44 050	185
Collins and Lapsley (1991)	Australia	1988	$A 1 441	1 233	75
Faze and Stevenson (1990)[c]	United Kingdom	1988	£1 820	3 293	58
Institut suisse de prophylaxie de l'alcoolisme (1990)	Switzerland	1988	SwF 514	342	51
Studies utilizing the international guidelines					
Single and others (1996[d] and 1998)	Canada	1992	Can$ 1 371	1 079	38
Collins and Lapsley (1996)	Australia	1992	$A 1 684	1 160	66
Harwood and others (1998)	United States	1992	US$ 98 000	98 000	384

Source: Exchange rates are based on International Monetary Fund, *International Financial Statistics Yearbook 1997*, pp. 14-15.

[a] Including all indirect and direct costs, unless otherwise indicated.

[b] Including costs of licit as well as illicit drug abuse.

[c] Including estimates of external costs only.

[d] E. Single and others, *The Costs of Substance Abuse in Canada* (Ottawa, Canadian Centre on Substance Abuse, 1996).

International guidelines for estimating the economic costs of drug use

Because of the difficulties in comparing the results of cost estimation studies in different countries, two international symposia were held in Canada under the auspices of the Canadian Centre on Substance Abuse with funding provided by the United Nations International Drug Control Programme and a number of provincial, national and international agencies specializing in addictions. The symposia resulted in the development of a set of guidelines for estimating the costs of substance abuse [15].

The guidelines begin with a discussion of the purposes of estimation studies. Instead of giving details on the exact procedures to be followed in every setting, the guidelines provide a general framework for the development of cost estimates and include a matrix of the types of costs to be considered and a detailed discussion of the following theoretical issues:

(a) Definition of substance abuse;

(b) Determination of causality;

(c) Comparison of the demographic and human capital approaches to cost estimation;

(d) Treatment and measurement of addictive consumption;

(e) Treatment of private costs;

(f) Measurement of intangible costs;

(g) Treatment of non-workforce mortality and morbidity;

(h) Treatment of research, education and law enforcement costs;

(i) Estimation of avoidable costs;

(j) Budgetary impact of substance abuse.

The guidelines conclude with a discussion of future directions that places emphasis on the need to include developing countries in economic cost studies and the implications of the guidelines for research agenda and data collection systems.

The development of the international guidelines has resulted in fewer differences in the methodological approach followed in different cost estimation studies, as the more recent studies have used the same basic cost-of-illness methodology. Nonetheless, there is still considerable variation in results (see table). The estimated costs of illicit drug abuse in 1992 range from US$ 38 in Canada [8], to US$ 66 in Australia [6] and to US$ 384 in the United States [13]. The higher estimates in the study conducted in the United States may reflect the fact that there is greater illicit drug use, as well as greater problems and costs arising from such use, in that country, but part of the variation in findings may have resulted from differences in how estimates were made of drug-attributed mortality and morbidity.

Types of costs and data requirements in cost estimation studies

Although studies vary with respect to how the economic costs of drug use are combined into different categories, the costs described below have often been included in prior cost estimation studies [4, 16] or have been recommended for inclusion [17]. The data requirements for each of those major types of costs are also noted. The general cost categories are presented in approximate order of their magnitude in prior cost estimation studies.

Indirect productivity costs

Indirect productivity costs refer to productivity losses due to premature mortality, lower productivity resulting from drug use (such as absenteeism) and the removal of some individuals from the legitimate market economy due to crime and crime careers. Estimating indirect productivity losses (and health-care costs) firstly requires estimates of mortality and morbidity attributable to drug use. Some deaths and hospitalizations are attributable simply because, by definition, they are caused by drug use (for example, drug-overdose deaths). In other cases, drug use is a contributory cause that accounts for a certain proportion of deaths and hospitalizations. To estimate the portion of specific chronic diseases attributable to drug use, information on the relative risk of drug use is combined with data on the number of persons using drugs at levels associated with a particular relative risk, in order to generate an estimate of the "aetiologic fraction" or attributable proportion of the cause of disease or death that can be reasonably attributed to drug use. For acute conditions where drug use is a contributory but not a necessary cause (for example, assaults, homicide, suicide and motor vehicle trauma attributable to drug use), the attributable fractions must be determined from special studies based on local information. Once the attributable proportion is estimated, it is then applied to the number of recorded deaths and hospitalizations due to the particular cause in order to generate an estimate of drug-attributable mortality and morbidity.

Thus, in order to estimate morbidity and mortality, prevalence data are required on drug use and on drug use by injection. Also required are the recorded number of deaths and hospitalizations, ideally by cause, age and gender; the list of conditions that epidemiological research has shown to be attributable to drug use; and the associated relative risks. Meta-analyses are available that review the epidemiological literature and estimate the relative risk of drug use for various causes of disease and death (see, for example, English and others [18] and Fox and others [19]). Estimates of the attributable fractions for certain causes of death (such as assaults, homicide and suicide) and disease, however, should be based on local information. To standardize the results in terms of per capita rates, data on population structure by age and gender and on life expectancy by age and gender are also required. In order to then estimate the value of lost productivity due to morbidity or premature mortality, data on mean income by age and gender are required (for estimating morbidity costs) and data on present value of lifetime earnings by age and gender are required (for estimating costs of premature mortality).

Health-care costs

Health-care costs comprise treatment in general and psychiatric hospitals, co-morbidity costs, ambulance services, residential care, treatment agencies, ambulatory care (physicians' fees and other professional services), prescription drugs and other health-care costs (such as household help and rehabilitation equipment). The required data include hospitalization costs; physicians' fees, costs of other professional services and number of cases seen by physicians and other professional service providers, by age and gender; ambulance costs (total costs, total number of trips, number of trips for drug-related causes); and costs of pharmaceuticals used to treat drug-related conditions (total number of prescriptions and number of prescriptions by cause).

Law enforcement costs

The costs of law enforcement consist of the portion of police, court, corrections and customs costs attributable to drug use. The costs for enforcement of drug laws per se are relatively uncontroversial and all of the associated costs may be included. An argument can be made to include some of the costs for enforcement of property crime (such as burglary or theft) and violent crime (such as assault and homicide); however, estimates of an appropriate attributable fraction of those crimes that can be causally attributed to drug use are generally lacking.

Prevention, research, training and averting behaviour costs

The costs of prevention, research, training and promotion of averting behaviour include drug prevention campaigns, training for physicians and other health professionals, specialized training on drug issues for law enforcement officials, and programmes for the promotion of averting behaviour (such as community crime prevention). The inclusion of such policy costs is debatable, as they concern discretionary expenditures in response to drug abuse rather than costs directly attributable to drug abuse. In most cost estimation studies, those costs are included but identified separately as policy costs. The inclusion of such costs requires data on the costs for prevention and specialized drug training for the health professionals and law enforcement officials.

Administrative costs for transfer payments

The cost of welfare payments, such as social welfare assistance or workers' compensation for persons disabled due to drug abuse, is limited to administrative costs. That is, the actual payouts to recipients are not included, as the productivity lost due to drug-related morbidity is also included. To count both the productivity lost due to drug-related illness and the welfare payments to persons unable to work because of the effects of drug abuse would represent double counting.

Other direct costs

Other costs include fire and accident damage attributable to substance abuse and direct workplace costs, such as the costs of drug-testing in the workplace or the attributable portion of such schemes as the Employee Assistance Programs and other health programmes.

Selected areas for further development

The choice of which costs to include is not a simple issue and it depends in part on the availability of data. If there are no data, it is perilous to exclude a particular cost item, as ignoring it would effectively treat the cost as zero. Indirect estimates, therefore, may be required for some cost items. For example, estimating productivity lost due to absenteeism, high job-turnover and accidents, in most studies, relies on estimates of the lower earnings of drug users or drug-dependent persons. Such estimates do not provide good control of confounding factors that may account for both drug use and lower productivity. Clearly, more exact estimation procedures would be desirable.

Another issue that requires further development is the attribution of crime to drug use. As Brochu and Pernanen note in their article in the present issue of the *Bulletin on Narcotics,* there is little doubt that drug abuse is a contributory cause for some property crimes and violent crimes. In most countries, high rates of illicit drug use have been found among criminal offenders. In Canada, for example, as many as 80 per cent of convicted criminal offenders reported having used illicit drugs during their lifetime, 50-75 per cent showed traces of drugs in their urine at the time of arrest and close to 30 per cent were under the influence of drugs when they committed the crime for which they were accused. Similarly, disproportionate numbers of drug addicts admitted for treatment have criminal records. Chronic or dependent use of heroin, cocaine or crack, amphetamines or hallucinogens is often implicated as a contributory cause of property crime, particularly burglary and theft. Assault, homicide and other crimes of violence are a result of disputes between sellers and buyers or other sellers in the illicit drug market. Illicit drug users are disproportionately involved in incidents of spouse and child abuse.

While drug abuse is related to crime, the relationship is not always causal, as Brochu and Pernanen point out. The fact that a crime is committed by someone using illicit drugs does not necessarily mean that the use of drugs caused the crime to be committed. There are several plausible causal connections. The pharmacological effects of drugs such as cocaine, other stimulants and phencyclidine (PCP) might induce violence; however, the pharmacological effects of most other illicit drugs would not lead to violent behaviour and some drugs may even have the opposite effect. Most addicts do not commit violent crimes, and those who do commit assault or other forms of violence began doing so before becoming drug-dependent. Thus, the pharmacological effects of drugs are at best only a partial explanation for violent behaviour.

As noted by Brochu and Pernanen, another potential causal connection between drug use and crime is the need for addicts to commit property crimes to support drug use. There are heroin and cocaine addicts who commit property crimes to purchase drugs, but the majority of illicit drug users are not dependent and even most dependent drug users do not commit property crimes. In the majority of cases, addicts who committed property crimes had been engaging in criminal behaviour prior to drug use, and many former addicts continue to commit property crimes even when they no longer use drugs. As with the pharmaceutical explanation, the connection between drugs and crime undoubtedly plays a role in many cases, but it is only a partial explanation.

Brochu and Pernanen also note that some crimes result from territorial disputes between rival distributors, and arguments and robberies involving buyers and sellers on the illicit market. That is most common in areas that are disadvantaged economically and socially, and that have traditionally high rates of violence.

Perhaps the most plausible causal connection is that both criminality and drug use, in particular drug-use dependency, are related to a similar set of socio-demographic and personality variables such as poverty, poor future career or income prospects and low investment in social values. Those factors may represent common underlying causes of both criminality and illicit drug use. Illicit drug use and crime may be mutually reinforcing, but the real cause of both drug use and criminal behaviour may be a complex set of underlying personality and social determinants.

In summary, there is little doubt that drugs are a contributing causal factor in many crimes, but the fact that a crime is committed by a drug addict or by an individual under the influence of drugs does not necessarily mean that the crime can be causally attributed to drug use. Some crimes, in particular crimes of violence, are undoubtedly caused by the use or marketing of illicit drugs. Credible estimates of the proportion of those crimes that are causally attributable to drug use are generally lacking; thus, the article by Brochu and Pernanen fills an important gap, at least with regard to the progress of research in Canada. Similar studies are required in other countries.

Estimation of the economic costs of pharmaceutical abuse is another area in which further research is needed. Most studies are limited to consideration of illicit drug abuse, largely because of difficulties in determining when medication use becomes abuse. A further difficulty in estimating the costs of pharmaceutical abuse is that health-care recording systems, such as the International Classification of Disease, sometimes fail to make a clear distinction between a disorder caused by licit drug use and a disorder caused by illicit drug use.

Further research is also required in order to take advantage of recent advances in willingness-to-pay methods that are used to value the cost of mortality. In the willingness-to-pay approach, information from surveys, insurance data sets or other sources is examined to determine how much people are willing to pay for relatively small changes in the risk of death. From those figures, an estimate of the value of life is produced. While that technique appears to have a reasonably sound theoretical basis, there continue to be considerable problems in terms of the accuracy and consistency of estimates obtained using such an approach. Another difficulty is that the cost estimates from studies utilizing willingness-to-pay methods cannot be compared with the gross domestic product. The results can only be meaningfully compared with the total value of life in a society, a figure that is invariably much higher than the gross domestic product and that generally lacks intuitive meaning. Nonetheless, a growing number of economic studies use willingness-to-pay methods, and future cost estimation studies may be able to utilize that emerging methodology to produce reasonable estimates of the largely intangible costs of premature mortality.

Prospects for expansion of cost studies to more countries

Economic cost studies utilizing the international guidelines are currently under way in several countries in Europe. The data required to conduct such estimates, however, are extremely difficult to obtain in most developing countries. Several developing countries, for example, Chile and Colombia, are planning to estimate the economic costs of drug abuse. A host of economic impacts arising from the production and distribution of illicit drugs will need to be taken into account in those studies, and the international guidelines will need to be revised to take those variables into account.

Prospects for a satellite account for substance abuse in the System of National Accounts

It has been suggested that the development of improved, internationally comparable methods for estimating the costs of substance abuse should be attempted, as far as possible, within the framework of the existing System of National Accounts [15]. The development of estimates of the costs of substance abuse in the framework of the System of National Accounts would be a further step towards the improvement and refinement of national accounting systems, increasing their relevance and usefulness. It took decades for the development of standardized measures such as the gross domestic product under the System of National Accounts, but the ultimate utility of such measures has been worth the time and effort.

Estimating avoidable costs of substance abuse

Economic cost estimates include both avoidable and unavoidable costs. The estimates do not represent the amount of money that could realistically be saved through effective government and social policy and programming. The counterfactual situation in economic cost studies, in which there are no problems associated with drug use, is hypothetical and generally not realizable under any circumstances. Even if completely effective policies could be found with no appreciable costs for enforcement, treatment and prevention programming, implementation would not be instantaneous and there would still be lingering adverse consequences from past use of the psychoactive substances.

Economic cost studies in Australia [5, 6] estimate the percentages of mortality and morbidity, and associated economic costs, that are avoidable. They utilize an “Arcadian normal” [20], which is the lowest age-standardized mortality rate for the relevant mortality or morbidity category among 20 comparable Western countries. The “Arcadian normal” is used to estimate the lowest percentage of preventable morbidity and mortality yet achieved in any of the chosen countries. While that is an extremely conservative assumption, such a method is nevertheless a useful tool for quantifying the percentage of preventable morbidity and mortality and their associated costs, which can be reduced and ultimately avoided.

Many of the estimated avoidable costs of drug abuse may be reduced or eliminated only over long lead times. There are several reasons for the slow reduction in avoidable costs. Firstly, lead times for policy implementation will not be effective instantaneously. Secondly, even after the implementation of effective policies, there will be long lead times before the health effects of policy changes are achieved. It may take years before the health status of former drug users is equivalent to persons who never used drugs. Finally, as some costs apply to premature mortality, it may take years before there is an appreciable decline in deaths attributable to drug use.

Estimating the cost-effectiveness of interventions

Estimation of the cost-effectiveness of interventions to eliminate, reduce or minimize the harm associated with drug abuse requires, firstly, comprehensive data reflecting the costs of abuse. Without such data, it is not possible to determine the total amount of resources that should be allocated to interventions. Some of those costs may be crude estimates that are necessary for policy purposes, to justify and support resources for interventions and their evaluations. Researchers cite burden-of-illness studies in support of research and of interventions; the same estimates should be required for the costs of illicit drugs. Prior to making any decisions on methodological issues, it should be determined whether the cost-effectiveness of only one particular intervention is being considered or whether the initial policy objective is to determine the relative cost of alternative interventions or programmes.

Cost-effectiveness, effectiveness and cost-benefit analysis

For cost-effectiveness, effectiveness and cost-benefit analysis, essential requirements include determining the costs that are to be included, estimating those costs which are considered essential and for which there are no reliable data and specifying how to measure whether the desired outcome or effectiveness has been achieved. In measuring effectiveness or outcome, it is especially important to specify the time period to be considered. If, for example, the intended outcome of a programme is for an addict to be drug-free, the time period by which effectiveness is to be measured must be determined. That specification can be compared with measures such as years of survival in therapeutic drug trials.

The choice of methodology depends on the question to be answered. If there has already been a decision to undertake an intervention, then the appropriate analysis is a cost-effectiveness, or value-for-money, analysis. In a cost-effectiveness analysis, either the required outcome or the amount of money available is specified, that is, one parameter is fixed. The objective is therefore to determine the least-cost way of achieving a specified objective, for example, providing a predetermined number of addicts with detoxification treatment or providing the largest number of addicts possible with detoxification treatment for a predetermined amount of money. In each case, one parameter is fixed and the relative merits of the intervention are not being evaluated.

Cost-benefit analysis should be undertaken only when the decision on whether to proceed with the programme, treatment or intervention has not yet been made. It is important to remember that cost-benefit analysis, valuing all the identifiable costs and benefits with a common unit of value, reaches a conclusion from a societal point of view, regardless of which groups or persons bear the costs and which groups or persons obtain the benefits. For that reason, the results may prove to be unacceptable to politicians and policy makers. Much of the criticism levelled at cost-benefit analysis fails to acknowledge that. It is worth noting that when the term cost-benefit analysis is used by non-economists, what is frequently meant is cost-effectiveness analysis.

Cost-effectiveness of prevention

Cost-effectiveness of prevention requires special consideration. For example, a significant proportion of the costs of illicit drugs borne by governments is associated with law enforcement, both preventive and punitive. Disaggregation of those two aspects of law enforcement costs could require comprehensive research, would be difficult and would inevitably involve some arbitrary judgements, and the results would not necessarily be transferable or generalizable.

Outcome measurement has special difficulties when assessing the cost-effectiveness of prevention. The question is whether outcomes, such as abstinence from illicit drugs, should be applied only to high-risk groups, to target groups by age and geographical location at greatest risk, or to those with an already high participation rate in consumption of illicit drugs. An alternative would be for outcome measures to be applied to the entire population. Another point for consideration is whether preventive strategies, to be most cost-effective, should focus on preventing initial consumption of illicit drugs or an early intervention for those who are already consuming illicit drugs. Relative cost-effectiveness studies may provide answers in terms of likely outcomes, but those outcomes may not be politically acceptable. The illegal status of many drugs of addiction in most societies may make prevention programmes more expensive than if those drugs were decriminalized. If costs of a preventive programme are calculated to be large in return for a very small gain, such programmes may still be funded so that governments can be “seen to be doing something”. That is, a decision to fund a programme may be socially and politically acceptable even when there is a very low cost-effectiveness ratio. While optimization according to cost-effectiveness ratios may not be acceptable, knowledge of ratios can nevertheless be used to inform present and future policy decisions.

Cost-effectiveness of treatment

Alternative treatment modalities are rarely subjected to economic evaluation, though such evaluation presents fewer problems and fewer ambiguities than other areas of drug abuse control. Firstly, the population is much more readily defined than the population for whom prevention strategies may be applicable. Secondly, a number of treatment modalities and therapeutic interventions are currently being provided in a range of countries, so that comparative cost-effectiveness studies within and between countries would be possible. Those treatment programmes include drug-assisted and non-drug-assisted detoxification, psychotherapy, addict support organizations and prescribed methadone. A portion of the costs of some of those programmes may be collected but there is a dearth of economic evaluations of treatment alternatives, making it difficult for policy makers to make informed decisions on such matters.

Law enforcement considerations

Difficulties associated with the disaggregation of law enforcement costs have already been mentioned. The argument concerning costs that are incurred only because a drug is illicit should also be acknowledged, although there is no agreement as to whether such costs should be calculated as part of the costs of illicit drugs. Punitive costs of law enforcement, in particular costs of incarceration, are usually only crude estimates, partly because there are very few data available and because the decisions relating to the inclusion and exclusion of costs are necessarily arbitrary. Although many crimes are associated with the consumption of illicit drugs, it is not appropriate to assign causality to all of those crimes.

Data requirements

Data requirements have been referred to throughout the present discussion. Both gross cost estimates and micro-costing data relating to drug abuse are often non-existent or very crude. In order to improve the use of economic data to enable informed policy decisions to be made, more comprehensive and detailed cost studies are required. There needs to be agreement of objective criteria for measuring outcomes, so that the cost-effectiveness of programmes can be compared over time. Because of the absence of cost data, it is not possible to estimate marginal costs; thus, informed decisions relating to economic size and expansion or contraction of programmes cannot be made.

It should be emphasized that the costs required for estimates include not only money, but also the intangible costs associated with crime, violence, suicide and premature deaths. Economic analysis is needed to provide information on the cost-effectiveness of interventions on both current costs and future costs. When comparisons are being made between the relative effectiveness of programmes designed to reduce drug abuse, some non-paid inputs must also be costed. For example, the costs of services provided by volunteers in non-governmental organizations should be estimated; otherwise real costs will not be identified and transferability will be limited.

The likelihood of compliance and the costs associated with non-compliance require estimation. Thus, there is a need for persons involved in economics and other disciplines, such as behavioural science and epidemiology, to collaborate in the design and evaluation of effective interventions. The estimation of outcome probabilities can be combined with economic estimates to inform policy decisions about programmes for prevention and intervention. While evidence such as that obtained from randomized controlled trials is not usually available to estimate outcome probabilities, the relative rigorousness of the available data should be acknowledged and addressed.

Determining the right policy mix

Drug policy inevitably involves a mixture of intervention strategies. Illicit drug use is discouraged through the application of criminal laws against the production, distribution and use of illicit drugs. Treatment is provided in various ways and preventive education may be conducted through school programmes and the mass media. Long-term investments include basic research on biomedical aspects of drug use, socio-behavioural risk factors and specialized training for health-care and law enforcement personnel. A key task for policy makers is determining the most appropriate mix of the various strategies, all of which are directed at reducing drug-related problems and associated costs. Enhanced research on the cost-effectiveness of specific policies and programmes not only helps in deciding which interventions should continue to receive support and funding, but it also helps policy makers determine the most appropriate mix of strategies to achieve the overall goals of a national drug strategy.

Estimating performance over time

Knowledge of the costs attributable to drug abuse at one point in time is of limited value. While the magnitude of the estimated costs might help in setting general priorities, indicating the importance of drug abuse on the political agenda in comparison with competing concerns, the true value of cost estimates can only be realized when estimates are available from a series of cost estimation studies indicating trends in total costs and the various cost components.

The utility of cost estimates over time is enhanced when the specific outcome indicators and performance measures are set out. That focuses the assessment of drug programming and policy on the specific impacts on cost components. For example, the initiation of enhanced outreach and treatment programming for intravenous drug users might be targeted to the reduction of drug-related crime and crime costs.

While the attribution of cost reductions to specific programmes and policies is at best a difficult undertaking, changes in specific costs attributable to drug abuse can help assess the effectiveness of programming and serve as a general barometer of the effectiveness of a drug strategy. The process is dialectic in nature: the policy and programme mix must be periodically evaluated and adjusted, based on changes in the economic costs and other considerations.

Efforts should be made to promote comparability, across countries and world regions, over time. The development of the international guidelines for estimating the costs of substance abuse is only a first step in that direction. The guidelines should be periodically reassessed and revised, based on the experience in applying them in different settings.

Conclusions

In the present article, the authors examined methodological issues involved in developing more rigorous and comparable data on economic aspects of substance abuse. After discussing the conceptual framework for estimating the economic costs of illicit drug use, it was noted that there were numerous contentious methodological issues and data requirements. Nonetheless, the prospects for future development of cost estimation studies have been improved by the recent development of international guidelines and the growing experience in cost estimation in a number of countries.

There is a great need for improved data on economic aspects of drug issues. Economic cost studies can identify the economic impact of drug abuse on the total value of goods and services in the economy and specify which sectors of the economy bear the greatest costs. Although not all of the economic costs are avoidable, reasonable estimates of avoidable costs are possible. The continuing development of economic analyses, such as cost-effectiveness studies and cost-benefit analyses of policies and programmes, should direct policy makers to make better-informed choices on where to invest and help them to assess performance.

Improving data on economic aspects of drug abuse will necessarily entail improving data collection systems. Economic cost studies to date have identified many knowledge gaps. Particular care must be taken in estimating productivity losses due to drug abuse and in developing a reasonable estimate of the proportion of crime that can be causally attributed to drug use.

There are many potential benefits to be gained by filling such knowledge gaps and improving the general level of understanding of the economic ramifications of substance abuse. Policy makers want to know the answers to four key questions:

(a) The costs of drug abuse to society;

(b) The proportion of those costs that are avoidable;

(c) How best to invest in order to avoid or minimize such costs;

(d) How well those investments are doing.

More complete and rigorous data on economic aspects of drug abuse would provide policy makers with a more comprehensive understanding of the ramifications of drug abuse. In addition, such data would provide synergy for greater multilateral and multisectoral cooperation in prevention.

With leadership from international drug agencies such as the United Nations International Drug Control Programme (UNDCP), a process has already begun that will enhance the development and use of economic data and, ultimately, improve the quality of decisions made on drug issues at the national, regional and international levels.

References

1. United Nations International Drug Control Programme, *World Drug Report* (London, Oxford University Press, 1997).

2. P. Giffin, S. Endicott and S. Lambert, *Panic and Indifference: the Politics of Canada's Drug Laws* (Ottawa, Canadian Centre on Substance Abuse, 1991).

3. A. Markandya and D. Perce, “The social costs of tobacco smoking”, *British Journal of Addiction*, vol. 84, 1989.

4. L. Robson and E. Single, *Literature Review of Studies on the Economic Costs of Substance Abuse* (Ottawa, Canadian Centre on Substance Abuse, 1995).

5. D. Collins and H. Lapsley, *Estimating the Economic Costs of Drug Abuse in Australia*, National Campaign against Drug Abuse Monograph Series, No. 15 (Canberra, Australian Government Publishing Service, 1991).

6. D. Collins and H. Lapsley, *The Social Costs of Drug Abuse in Australia in 1988 and 1992*, National Drug Strategy Monograph Series, No. 30 (Canberra, Australian Government Publishing Service, 1996).

7. M. Adrian, P. Jull and R.Williams, eds., *Statistics on Alcohol and Drug Use in Canada and Other Countries* (Toronto, Addiction Research Foundation, 1989).

8. E. Single and others, “The economic costs of alcohol, tobacco and illicit drugs in Canada, 1992”, *Addiction*, vol. 93, 1998, pp. 983-998.

9. Institut suisse de prophylaxie de l’alcoolisme, *Le Problème de la drogue—en particulier en Suisse—considéré sous son aspect social et préventif* (Lausanne, Office fédéral de la santé publique, 1990).

10. C. Fazey and R. Stevenson, *The Social and Economic Costs of Drug Abuse in the U.K. and the Netherlands* (London, Commission of the European Communities, 1990).

11. D. Rice and others, *The Economic Cost of Alcohol and Drug Abuse and Mental Illness 1985*, report submitted to the Office of Financing and Coverage Policy of the Alcohol, Drug Abuse and Mental Health Administration (San Francisco, Institute for Health and Aging, University of California, 1990).

12. W. Manning and others, *The Costs of Poor Health Habits*, RAND Study (Cambridge, Harvard University, 1991).

13. H. Harwood, D. Fountain and G. Livermore, *The Economic Costs of Alcohol and Drug Abuse in the United States 1992* (Rockville, Maryland, United States Department of Health and Human Services, 1998).

14. D. Rice, *Estimating the Cost of Illness*, Health Economics Series, No. 6 (Rockville, Maryland, Department of Health, Education and Welfare, 1966).

15. E. Single and others, *International Guidelines on Estimating the Costs of Substance Abuse* (Ottawa, Canadian Centre on Substance Abuse, 1995).

16. B. Choi, L. Robson and E. Single, “Estimating the economic costs of the abuse of tobacco, alcohol and illicit drugs: a review of methodologies and Canadian data sources”, *Chronic Diseases in Canada*, vol. 18, 1997, pp. 149-165.

17. M. French, J. Rachal and R. Hubbard, “Conceptual framework for estimating the social cost of drug abuse”, *Journal of Health and Social Policy*, vol. 23, 1991, pp. 1-22.

18. D. English and others, *The Quantification of Drug Caused Morbidity and Mortality in Australia, 1992* (Canberra, Commonwealth Department of Human Services and Health, 1995).

19. K. Fox and others, “Estimating the costs of substance abuse to the Medicaid Hospital Care Program”, *American Journal of Public Health*, vol. 85, 1995, pp. 48-54.

20. B. Armstrong, “Morbidity and mortality in Australia: how much is preventable?”, J. McNeill and others, eds., *A Handbook of Preventive Medicine* (Sydney, Edward Arnold, 1990).

Economic evaluation of policies and programmes: further uses of estimates of the social costs of substance abuse

D. COLLINS

Adjunct Professor, Department of Economics, Macquarie University, Sydney, Australia

H. LAPSLEY

Senior Lecturer in Health Economics, Faculty of Medicine, University of New South Wales, Sydney, Australia

ABSTRACT

The present article identifies the theoretical areas in substance abuse estimation that have not been sufficiently addressed. Those include issues relating to the definition of social costs, a more comprehensive labour market analysis, the social benefits of drug consumption and the distributional impacts of substance abuse. Examples are presented of types of cost estimates, how the results of estimates can be interpreted and the policy use of each type of cost.

Data requirements are identified and the process of proceeding from aggregate estimates to disaggregated evaluation is reviewed. Issues of attribution are considered, and the importance of calculation of avoidable costs of substance abuse is explained.

General issues are reviewed with regard to benefit-cost analysis and evaluation criteria applicable to substance abuse. The article presents as a case study the economic evaluation of Quit Victoria. It uses the calculations of the social costs of tobacco to provide the basis of benefit-cost evaluation of Quit Victoria. The study resulted in a positive benefit-cost ratio under all assumptions.

The article concludes with a review of the issues to be addressed in the economic evaluation of a medically supervised injecting room that is being undertaken in New South Wales, Australia. It emphasizes the importance of estimating social costs in project appraisal and public policy-making.

Introduction

The objectives of the Third International Symposium on the Economic and Social Costs of Substance Abuse, held in Banff, Canada, from 31 May to 3 June 2000, were set out in the August 1999 planning concept paper of the Canadian Centre on Substance Abuse. The short-term goal was to test and improve the international guidelines formulated by the Centre [1]; the medium-term goals were to improve the database for cost estimation, to fill existing gaps in methodology and to extend the scope of such studies to include estimates of avoidable costs; and the long-term goal was to expand cost estimation to comprehensive cost-benefit analyses of alternative policy options.

The present paper deals specifically with the long term-goal of the Symposium, the economic evaluation of policy options. Before discussing issues of economic evaluation, however, it should be pointed out that much theoretical and data development remains to be undertaken before estimation of the costs of substance abuse can be considered to have reached a satisfactory level.

Theoretical areas in need of further development include:

(a) *Definition of social (external) costs of substance abuse.* Consensus has not yet been reached on the definition of social costs, even though it is of crucial importance to the estimation of such costs. For example, the question of addiction and its impact on informed and rational decision-making has yet to be resolved satisfactorily in spite of the popularity among economists of the theory of rational addiction [2]. Another important issue is whether costs imposed by substance abusers on members of their immediate families should be considered to be private or social costs;

(b) *Labour market analysis.* The issue is whether labour markets fully adjust factor returns so that the costs of substance abuse are internalized upon the abuser and can therefore be treated in some circumstances as private costs. If labour markets do not perfectly adjust in this way (which is more plausible), the question is what proportion of drug-related productivity losses can be considered to be social costs;

(c) *Social benefits of drug consumption.* Some types of substance use can lead to private and/or social benefits (as is the case with moderate alcohol consumption, which is neither addictive nor abusive). The question is whether such benefits should be set against the costs of substance abuse or whether they should be simply ignored, as they are in most studies;

(d) *Distributional impact of substance abuse.* The Second International Symposium on the Economic and Social Costs of Substance Abuse, held in Montebello, Canada, from 2 to 5 October 1995, recommended that future studies should indicate the incidence of the social costs of substance abuse among broad community groups (substance abusers, other individuals, business and government). Such a process involves difficult methodological issues akin to those involved in tax incidence analysis. However, from a political point of view, the incidence of the costs of substance abuse is particularly important since such a high proportion of those costs is likely to be borne by individuals and businesses, rather than by government. That area of analysis is in need of substantial further development. There may well be scope for extending it to include an analysis of the incidence of such abuse costs in different household income categories, that is, the distributional consequences of substance abuse.

Areas of primary importance for future data development are:

(a) Development of attribution factors for drug-related crime;

(b) Impact of substance abuse upon productivity (both absenteeism and on-the-job productivity);

(c) Further development of epidemiological information on passive smoking;

(d) Measurement of the costs of the abuse of pharmaceuticals;

(e) Analysis of the impact of tobacco smuggling;

(f) Valuation of intangible costs such as pain, other forms of suffering and loss of life;

(g) Estimation of avoidable costs;

(h) Costs of fires attributable to cigarette smoking;

(i) Costs of drug-related litter and other forms of pollution.

There is no dispute with the primary objective of the Third Symposium, which was to refine the development and use of the social costs of substance abuse. However, the discussion to date has largely ignored the further uses to which the wealth of data generated in the process of producing estimates of social costs can be profitably applied, in particular in the area of project appraisal. Brown and Jackson [3] used "project appraisal" as a general term referring to attempts by applied welfare economists to evaluate the efficiency of alternative projects or more widely the efficiency of alternative policies. That is the sense in which the term is used in the present article.

Social cost estimates and their uses

Estimation of social costs performs the important function of indicating the size of the substance abuse problem and its component parts. Extending the analysis to estimation of avoidable costs gives valuable information about the potential economic returns available to policies and programmes designed to curb substance abuse. The aggregate estimates have been used in various countries as a powerful political argument for mobilizing public resources to be used against drug abuse. Such aggregate estimates do not, however, indicate to which particular programmes and policies those resources should be directed.

Table 1 presents a summary of various types of exercises involved in the estimation of the social costs of substance abuse, together with the interpretation and significance of each type of exercise.

Table 1. Cost estimates of substance abuse

Type of estimate	*Interpretation of results*	*Example of policy use*
Aggregate costs	Total external costs of substance abuse compared with a situation of no substance abuse	Indication of the size of the substance abuse problem
Avoidable costs	Potential economic benefits from substance abuse harm minimization strategies	Determination of the appropriate level of resources to be devoted to harm minimization strategies
Incidence of costs	Distribution of the external costs of substance abuse among various community groupings	Mobilization of support from various groups (such as the business community) for programmes against substance abuse
Disaggregated costs	External costs of substance abuse disaggregated by category	Economic evaluation (cost-benefit or cost-effectiveness analysis) of harm minimization programmes
Budgetary impact	Impact of substance abuse on government expenditure and revenue	Assessment of the case for industry making payments to government as compensation for abuse of substances produced by the industry

Source: Adapted from J. D. Collins and H. M. Lapsley, "Estimating and disaggregating the social costs of tobacco", *The Economics of Tobacco Control: Towards an Optimal Mix,* I. Abedian and others, eds. (Cape Town, Applied Fiscal Research Centre, University of Cape Town, 1998).

Economic evaluation (the topic of the present paper) and, in particular, cost-benefit analysis and cost-effectiveness analysis applied to substance abuse issues are logical developments of the types of analysis referred to in table 1.

Studies estimating the social costs of substance abuse are based on the collection and development of data on all aspects of the economic impact of substance abuse. As an indication, a typical study concerning the aggregate cost of substance abuse would involve the compilation and/or use of data on some or all of the following:

Epidemiological information on substance abuse;
Mortality classified by disease;
Morbidity classified by disease;
Drug use surveys;
Use and abuse of pharmaceuticals and associated costs;
Public and private hospital occupancy;
Public hospital costs of diagnosis-related groups;
Occupancy and costs of public and private nursing homes;
Publicly funded medical services;
Health expenditure and sources of funding;
Health insurance;
Consumption expenditure;
Labour force;
Labour force absenteeism;
Award rates of pay;
Employee earnings and hours of work;
Unpaid work;
Household incomes and wages, salaries and supplements;
Company incomes and turnover;
Valuation of pain, suffering and loss of life;
Output of manufacturing industry;
Tobacco and alcohol market shares;
Population size and structure;
Demographic impact of substance abuse;
Income tax and indirect tax revenue;
Customs duties;
Subsidies and assistance to relevant industries;
Motor vehicle usage and accidents;
Drug-related crime and law enforcement;
Property insurance;
Consumer prices;
Substance abuse programme costs.

Studies of the aggregate cost of substance abuse provide a high proportion of the data needed for programme evaluation. If the physical output of programmes (for example, reduction in smoking prevalence, liver cirrhosis or drug psychoses) can be identified, aggregate cost studies will normally yield virtually all the data needed for the economic evaluation of the benefits of individual projects.

Such an extension of the use of aggregate data rarely occurs and those project appraisals attempted tend to be undertaken independently of the aggregate cost studies. (A surprising number of evaluations are undertaken by teams that appear to possess little in the way of economic expertise.) A gold mine of economic data is available for project appraisal but appears to be rarely used for such purposes.

A logical development of the *Guidelines* of the Canadian Centre on Substance Abuse [1] would be an extension into the area of project appraisal. In the present article, some of the theoretical and practical problems of such evaluation are considered, using examples from Australia with which the authors have been, or are currently, involved.

From aggregate estimates to disaggregated evaluation

Aggregate estimates of the social costs of substance abuse give no indication of the benefits potentially available to harm reduction programmes since:

(a) Some of those costs relate to past substance abuse (for example, morbidity attributable to smoking) and are therefore unavoidable costs;

(b) There can be no expectation that it will be possible to completely prevent the abuse of any particular substance. Even for periods well into the future, when the effects of past abuse have “washed out” of the system, it may be possible to reduce the costs of substance abuse but not to eliminate them.

It is necessary therefore to estimate the avoidable costs of substance abuse in order to be able to indicate the extent of potential benefits to harm minimization programmes. Estimates of avoidable costs, however, fail to indicate how those cost reductions might be achieved or whether the social benefits resulting from such programmes would exceed their social costs. Those issues can be settled only by a process of project appraisal.

Project appraisal evaluates the efficiency of alternative projects or alternative policies. Its aims are to determine, by enumerating the benefits and costs of alternative projects or policies, the appropriate level of public resources to be devoted to the problem, and the nature of the solutions to the problem. Its objective is to maximize the social rate of return resulting from the use of public resources so that the resources can be used as efficiently as possible.

Project appraisal is approached from the perspective of the community as a whole, not from the perspective of individuals, firms or the public sector. Such a social perspective complicates the analysis substantially, since private project appraisal avoids many of the difficult theoretical and practical issues that social appraisers must deal with, in particular the valuation of benefits and choice of discount rate. Furthermore, since the viewpoint is that of the community as a whole and not just that of the government, the issues are much more complex than those of public expenditure funding and public revenue benefits.

In principle, the process of project appraisal should lead to the allocation of resources to all public programmes that yield at least a minimum rate of return. That rate of return should take account of the rates of return, calculated on a consistent basis, on investment in the private sector, to ensure an efficient allocation of resources between the private sector and the public sector. In practice, there are many types of public goods and services that the private sector would never supply (unless through private provision facilitated by public funding). Even if it were possible to calculate public and private rates of return on such a consistent basis, political constraints on public expenditure levels mean that the objective of project appraisal is usually to achieve the efficient allocation of previously determined public expenditure levels between competing public sector uses. Governments generally purport to be attempting to reduce the size of the public sector on the grounds that private expenditure is more efficient than public expenditure. The empirical basis for such a generalization is often extremely questionable and it usually appears to be based more on ideology than on analysis. Combined with the equally questionable doctrine that personal and company income tax rates should be reduced as much as possible, it has had a powerful effect on the imposition of constraints on public expenditure. At that level of evaluation, the appropriate evaluation tool is benefit-cost analysis (BCA).

The scope of project appraisal can be extended into programme budgeting, which is a system of managing government expenditure by attempting to compare the programme proposals of all government agencies authorized to achieve similar objectives [4]. It appears that, in the area of health, such lofty objectives are rarely sought, let alone achieved. The paucity of individual project appraisals makes the implementation of programme budgeting seem a distant objective.

In many cases, the objectives of public expenditure analysis may be even more modest. The objective of the evaluation exercise may be predetermined (for example, a reduction of 10 per cent in juvenile smoking prevalence) so that the analysis is reduced to cost comparisons of alternative programmes designed to achieve the same objective. In other situations, it may be considered that it is so difficult to value a programme’s output that BCA is impossible. In those circumstances, a cost-effectiveness analysis (CEA) is more appropriate [5].

CEA can be defined as a detailed comparison of the costs of alternative techniques for achieving the same predetermined objective. In practice, CEA can be used to determine how a given objective can be achieved at the least cost, or how a desired output can be maximized for a given cost. The objectives and outputs of programmes subject to CEA are almost always one-dimensional since, if alternative programmes yield multiple outputs in different ratios, it becomes necessary to assign values to each type of output. That then moves the analysis into the realm of BCA. Murray and others [5] discuss a broader view of cost-effectiveness in terms of allocating a fixed budget between interventions in such a way as to maximize health in a society. They acknowledge that only a few applications of this broader use—in which a wide range of preventive, curative and rehabilitative interventions that benefit different groups within a population are compared in order to derive implications for the optimal mix of interventions—can be found.

The advantage of CEA in its usual and more limited sense is that there is no need to value output benefits. That makes the analysis much simpler than BCA, since it is necessary to identify only the costs of alternative interventions. That is generally a much more straightforward process than the valuation of programme benefits, even though significant problems may arise in the allocation of overhead costs.

The major disadvantage of CEA is that the policy objective is predetermined rather than arising from the analysis. CEA is of no assistance in determining policy objectives. As Murray and others [5] pointed out, the implicit assumption of CEA that the required additional resources would need to be transferred from another health intervention or from another sector is rarely discussed.

A further extension of evaluation techniques comes in the form of cost-utility analysis (CUA). While CUA is the least common method of economic evaluation identified in the present article, its use within the health-care sector warrants some discussion. Such a method of analysis calculates the cost per specified health effect (of a programme, a technology or a pharmaceutical intervention) and expresses outcomes as uniform units of health. Those units are presumed to have similar values across all conditions. The health effects are weighted to reflect individual or societal preferences for different health outcomes.

The most common weighting units are quality-adjusted life years (QALYs) and disability-adjusted life years (DALYs). The QALY measurement attempts to compare treatment priorities, identifying and measuring the utility of using resources to treat people of different health status, with different likely outcomes of treatment. QALY is regarded as a unit of health that combines extension of life with a measure of its worth. Its use is focused on societal decisions relating to which good or service to produce relative to one another, that is, allocative efficiency. There have been a number of "league tables" developed comparing QALY measures of quality and quantity of life years gained. It is possible, for example, to compare QALYs from resources spent on smoking cessation programmes with resources spent on organ transplantation.

The DALY measurement combines healthy life years lost because of premature mortality with healthy life years lost because of disability. It is a useful economic tool as the resource implications of each DALY component can be identified and estimated. The total loss of DALYs worldwide reflects the global burden of disease.

CUA can be considered a special form of CEA in which effects are measured in health status; it can contribute to societal decision-making by determining allocative efficiency. While it is still a matter of contention whether QALYs and DALYs can be effective and acceptable public policy tools, they are increasingly being used to contribute to economic evaluation in the health-care sector.

General issues in benefit-cost analysis

Social benefit-cost analysis attempts to describe and quantify the social benefits and costs of a policy or programme expressed in terms of a common monetary unit. The current value of the flow of social benefits over the life of a project is compared with the current value of the flow of expenditures that have yielded those benefits. Discounting techniques are used to permit comparison of the current value of differential flows of the benefits and expenditures over time. The costs and benefits, once valued in comparable terms, are then compared in terms of a criterion such as a benefit-cost ratio or some measure of the project's rate of return.

The theoretical and practical issues involved in BCA have been discussed at length in the literature and it would be inappropriate to undertake a full review here; however, a brief recapitulation of the major issues is presented below as they form the basis for the discussion by the authors of an example of BCA in the field of drug control.

Valuation of benefits and costs

As indicated earlier, exercises focusing on aggregate costs of substance abuse should identify and place values on all the costs of abuse of the substance under review. Any reduction of those costs resulting from implementation of a particular programme will represent benefits of that programme. The theoretical and practical issues involved in the valuation of programme benefits should already have been addressed in the study on the costs of substance abuse. An approach based on human capital should provide all the necessary information, including discounted future costs. The approach based on demographics will not yield information directly about future substance abuse costs; thus, extra analysis will be necessary.

Estimates of the aggregate costs of substance abuse should already have made the necessary distinctions between private and social costs and should have ensured that double-counting of costs was avoided by including only real (not pecuniary) costs. The estimates should already have taken account of valuation problems, including the impact of private market imperfections such as monopoly power or managed exchange rate regimes, and the difficulties of placing valuations on intangibles such as pain, other forms of suffering and loss of life. (For a review of issues involved in estimating the costs of tobacco use, see J. Lightwood and others [6]).

It will still be necessary to undertake the extra research involved in identifying programme costs, but that should prove to be a relatively straightforward exercise.

Discounting of benefits and costs

The choice of discount rate, by which the time streams of benefits and costs are to be reduced to values expressed in monetary units in a common year, is one of the major theoretical problems of BCA. It is important for long-lived projects, for which the benefit-cost ratio can be particularly sensitive to the choice of discount rate. For example, it has been concluded that the issue of whether the lifetime health-care costs of smokers are higher than those of non-smokers depends upon the discount rate that is applied [7].

For a typical public investment project, the higher the discount rate is, the lower the benefit-cost ratio will be. As discussed below, serious theoretical problems remain in the choice of discount rate. That may not be as great a problem, however, for drug programme evaluations as it can be for other forms of public investment analysis.

The most usual application of social BCA is in the evaluation of large public investment projects, such as roads, bridges and power stations, which typically involve a large initial investment, operating costs over the predetermined life of the project and then, possibly, significant expenditure at the end of the project life if, for example, the facility has to be scrapped. The highly uneven time stream of costs is to be compared with a relatively steady flow of benefits. Expenditure patterns over time on drug programmes (such as the Victorian Smoking and Health Program in Australia, discussed below) are quite different, however; there, relatively steady expenditure flows from the beginning lead, after a period of time, to a relatively steady flow of benefits. If the benefit and cost flows of the projects to be compared are relatively steady over time, the ranking of the projects may be relatively insensitive to the choice of discount rate.

Serious issues about the choice of discount rate remain unresolved. Essentially, the choice is between the social opportunity cost rate and the social time preference rate.

The social opportunity cost rate is a measure of the best alternative use to which the community might have put the resources used in the project under review. That reflects the rates of return available to alternative investments, in the private sector or the public sector. The social time preference rate is a measure of the community's valuation of current consumption against future consumption. A society that places the interests of the current generation above those of future generations will have a high social time preference rate. The two rates are, in an imperfectly competitive real world, most unlikely to be the same. One would expect the social time preference rate to be lower than the social opportunity cost rate. The issue of which category of discount rate to choose has not been resolved [3].

Evaluation criteria

The results of BCA are presented using one or more of the following three measures:

(a) *Net present value* is the present value of the future time stream of net benefits *(*that is, benefits less costs*)*. The future benefits and costs are discounted to present values by use of what is deemed to be an appropriate discount rate. All other things being equal, the higher the discount rate that is chosen, the lower the net present value will be;

(b) *Benefit-cost ratio* is the ratio of the present value of a programme's benefits to the present value of its costs. Once again, because the process involves discounting the size of the calculated measure, it is crucially dependent on the choice of discount rate;

(c) *Internal rate of return* is the discount rate that equates the net present value to zero. That measure has the advantage of not being dependent upon a choice of discount rate—the calculation generates the internal rate of return directly. The logic of such a measure is that it effectively represents the rate of return generated by the programme or project.

As a result of different sets of implicit assumptions underlying the calculation of the criteria, the ranking of the projects under review may not be consistent for all three criteria.

Distributional issues

In public policy deliberations, most countries take account of distributional considerations, expressed broadly as the extent of poverty and wealth in the community. Different substance abuse programmes may have different distributional impacts. It is possible (although there is no supporting empirical evidence) that the damaging effects of smoking on health are more highly concentrated on the poor than are the effects on health of alcohol abuse. Thus, anti-smoking programmes might have what the community considers to be more desirable distributional consequences than anti-alcohol programmes. BCA can assign distributional weights designed to reflect what are perceived to be society's distributional values, perhaps giving more weight to benefits accruing to the poor than to those accruing to the more affluent.

Economic evaluation of Quit Victoria

The authors have recently attempted a benefit-cost analysis of an ongoing programme in Australia called the Victorian Smoking and Health Program, also known as Quit Victoria [8]. The benefit-cost analysis is based on data generated by an aggregate cost study presented in the same paper that also attempts, for the first time in Australia, to estimate the social costs of tobacco use for the state of Victoria. This section of the present article draws heavily upon the methodological discussion in the Quit Victoria report and presents a summary of the results of the analysis. Many of the problems confronted in the Quit Victoria BCA are common to evaluations of programmes in the area of substance abuse.

Quit Victoria was established in 1985 with the objective of reducing the harmful effects of tobacco use in Victoria. Its main source of funding is the Victorian Health Promotion Foundation, which was originally funded through the state tax on tobacco. As a result of the fact that state tobacco taxes have recently been declared to be unconstitutional, the Foundation is now funded through direct grants from the Health Department of Victoria. Other funders of Quit Victoria are the Anti-Cancer Council of Victoria and the National Heart Foundation (Victorian Division). Quit Victoria conducts media campaigns, school and community education and anti-smoking sponsorships and supports legislation designed to reduce tobacco consumption.

Any estimate of the benefits of a public health programme such as Quit Victoria relies upon identification of an implicit mechanism of the link between implementation of the programme and the harm reduction outcome. The mechanism is, in principle, as follows:

Programme→reduction in consumption→reduction in harm

There are two fundamental links in the above-mentioned mechanism:

(a) *Programme leads to a reduction in consumption.* This raises two questions:

(i) What reduction in consumption is attributable to the programme?

(ii) Over what period of time does that reduction occur?

(b) *Reduction in consumption leads to reduction in harm.* Three questions arise:

(i) What reduction in harm is attributable to the reduction in consumption?

(ii) Over what period of time does the reduction in harm occur?

(iii) Over what period of time does the reduction in costs (whose estimation is the objective of the exercise) occur?

Any such exercise must therefore address the dual issues of quantifying the health effects of anti-smoking programmes and determining the periods of time over which those programmes take effect.

A further significant question is whether programmes such as Quit Victoria produce permanent reductions in smoking prevalence or whether smoking prevalence would return to former levels if such programmes were to be abandoned. Is the effect of programmes such as Quit Victoria like a screw, remaining in its new position when the driving force is removed, or is it like a spring, reverting to its former position? The answer to that question will determine whether there is a need for ongoing anti-smoking programmes or whether such programmes can cease once the desired reduction in smoking prevalence is achieved. The answer is also crucial to the process of estimating the rates of return on anti-smoking programmes.

Expenditures or programmes such as Quit Victoria may lead to positive “externalities”, such as reductions in the prevalence of smoking in other states. In estimating the impact of such a programme, account should be taken of such externalities. Other states may benefit from the money spent on the programme by the state of Victoria (for example, when Quit Victoria messages on billboards at the Melbourne

Cricket Ground are seen on national cricket telecasts) and Victoria may benefit from expenditure by other states and by the Government of Australia. No attempt was made to estimate those externalities in the Quit Victoria study. In the case of Quit Victoria, all such externalities are positive, that is, the programmes confer benefits (rather than impose costs) on other states. The same is probably true for virtually all harm minimization programmes whose effects spread beyond the jurisdictions funding those programmes.

A related issue is whether the impact of Quit Victoria programmes, in reducing smoking prevalence in one state, could be evaluated using the changes in smoking rates in other states as a comparator. If their prevalence was also reduced, a comparison of Victoria with New South Wales, for example, might well underestimate the effectiveness and positive outcome of expenditure of the Quit Victoria programme.

The calculations of the social costs of tobacco in Victoria in 1988 and 1992 [8] provide the basis for BCA evaluation of Quit Victoria.

Social BCA involves comparison over the life of a programme of its social costs and social benefits. The Quit Victoria study presents estimates of social costs in Victoria for the years 1988 and 1992, on the basis of which an attempt to predict future social costs can be made. The study makes what is deemed to be a conservative assumption that the social costs of tobacco smoking rose linearly between 1988 and 1992 and that, in the absence of anti-smoking programmes, they would remain at 1992 levels, expressed in real terms, in the future.

The problem arose concerning estimation of the reduction in the social costs of tobacco use in Victoria that could be attributed to Quit Victoria expenditure. Generally, the best measure of smoking in any community is taken to be the smoking prevalence rate, that is, the percentage of the relevant population who smoke. Over the life of Quit Victoria, the total smoking prevalence rate in Victoria has declined substantially, mirroring the decline in the national rate. How much of the decline in Victoria is attributable to Quit Victoria?

The smoking prevalence rate in Victoria is subject to a wide range of influences. The major influences can be summarized as follows:

(a) Quit Victoria;

(b) Other anti-smoking policies in Victoria, for example, restrictions on cigarette advertising;

(c) Anti-smoking campaigns of the federal Government and of other state governments;

(d) Federal and state tobacco tax policies;

(e) Other federal and state regulatory policies, for example, sponsorship bans and restrictions on smoking in the workplace;

(f) Broader economic and non-economic factors, such as changes in living standards or in attitudes towards healthy living.

Can such varied influences on the Victorian prevalence rate be quantified separately? Little evaluation work has been undertaken of anti-smoking programmes and policies in Australia. Perhaps the most intensive study in that area is the recently-published evaluation of the National Tobacco Campaign [9]. The results published to date, however, provide no useful quantitative estimates of the Campaign’s impact on the national prevalence rate, and little other verifiable information on the topic is available. There is an urgent need for research disaggregating the effects of the various factors influencing prevalence rates; however, it is currently not possible to identify those effects.

The approach adopted in the study was to identify the decline in the prevalence rate in Victoria over the relevant period and to consider the proportion of the decline that could be attributed to the activities of Quit Victoria. The range of attributable proportions adopted in the study was 10-30 per cent, that is, it was assumed that 10-30 per cent of the decline in the prevalence rates in Victoria was attributable to the activities of Quit Victoria. The actual attributable decline is most likely to lie within that range. Chaloupka and Warner [10] provide international information on the impact of anti-smoking programmes.

Declines in prevalence will lead to reductions in social costs but the lags involved in that process are difficult to identify. A decline in smoking prevalence may lead to a virtually instant decline in some costs, such as those arising from fire-related deaths, injuries and damage. On the other hand, other types of costs may only be responsive to declines in smoking prevalence with a considerable lag. For example, reduced smoking prevalence may lead to a decline in costs related to lung cancer only after a period of many years. It appears impossible, on the basis of currently available research, to estimate the relevant average lag period. For evidence on individual lags, see Armstrong [11]. The Quit Victoria study adopted the approach of making an educated guess of the range of values in which the actual lag might lie and to test the sensitivity of the results to the adoption of different lags. All other things being equal, the longer the assumed lag, the lower the calculated rate of return will be, since the total benefits will be lower and they will accrue later in the life of the programme.

The present study, therefore, assumes that a given proportionate reduction in the prevalence rate will lead to the same proportionate reduction in social costs after a relevant period. Results are calculated on assumed average lags of 6, 8 and 10 years.

There is no problem identifying the annual costs incurred to date in financing Quit Victoria. There is, however, a problem in forecasting expenditure, compounded by the fact that the future prevalence rate will presumably be a function of future expenditure. There are three possible scenarios to be considered:

(a) The “spring” scenario: maintenance of Quit Victoria expenditure at the current real level will keep the smoking prevalence rate at the current levels;

(b) The “screw” scenario: reduction of the Quit Victoria expenditure level to zero will not lead to any change in the current prevalence rate;

(c) The “modified spring” scenario: maintenance of the current expenditure level of Quit Victoria will lead to further declines in prevalence.

There are other possible scenarios but they are likely to be variants of the above-mentioned scenarios.

The scenario assumed to be the least favourable to Quit Victoria is the “spring” scenario. Compared with that scenario, the “screw” scenario would imply lower expenditure for the same benefits while the “modified spring” scenario would imply higher benefits for the same expenditure. The most conservative approach, the “spring” scenario, was used for the present study. Thus, it was assumed that the expenditure level of Quit Victoria would be maintained at the real 1998 level and that the prevalence rate would remain constant at the 1998 level.

From that, an important question arises: over what period should such an analysis for Quit Victoria be undertaken? In investment analysis, it is usually possible to determine the life of the project. In the case of Quit Victoria, the programme has no predetermined life and it was difficult to predict future public policies towards smoking in Victoria. Given that the benefits of Quit Victoria programmes may accrue over many years and that they would certainly continue to accrue for at least a period of time after one programme had ended, the adoption of too short a period of analysis would lead to an underestimation of the programme benefits. On the other hand, it is extremely difficult to predict future developments in medical technology and, thus, future medical costs. Technological improvements may lead to cost reductions, for example, as a result of the development of new vaccines. Alternatively, they may lead to cost increases, for example, as a result of the development of more effective but more expensive medical treatments. The higher the adopted discount rate, the lower the value placed on future benefits will be. For example, one dollar's benefits to be received in 30 years will have a current value of 17 cents at a discount rate of 6 per cent, but only 6 cents at a discount rate of 10 per cent. For the purposes of the Quit Victoria study, a period of analysis of 30 years was adopted.

The absence of adequate research results with regard to the links between programmes like Quit Victoria, smoking prevalence and the social costs of smoking means that a study such as the present one must identify its underlying assumptions and test the sensitivity of the results to changes in those assumptions. In table 2, below, details are presented of the three sets of assumptions underlying the evaluation of Quit Victoria in the present study.

Table 2 shows, for each of the issues identified in the discussion above, the most conservative assumption (in terms of reducing the indicated rate of return to the Quit Victoria programme), the least conservative assumption and the most plausible assumption (in terms of its approximation to reality). Thus, in that context, the term conservative denotes "yielding a comparatively low rate of return".

Table 2. Assumptions underlying the Quit Victoria evaluation

Set of assumptions	*Attributable decline in smoking prevalence (percentage of total decline)*	*Time lag between reduction in prevalence and reduction in social costs (years)*	*Characterization of the effect of Quit Victoria expenditure*[a]	*Discount rate (percentage)*
Most conservative	10	10	Acts as a spring	8
Least conservative	30	6	Acts as a screw	4
Most plausible	20	8	Acts as a spring	6

[a] If the effect of the expenditure on Quit Victoria is assumed to act as a "spring", it is necessary to maintain the Quit Victoria expenditure at the current level in order to maintain the current rate of smoking prevalence. If the expenditure is assumed to act as a "screw", the prevalence rate will remain at its current level even though all Quit Victoria expenditure will cease.

In table 3, evaluation results are presented on the basis of the three sets of assumptions. The following four evaluation measures are calculated for each set of assumptions:

(a) Net present value in 1987 prices: the value of the time stream of Quit Victoria social benefits less the time stream of its costs with both benefits and costs discounted back to the values in 1987, the year in which Quit Victoria was instituted;

(b) Net present value in 1999 prices: net present value in 1987 prices converted to 1999 prices by the application of the change in the relevant consumer price index between 1987 and 1999;

(c) Benefit-cost ratio: the ratio of the present value of Quit Victoria benefits to the present value of the costs;

(d) Internal rate of return: the discount rate that equates net present value to zero or the benefit-cost ratio to unity; in effect, the social rate of return generated by the programme.

Table 3. Social benefits of Quit Victoria

	Net present value (millions of Australian dollars)			
Set of assumptions	*1987 prices*	*1999 prices*	*Benefit-cost ratio*	*Internal rate of return (percentage)*
Most conservative	156	224	5.4	24.1
Least conservative	1 675	2 416	51.1	55.9
Most plausible	632	911	15.8	37.9

Two major conclusions can be drawn from table 3.

(a) The evaluation results are extremely sensitive to the choice of assumption sets;

(b) Even on the most conservative set of assumptions, expenditure on Quit Victoria yields extremely high rates of return.

The estimated social benefits of Quit Victoria are high even under the most conservative analysis. The Quit Victoria programme would yield negative net benefits only if the impact of the programme accounts for no more than 1.5 per cent of the overall reduction in smoking prevalence over the relevant period. It appears implausible that the Quit Victoria programme could have such a minimal impact.

The conclusion to be drawn is that it is extremely difficult to conceive of any other public expenditure, outside the area of public health, that would yield social rates of return of the same order of magnitude as those of Quit Victoria.

There seems no doubt that expenditure on Quit Victoria to date has yielded, and will continue to yield, very high social benefits to the residents of Victoria. Presented below are some of the conclusions that can be drawn from the analysis about appropriate future policies towards the programme.

Given the steady decline in smoking prevalence rates in Australia over the last decade, the question that arises is whether there is any reason to believe that it is possible to reduce smoking prevalence further. It would be necessary to determine the minimum smoking, or baseline, rate that could be achieved in a fully informed community at Australia's stage of economic development.

Perhaps the best indication of the baseline rate of Australia is smoking behaviour among general medical practitioners in Australia [12]. Given that zero smoking prevalence is not achievable, the smoking rates of general practitioners could be interpreted as the best achievable in a fully informed population. About 9 per cent of male general practitioners and about 4 per cent of female general practitioners are smokers and only about 4 per cent of doctors under the age of 30 smoke. Surveys of physicians in the United States of America have shown similar results. About 2-3 per cent of medical students in Australia smoke. That may be considered to be the minimum baseline rate, as medical students (because of their age) have all been exposed to tobacco education prior to their decision to smoke. Those figures compare with an actual prevalence rate in Victoria of 25.3 per cent in 1997.

There is substantial scope for continued improvement in the overall prevalence rate, especially if the Australian community were to become fully informed about all relevant aspects of smoking, including not only the health impact of smoking, but also the highly addictive properties of nicotine.

There seems to be no obvious reason to assume that Australia has, in 2000, reached the minimum achievable prevalence rate. With regard to such rates, the "most plausible" evaluation adopted in the present study may well be too conservative. The case for continued Quit Victoria expenditure is extremely strong. Again, if the effect of Quit Victoria expenditure is like a spring, the rate of return to future expenditure would be high, even though it represents a holding operation—preventing the prevalence rate from rising again. Thus, there seems to be every justification for continuing the Quit Victoria programmes in the future.

Given the high calculated rate of return on Quit Victoria expenditure, it would appear that the current level of expenditure on Quit Victoria is far too low. If public resources are being directed to programmes that yield lower marginal rates of return than those attributable to Quit Victoria, then it is clear that resources are being misallocated.

What are the issues to be determined in such a BCA exercise? The following represents the major issues that arose in the Quit Victoria study:

(a) What the aggregate social costs of abuse of the substance under review are;

(b) What programme expenditures, both past and future, amount to;

(c) To what extent the reduction in prevalence could be attributed to the programme;

(d) Whether the effects of anti-abuse programmes are permanent or temporary (can be likened to a screw or to a spring);

(e) Whether the programme confers externalities on other jurisdictions;

(f) What the optimum size of the programme is;

(g) What future medical costs will be;

(h) How long a period the BCA should cover;

(i) What the time lags between reduced prevalence and reduced costs are;

(j) What the discount rate should be;

(k) What the evaluation criterion should be.

Issues in the economic evaluation of a medically supervised injecting centre

The state of New South Wales in Australia has recently passed legislation to enable an 18-month trial of a medically supervised injecting centre, to be situated in an inner-city area in Sydney with high usage of illicit drugs. Heroin will not be supplied, so to that extent the role of the centre can be seen as similar to the provision of clean needles and syringes, as a public health strategy to reduce blood-borne virus transmission. The trial is to be accompanied by a substantial evaluation component involving quantifiable measures, including:

(a) Public health (e.g. drug overdose and blood-borne virus transmission);

(b) Health of clients, (e.g. fewer deaths, utilization of treatment services);

(c) Public amenity (e.g. public injecting and disposal of injecting equipment);

(d) Criminal activity;

(e) Economic analysis, including cost analysis.

The evaluation protocol is still being developed. It is anticipated that, when the centre opens in 2000, it will be accompanied by a comprehensive collection of clinical, epidemiological and economic data. The evaluation is an example of economic data contributing to overall appraisal of a drug policy intervention. The economic analysis to be undertaken for the centre can be directly related to the types of estimates listed in table 1. Firstly, aggregate costs have been used to identify and quantify the extent of the problem, contributing to the political decision to undertake the trial. The second type of estimate, avoidable costs, forms an important component of the economic evaluation, as the anticipated reduction in costs relating to ambulance attendances, treatment of blood-borne viruses and overdose deaths are examples of quantifiable avoidable costs.

The incidence of costs is also relevant to the study. For example, the business community bears some of the costs of substance abuse, while people who live in streets where injecting drug users congregate bear certain costs, including reduced services owing to factors such as public nuisance and lower property values, as the character of their suburb changes.

CEA, requiring disaggregated costs, will identify the extent of harm minimization achieved by the operation of the centre. Some externalities can be anticipated to be identified during the evaluation of the trial phase. After the initial trial, CEA may be useful in making informed decisions relating to economic evaluation of alternative strategies to reduce harm associated with illicit drugs.

The final type of estimate in table 1 is budgetary impact. While expenditure on government-provided programmes is relevant to that category, there is no revenue impact, as heroin is illegal and therefore not taxable.

There are a number of economic issues associated with the trial in addition to those listed in table 1. They include economies of scale (for example, the state of New South Wales is trialling only one centre while the state of Victoria plans to trial five centres), substitution of different levels of health professionals that influence costs (New South Wales, for example, has mandated medical supervision, which is not required in Victoria), and whether exposure to, and uptake of, treatment programmes may result in increased costs to public programmes in the short term.

Conclusion

The present article has addressed what is suggested to be the next stage of the development and use of guidelines concerning the costs of substance abuse. It is envisaged that, as those cost estimates become further refined, their use should also extend into economic evaluation, to enable cost comparisons of policy effectiveness over time and to facilitate economic appraisals of programmes.

The completed case study of Quit Victoria, a cost-benefit evaluation, provides one example of the use of disaggregated cost data, which demonstrates high social benefits from the programme. The second example of economic evaluation, the medically supervised injecting centre, which is yet to be undertaken, involves a range of cost data and a range of specified outcome measures. The economic evaluation of the centre will underline the importance of including cost data and economic criteria in the initial study design.

As economic issues relating to substance abuse and project appraisal of harm reduction strategies become more prominent in public policy-making, the next stage of social cost estimation becomes important. If the extensive data inputs and analytical outputs of studies concerning the costs of substance abuse are not used in project appraisal, substantial (and expensive) data resources will continue to be underutilized.

References

1. E. Single and others, *International Guidelines for Estimating the Costs of Substance Abuse* (Ottawa, Canadian Centre on Substance Abuse, 1996).
2. G. Becker and K. Murphy, “A theory of rational addiction”, *Journal of Political Economy*, 1988, p. 675.
3. C. V. Brown and P. M. Jackson, *Public Sector Economics* (Oxford, Blackwell Publishers, 1990).
4. D. N. Hyman, *Public Finance: A Contemporary Application of Theory to Policy* (London, Dryden Press, 1996).
5. C. J. L. Murray and others, “Development of WHO guidelines on generalised cost-effectiveness analysis”, GPE discussion paper No. 4, 1999.
6. J. Lightwood and others, “Estimating the costs of tobacco use”, *Tobacco Control in Developing Countries*, P. Jha and F. J. Chaloupka (Oxford University Press, 2000).
7. J. J. Barendregt, L. Bonneux and P. J. van der Maas, “The health care costs of smoking”, *New England Journal of Medicine*, vol. 337, No. 15 (1997), p. 1052.
8. D. J. Collins and H. M. Lapsley, *The Social Costs of Tobacco in Victoria and the Social Benefits of Quit Victoria* (Melbourne, Australia, Victorian Smoking and Health Program, 1999).
9. Australia’s National Tobacco Campaign, *Evaluation Report, Volume 1* (Commonwealth Department of Health and Aged Care, 1999).
10. F. J. Chaloupka and K. E. Warner, *The Economics of Smoking*, National Bureau of Economic Research Working Paper No. 7047 (Cambridge, Massachusetts, 1999).
11. B. K. Armstrong, “Morbidity and mortality in Australia: how much is preventable?”, *A Handbook of Preventative Medicine*, J. McNeill and others, eds. (London, Edward Arnold, 1990).
12. M. Winstanley, S. Woodward and N. Walker, *Tobacco in Australia: Facts and Issues 1995* (Melbourne, Australia, Victorian Smoking and Health Program, 1995), p. 15.

The cost to employers of employee alcohol abuse: a review of the literature in the United States of America

H. J. HARWOOD

M. B. REICHMAN

The Lewin Group, Falls Church, Virginia, United States of America

ABSTRACT

It is widely recognized that alcohol and drug abuse by workers can adversely affect their performance and the productivity of the workplace. The specific ways in which substance abuse can be harmful are well understood. Major elements of the costs incurred (for example, in lost productivity and earnings of workers and in deaths at the workplace) are captured in the most recent cost studies, as well as in the international guidelines for estimating the economic costs of substance abuse. However, no studies have rigorously measured the full economic burden on the workplace alone, because of the theoretical and empirical difficulties arising from the spread of the impact of substance abuse, via the markets, among employers (through lost profits), workers (through lost earnings and benefits) and consumers (through higher prices for goods and services). Both employers and workers recognize the nature of the problem and have worked together through bodies such as the International Labour Organization to find common solutions and formulate multilateral policies. Data for the United States of America show that policies are frequently established at the workplace to reduce alcohol and drug abuse by workers.

Introduction

It is widely recognized that alcohol abuse by workers can adversely affect their performance and the productivity of the workplace. The specific ways in which alcohol abuse can be harmful are fairly easy to understand and their impact can be readily analysed and quantified.

The cost methodology proposed under the international guidelines for estimating the economic costs of substance abuse recognizes and takes into account the impact of alcohol abuse, without providing for a detailed breakdown of workplace costs. However, recent studies of the cost of alcohol abuse to society have covered major elements of such costs, including lost productivity and earnings of workers and deaths at the workplace. In general, published studies that attempt to estimate the impact of alcohol abuse on the workplace alone have not been available.

The lack of such studies may be due to the major theoretical and empirical problems associated with attempts to estimate the costs incurred at the workplace as a result of alcohol abuse. Employers, as actors in markets for labour, other inputs, capital and goods and services, play a key role in spreading the impact of alcohol abuse among employers (lost profits), workers (lost earnings and benefits) and consumers (higher prices for goods and services). The distribution of costs between the parties cannot be determined by using theoretical concepts, but is subject to market conditions and the bargaining ability of employers, workers and consumers.

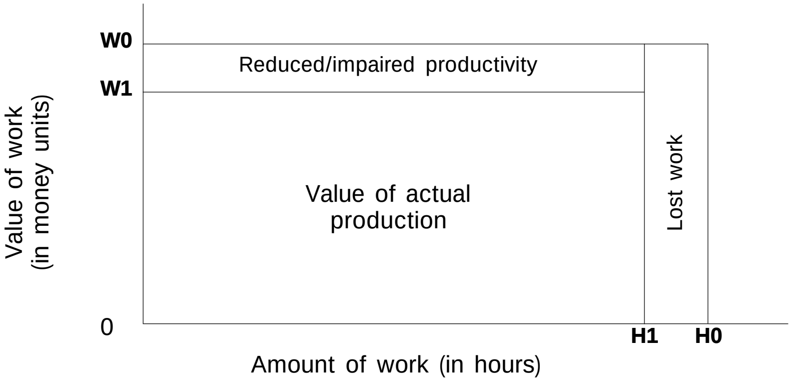
Despite uncertainty about the causal link between alcohol abuse and its apparent economic consequences, both employers and workers recognize the nature of the problem. They have worked together through bodies such as the International Labour Organization (ILO) to formulate multilateral policies to address the problem. Data for the United States of America show that workplaces have largely embraced policies to reduce alcohol abuse by workers.

The present paper is based on a literature search conducted in the following databases: ETOH, HealthSTAR, Information on Drugs and Alcohol (IDA), MEDLINE and Substance Abuse Information Database (SAID). The publications of the following bodies were also searched: the Canadian Centre on Substance Abuse (CCSA), the United States Department of Labor, ILO, the National Institute on Drug Abuse (NIDA), the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the United States Substance Abuse and Mental Health Services Administration (SAMHSA). The databases and lists of publications were searched using various combinations of the following search terms: cost, economics, employers, employees, workers, workplace, alcohol and alcoholism.

Economic theory and employee alcohol abuse

Economic theory may be directly applied to analyse the impact of alcohol abuse by workers in the workplace, as illustrated in figure I. Worker productivity can be described in terms of the amount and value of the work done. In a market economy, the wages paid to workers are expected to equal the value of their productivity to the enterprise. Earnings are by definition the amount of time worked multiplied by the wages paid per amount of time.

Figure I. Impact of alcohol abuse on the amount and quality of production


W0
W1
Reduced/impaired productivity
Value of work (in money units)
Value of actual production
Lost work
0
H1
H0
Amount of work (in hours)

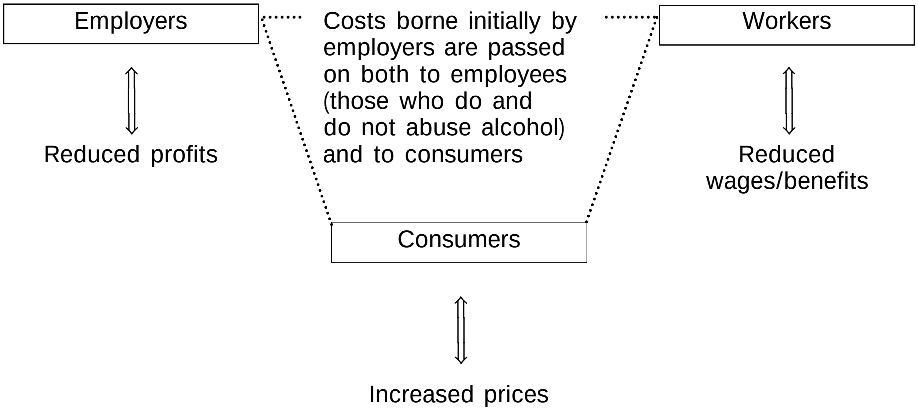
A review of the literature reveals that alcohol abuse can affect both the amount of time worked (through absenteeism, tardiness or extended sick leave) and the productivity of workers (by making them less quality-conscious and more mistake-prone). The problems experienced by the affected workers may also have an adverse impact on the productivity of their co-workers and managers. In figure I, lost work time is reflected in a shift from H0 to H1, and reduced or impaired productivity in a shift in wages from W0 to W1. Expected productivity is W0 x H0; actual productivity is W1 x H1. The difference in value represents the performance decline due to alcohol abuse.

The challenge to management arising from alcohol abuse by workers is to achieve levels of productivity as close as possible to the expected levels. That means minimizing lost work and poor productivity due to alcohol abuse. Matters are complicated, however, by the fact that employees miss work and underproduce for a number of reasons.

Costs to the workplace arise when the actual productivity of workers is less than expected for a given wage level. However, if a worker that abuses alcohol is less productive in a job that requires less productivity and pays less, there may be no difference between expected and actual employee productivity; hence, there is no cost to the workplace. It is the "hidden" alcohol abusers that cost the workplace money.

Economic theory notes two further complications (figure II). First, workplaces generally establish wages based on the value of productivity of all workers in a group, the average productivity. A group of workers with hidden alcohol abusers has a lower average productivity because of the few abusers, and the entire group therefore receives a lower wage. Thus, the workplace balances expected and actual productivity, although at the cost of paying the majority of workers less than the value of their individual productivity.

Figure II. Redistribution of alcohol-related costs in the workplace


Employers
Costs borne initially by employers are passed on both to employees (those who do and do not abuse alcohol) and to consumers
Workers
Reduced profits
Reduced wages/benefits
Consumers
Increased prices

Secondly, should the workplace misread (overestimate) average worker productivity, the costs would be higher because higher average wages would be paid. The increased costs may translate into reduced profits for the workplace; however, they may also cause an increase in the price of the goods and services provided. Thus, management failure to recognize productivity problems due to worker alcohol abuse (or any other factor) will mean either reduced profits or increased prices, or some combination of the two. The costs of worker alcohol abuse are therefore borne by workers, consumers and the workplace, and their respective shares are indeterminate, theoretically.

Behaviours and outcomes associated with employee alcohol abuse

A review of publications and other materials available from bodies such as the Department of Labor, ILO and NIDA revealed a range of workplace behaviours possibly related to alcohol. Employers should be aware of the consequences of alcohol abuse by workers at the firm level, including [1]: increased absenteeism and tardiness; increased insurance claims for treatment and services for alcohol abuse and its various sequelae; increased use of workers' compensation and sick leave; developing and implementing a substance abuse policy; testing for alcohol and drugs; development and administration of an employee assistance programme; accidents and damage; theft and fraud; increased turnover and replacement; diverted supervisory, managerial and co-worker time; friction among workers; poor decision-making; damage to a company's public image; and increased liability.

The above-mentioned consequences of alcohol abuse for the workplace are all confirmed by the substantial literature on the effects of alcohol on cognition and performance. The literature is compiled and summarized by NIAAA in publications such as the tri-yearly report entitled *Alcohol and Health* and the NIAAA peer-review journal *Alcohol Research and Health* (formerly *Alcohol Health and Research World*).*

Recent evidence showing the relationship of alcohol abuse to adverse workplace outcomes such as those enumerated above was collected in the United States from the National Household Survey on Drug Abuse (NHSDA). Table 1 presents data from the 1997 NHSDA on the workplace outcomes and behaviours that are associated with heavy alcohol use (generally not in the workplace) by workers.

Table 1. Adverse workplace outcomes reported for 1997 by full-time workers aged 18 to 49

Workplace outcome	*Employees reporting heavy alcohol use (percentage)*	*Employees not reporting heavy alcohol use (percentage)*
Worked for three or more employers in the past year	8.0	4.4
Missed two or more days of work in the past month due to illness or injury	12.4	8.5
Skipped one or more days of work in the past month	11.3	5.1
Voluntarily left an employer in the past year	19.7	15.8
Fired by an employer in the past year	0.9	1.4
Had a workplace accident in the past year	8.5	5.3

Source: Substance Abuse and Mental Health Services Administration, Office of Applied Studies, *National Household Survey on Drug Abuse* (Washington, D.C., 1997).

*See, in particular, vol. 19, No. 2, on alcohol and cognition (1995).

A clear increase in adverse workplace outcomes was reported by employees who were heavy users, defined as drinking five or more drinks on five or more occasions in the past month. Such outcomes are not unique to alcohol, however. The higher rates reported by heavy users indicate that they are more likely to have problems, but it is necessary to study whether the heavy users have other characteristics that may also be linked to those outcomes, such as their demographic characteristics. In fact, the population of heavy alcohol users is disproportionately composed of young males, who are also at risk of experiencing adverse workplace outcomes. An analysis will need to adjust for other factors (the attribution factor) to develop meaningful estimates of the impact of heavy alcohol consumption on workplace problems.

Alcohol-related job performance problems can be caused not only by alcohol consumption while on the job (often a cause for dismissal), but also by heavy drinking when not at work [2]. A 1986 study analysing the "hangover effect" in airline pilots is frequently cited in the literature [3, 4]. Using a flight simulator, Yesavage and Leirer studied the performance of airline pilots under three different alcohol conditions [5]. A single episode of moderate drinking (leading to an increase in blood alcohol concentration (BAC) to 0.10/100 millilitres) produced a dramatic increase, from 10 to 89 per cent, in the number of pilots unable to perform all operations correctly. Fourteen hours later, after the BAC of the pilots had returned to 0.0/100 millilitres, 68 per cent still could not perform all operations correctly.

Nevertheless, statistics regarding the impact of alcohol abuse on workplace outcomes must be interpreted with caution. While certain types of behaviour and outcome (for example, increased accidents) may be attributed, in part, to employee alcohol abuse, other factors undoubtedly contribute as well. For example, a 1989 study by Bernstein and Mahoney [6] found that up to 40 per cent of industrial fatalities and 47 per cent of industrial injuries can be linked to alcohol consumption and alcoholism. Researchers must disentangle the effects of alcohol from the effects of other factors such as gender, race, ethnicity, age, educational attainment, occupation and industry. The United States Bureau of Labor Statistics, in its census of fatal occupational injuries, reported 28,068 fatal occupational injuries in the United States from 1992 through 1996. The risk of occupational injury and death varies significantly across industries, with mining and quarrying, agriculture, construction, transportation and public utilities all reporting substantially higher rates of unintentional injuries and deaths [7].

Another example is provided by a large-scale study concerning illicit drug use and job performance in the United States Postal Service [8]. The authors concluded that drug use as exhibited by a positive pre-employment test was significantly associated with turnover, absenteeism and other negative workplace outcomes. Butler cites the following results from another study conducted within the United States Postal Service: "Applicants who had tested positive for illicit drugs were 2.67 times as likely to experience problems requiring employee assistance programme intervention, 2.44 times as likely to be formally disciplined, and 1.88 times as likely to accumulate an elevated dollar amount of medical claims than those who had tested negative for drug use" [3]. All such outcomes incur costs for the employer, the employee and others. A portion of the costs may be directly attributed to employee substance abuse. However, a cost-of-illness study that immediately assigns the full cost to drug abuse is misleading.

Prevalence of alcohol abuse among United States workers

NHSDA, conducted annually since 1990 by SAMHSA, is the primary source of statistical information on the use of alcohol, tobacco and illicit drugs in the United States. The 1994 and 1997 NHSDA included a special module designed to collect data on the use of illicit drugs and alcohol by United States workers and on company policies on drug and alcohol use. An analysis of the 1997 NHSDA was released in 1999 by the SAMHSA Office of Applied Studies [9]. The analysis included the following findings regarding the prevalence of current heavy alcohol use* among workers employed in various occupations and establishments of different sizes (see table 2):

(a) Among full-time workers, 6.2 million, or 7.6 per cent, reported current heavy alcohol use;

(b) Heavy alcohol use was more common among 18- to 25-year-olds *(*11.7 per cent), males *(*11.1 per cent), Hispanics *(*9.8 per cent) and whites *(*8.1 per cent), those with less than a high school education *(*14.7 per cent), and those who reported personal income of $9,000-$19,999 *(*9.4 per cent);

(c) Workers in medium-size establishments *(*25-499 employees) reported the highest rate of heavy alcohol use *(*8.3 per cent);

(d) Current heavy alcohol use varied across occupational categories, the highest rates being accounted for by food preparation workers, waiters, waitresses and bartenders *(*15 per cent); handlers, helpers and labourers *(*13.5 per cent); and construction workers *(*12.4 per cent).

Table 2. Current heavy alcohol use among full-time United States workers aged 18 to 49, 1994 and 1997

	Percentage of heavy alcohol users	
Item	*1994*	*1997*
Establishment size		
1-24 employees	9.6	7.0
25-499 employees	7.9	8.3
Over 500 employees	7.3	7.4
Occupational category		
Food preparation, waiters, bartenders	12.2	15.0
Handlers, helpers, labourers	15.7	13.5
Construction	17.6	12.4
Precision production and repair	13.1	11.6
Other services	5.1	11.4
Transportation and materials moving	13.1	10.8
Machine operators and inspectors	13.5	9.0
Protective services	6.3	7.8
Executive, administrative, managerial	6.5	7.1
Extraction and precision production	12.9	5.5
Technicians and related support	6.2	5.1
Administrative support	3.5	5.1
Professionals (specialists)	4.4	4.4
Sales	8.3	4.1

Source: SAMHSA, Office of Applied Studies, *National Household Survey on Drug Abuse* (Washington, D.C., 1994 and 1997).

*Heavy alcohol use is defined as consuming five or more alcoholic beverages on the same occasion on five or more occasions in the past 30 days.

It is clear that many people who abuse alcohol are employed. Though the rate varies across establishment size, occupational categories and regions, all employers are vulnerable to the adverse effects of employee alcohol abuse. The first step that employers can take to mitigate the impact of alcohol abuse in the workplace is to understand and recognize what the impact is.

Workplace efforts to reduce employee alcohol abuse

Workplaces in the United States generally take specific measures to prevent or discourage employee alcohol abuse. Recent workplace surveys have found that employers (in particular medium- and large-scale employers) have several policies or services intended for that purpose. They include the distribution of basic educational and information materials, the provision of employee assistance programmes and insurance for treatment of alcoholism, policies to counter alcohol abuse and even the testing of workers for improper use of alcohol.

The United States Department of Labor [1] suggests a five-part programme to control substance abuse in the workplace. Each component is also found in the recommendations established by ILO, including the following: a written substance abuse policy; an employee education and awareness programme; a supervisor training programme; an employee assistance programme; and drug- and alcohol-testing, as appropriate.

Employer efforts to prevent or reduce the incidence of substance abuse among employees vary by establishment size and industry (see table 3). In general, the

Table 3. Workplace policies covering full-time United States workers aged 18 to 49, 1997

(Percentage)

Item	*Information about drug and alcohol use*	*Written policy about drug and alcohol use*	*Employee assistance programme*
Establishment size			
1-24 employees	50.7	45.9	26.8
25-499 employees	87.1	84.6	60.6
Over 500 employees	92.3	90.4	75.4
Occupational category			
Protective services	95.2	92.6	79.7
Transportation and materials moving	85.9	83.3	46.5
Extraction and precision production	85.5	81.5	65.0
Technicians and related support	78.6	75.3	61.7
Administrative support	81.4	77.7	54.7
Machine operators and inspectors	79.6	76.3	52.7
Precision production and repair	75.3	72.4	48.4
Professionals (specialists)	75.2	72.4	56.4
Handlers, helpers, labourers	71.7	70.0	51.3
Executive, administrative, managerial	71.5	70.3	49.7
Other services	67.1	62.8	36.2
Food preparation, waiters, bartenders	66.8	63.4	26.5
Sales	62.2	59.3	43.3
Construction	60.0	55.6	30.3

Source: SAMHSA, Office of Applied Studies, *National Household Survey on Drug Abuse* (Washington, D. C., 1997).

larger the establishment, the more likely it is: *(a)* to provide information about alcohol and drug use; *(b)* to have a written policy about drug and alcohol use; *(c)* to offer an employee assistance programme; and *(d)* to test current employees or applicants for alcohol use [1, 10, 11]. In 1997, in the occupational categories in which employers were most likely to offer information about, and to have a written policy on, drug and alcohol use, the percentage that did so for drugs and alcohol, respectively, were: in protective services, 95.2 and 92.6 per cent; in transportation and the moving of materials, 85.9 and 83.3 per cent; and in extraction and precision production, 85.5 and 81.5 per cent [9]. The occupational categories in which employers were most likely to offer an employee assistance programme were protective services (79.7 per cent), extraction and precision production (65 per cent) and technicians and related support (61.7 per cent).

Workplace testing for alcohol use, though still less common than other policies for dealing with alcohol abuse, is increasing in prevalence. A national survey of workplaces conducted in 1995 found that 21.7 per cent tested applicants and 28.4 per cent tested current employees (see table 4).

Table 4. National estimate of alcohol testing among private non-agricultural workplaces by size, industry and census region, 1995

(Percentage)

	Workplace with testing programmes, by test group		
Characteristics	*Applicants*	*Current employees*	*Applicants and current employees*
All workplaces	21.7	28.4	36.0
Workplace size			
50-99 employees	19.2	25.0	31.4
100-249 employees	19.7	26.8	33.6
250-999 employees	26.5	34.1	42.9
Over 1,000 employees	33.3	40.0	55.2
Industry			
Manufacturing	33.7	38.1	49.5
Wholesale/retail	19.4	26.3	33.1
Communications, utilities, transportation	30.3	47.4	52.7
Finance, insurance, real estate	5.8	7.3	11.0
Mining and construction	29.3	39.0	45.5
Services	13.3	20.2	26.5

Source: [10].

Alcohol testing programmes can be divided into two main types: conditional and unconditional [10]. Conditional testing consists of assessments conducted for certain employees, taking into account factors such as an accident or the determination of reasonable cause. Unconditional testing includes random and regular assessments that any employee may be required to undergo regardless of job performance or behaviour. On the whole, conditional testing is more prevalent than unconditional testing. Within workplaces that conduct conditional testing, 73.9 per cent reported testing employees following an accident and 77.9 per cent reported testing after establishing reasonable cause. Among workplaces conducting unconditional testing, random testing (48.4 per cent) is far more prevalent than regular testing (12.4 per cent).

Though alcohol testing continues to increase in prevalence, very few alcohol-testing programmes exist in workplaces that do not also test for illicit drugs. Furthermore, alcohol testing rarely occurs as a single strategy to reduce employee alcohol abuse. Testing programmes almost always occur in conjunction with formal policies and employee assistance programmes [10, 11].

Alcohol testing, on its own, is considered to be an ineffective and misleading way for an employer to combat employee alcohol abuse. A positive alcohol test does not necessarily indicate impairment. Testing programmes should be implemented in conjunction with a formal written policy that establishes the purpose for testing, rules, regulations, rights and responsibilities of all the parties concerned [3, 10, 12].

The data clearly show that employers take the problem of alcohol abuse among employees very seriously. The majority of employers, especially in large workplaces, have some form of substance abuse policy or programme. Employers have not had the opportunity to base their decisions to develop and implement substance abuse policies on empirical evidence. There is an abundance of material showing the prevalence of alcohol abuse among employees and its impact on the workplace. Such data are available at both the national and industry level [3, 9, 12]. In general, however, there is a lack of data for an analysis of the impact of alcohol abuse on specific workplaces and on the employers concerned.

Can employers measure the cost of employee alcohol abuse?

While employers have access to a wealth of information regarding the prevalence of alcohol and other drug use among workers, most estimates of the prevalence and consequences (including the cost) of alcohol and other drug use are derived from large surveys of households, such as NHSDA [9]. While information at the national level is useful, employers must base policy decisions regarding employee substance use and abuse on assessed needs in their own workplaces. Estimates of prevalence and cost at the national level may be of little use to employers in deciding on policy at the workplace or company level. Table 2 shows that rates of employee alcohol abuse vary widely across occupational categories, regions and demographic characteristics. At the very least, employers should take into account the basic demographic make-up of their workforce, determined by such factors as gender, age and educational attainment, as well as industry and occupational category, when assessing data at the national level.

The literature provides few examples of the cost of alcohol abuse in specific companies and workplaces. The vast majority of the literature on alcohol and other drug abuse in the workplace pertains to incidence and prevalence rather than evaluations of cost [3, 9, 13, 14]. Several studies have relied on self-reporting of alcohol and drug use and related workplace outcomes through employee surveys [13, 14]. Employee surveys can be valuable because workers are in the best position to assess the problem, provided that steps are taken to ensure the anonymity and confidentiality of reported results [15].

A 1995 study by French and others not only measured the prevalence of smoking, alcohol, illicit drug use and prescription drug use, but also reported workplace consequences such as reduced performance and absenteeism [13]. The study used an employee survey administered to more than 1,200 randomly selected employees at five different workplaces (manufacturing services, manufacturing, municipal Government, financial services and health-care services). The survey asked employees whether their drinking and related problems (such as a hangover, an illness or an accident) had caused any of the following workplace consequences: poor performance, tardiness or leaving work early, absence, accident and injury, being high at work and needing emergency care. At three of the five workplaces, approximately 20 per cent of the drinkers reported poor performance attributable to the effects of alcohol during the past year. Fewer than 10 per cent of drinkers at each site indicated that they had been tardy or absent, or that they had left work early because of alcohol use. None of the respondents reported being hurt at work in an accident due to alcohol use, and less than 1 per cent at each site revealed that they had been high at work during the past year.

While the authors do not attempt to estimate the costs of employee drinking, they do use an employee survey to assess the workplace consequences. From a policy perspective, the value of the work of French and others is that it presents empirical results from a unique and relatively small data sample. More studies of that kind would assist employers in measuring the impact on performance and possibly the costs of substance abuse in the workplace.

In 1988, the American Productivity and Quality Center, a non-profit body that works with organizations to improve productivity and quality, undertook a research project funded by the United States Department of Labor [1]. The purpose of the project was to determine whether the costs associated with substance abuse in the workplace could be tracked. On the basis of discussions with and reviews by experts from organizations such as the American Federation of Labor and Congress of Industrial Organizations, the National Academy of Sciences, NIAAA, NIDA and the Social Security Administration, the research team developed a “total costs model” that organizations could use to calculate both direct and indirect costs. The model requires the collection both of objective financial data and of more subjective performance data. The data collected were the costs incurred in dealing with the consequences of alcohol abuse listed above in the section entitled “Behaviours and outcomes associated with employee alcohol abuse”. The only items not covered in the data collected were the increased use of workers’ compensation and sick leave; friction among workers; poor decision-making; damage to a firm’s public image; and increased liability. The performance data were collected via two subjective employee surveys. Sample questions included the following: “From your observations, how do substance abuse problems interfere with job performance?”; and “Have you observed substance abusers having any of the following problems: missed deadlines, excessive mistakes, decreased efficiency, etc.?”

Three companies agreed to conduct a pilot test of the model, without having full access, in each case, to all of the data included in the model. However, the research team refined the model to make it adaptable to different types of organization with different data capabilities. The companies that tested the model indicated that it provided valuable information upon which to base future cost-benefit estimates. Two of the three companies indicated that they would continue to use the model. On the basis of the pilot test of the model, the research team concluded that the costs typically tracked by employers are only the tip of the iceberg. While the model proved to be a good way for the three companies that tested it to assess the impact of substance abuse on their workplaces, certain types of impact and costs still need to be measured. The data collected usually fail to capture the so-called ripple effect, that is, “costs incurred when mistakes are made by substance abusers and problems and/or errors then occur down the line or ripple throughout the rest of the organization” ([1], p. 65).

The model has several limitations. The way in which the performance data are collected could bias individual responses. The model calls for collecting data about the job performance of employees from supervisors in an interview format. Data collected in such a manner are inherently subjective and thus difficult to aggregate and compare across companies. In addition, a small sample size of only three companies, with an unknown number of employees being interviewed at each company for performance data, makes it impossible to generalize from conclusions drawn about the usefulness of the total costs model.

The total costs model was the only one of its kind found in the literature reviewed. Although some organizations may not have all the required data at their disposal, such a model would provide valuable guidance to employers in evaluating the cost of substance abuse in their workplaces.

Workplace costs and international guidelines

The international guidelines for estimating the economic costs of substance abuse [16] are designed to cover the types of cost that result from worker abuse of alcohol. Table 5 shows the relationship between the major cost categories of the guidelines and a range of specific impacts identified in the literature review. Most of the impacts fit into the major cost categories of the guidelines.

Despite the conceptual fit, there are two problems. First, some of the specific impacts have not yet been evaluated in cost-of-illness studies. For a variety of reasons, such as lack of data and limited resources, studies have not yet determined those costs. That applies, in particular, to workplace-specific costs, such as those arising from absenteeism and tardiness, increased turnover and replacement, diverted co-worker and managerial time and friction among workers.

Table 5. Cost categories, international guidelines and estimating the impacts of alcohol abuse

International guideline (cost category)	*Impacts dealt with in the literature (cost category)*
Consequences to health	Increased insurance claims for alcohol abuse treatment and services and for the various sequelae of alcohol abuse
Prevention, research etc.	Development and administration of an employee assistance programme
Premature mortality	Accidents and damage
Lost employment or productivity	Increased absenteeism and tardiness; increased use of workers' compensation and sick leave; increased turnover and replacement; diverted supervisory, managerial and co-worker time; friction among workers; and poor decision-making
Property destruction (fire, accident)	Accidents and damage and increased liability
Legal consequences	Theft and fraud
Outside the guidelines	Developing and implementing a substance abuse policy and testing for alcohol or drug abuse

Secondly, the costs may have been subsumed under broader estimation categories and dealt with, in whole or in part, in recent cost-of-illness studies. For example, the costs of workplace accidents are generally included in the costs of the following: all accidents, injuries and trauma; health care; lost work due to morbidity, disability and premature death; and property destruction.

Another factor that must be taken into account in carrying out studies in different countries is the difference in economic and social systems. In some countries, for example, health services are paid for in part by insurance provided at the workplace (often with cost-sharing by the workers), while in others, national health systems are financed through payroll taxes and other tax revenues. Thus, issues related to the burden or location of costs need to be understood before workplace costs can be estimated.

Conclusion

There is ample evidence that alcohol problems have major impacts beyond the workplace. A wide range of such impacts have been studied and their costs estimated, but even more remain to be considered. Studies based on the international guidelines for estimating the costs of substance abuse capture many of the costs of workplace alcohol abuse, even though a detailed breakdown and analysis of workplace costs may not be provided.

In general, the available published studies fail to estimate the costs of alcohol abuse specifically related to the workplace, in particular those that fall upon employers. The studies of costs to society include important components of costs beyond the workplace, but they explicitly refrain from attributing specific amounts to workplaces and, in particular, to employers.

Economic theory shows that it is very difficult and potentially meaningless to develop estimates of the cost of alcohol problems for a particular workplace or population of workers. That is because the impacts felt and costs incurred beyond the workplace are shared or redistributed in uncertain proportions between employers, workers, customers and consumers.

Alcohol abuse by workers is clearly a matter of great concern to employers in the United States, most of whom have established policies, services, benefits or initiatives dealing with the issue. Such efforts are designed to counter the potential economic harm caused by alcohol in the workplace and by its negative impact on the efficiency, quality and costs of the business organization.

References

1. United States Department of Labor, *Substance Abuse in the Workplace: a report and Total Costs Model Prepared for the United States Department of Labor* (Washington, D.C., 1989).

2. G. M. Ames, J. W. Grube and R. S. Moore, "The relationship of drinking and hangovers to workplace problems: an empirical study", *Journal of Studies on Alcohol*, vol. 58, No. 1 (1997), pp. 37-47.

3. B. Butler, *Alcohol and Drugs in the Workplace* (Vancouver, Butterworths, 1993).

4. L. Alyanak, “Substance abuse at work”, *World of Work: The Magazine of the ILO*, No. 30 (International Labour Organization, Geneva, 1999).

5. J. A. Yesavage and V. O. Leirer, “Hangover effects on aircraft pilots 14 hours after alcohol ingestion: a preliminary report”, *American Journal of Psychiatry*, vol. 143, No. 12 (1986), pp. 1546-1550.

6. M. Bernstein and J. J. Mahoney, “Management perspectives on alcoholism: the employer’s stake in alcoholism treatment”, *Occupational Medicine*, vol. 4, No. 2 (1989), pp. 223-232.

7. National Safety Council, *Safety Facts 1999* (Chicago, Illinois, 1999).

8. J. Normand, S. Salyards and J. Mahoney, “An evaluation of pre-employment drug-testing”, *Journal of Applied Psychology*, vol. 75 (1990), pp. 629-639.

9. Z. Zhang, L. X. Huang and A. Brittingham, *Worker Drug Use and Workplace Policies and Programs: Results from the 1994 and 1997 National Household Survey on Drug Abuse* (Rockville, Maryland, SAMHSA, Office of Applied Studies, 1999).

10. T. D. Hartwell, P. D. Steele and N. F. Rodman, “Workplace alcohol-testing programs: prevalence and trends”, *Monthly Labor Review*, vol. 121, No. 6 (1998).

11. H. Hayghe, “Survey of employer drug programs”, in *Drugs in the Workplace: Research and Evaluation Data*, vol. II, NIDA Research Monograph Series No. 100, S. W. Gust, J. M. Walsh, L. B. Thomas and D. J. Crouch, eds. (Bethesda, Maryland, NIDA, 1991), pp. 177-207.

12. International Labour Organization, *Management of Alcohol- and Drug-related Issues in the Workplace* (Geneva, 1996).

13. M. T. French and others, “Prevalence and consequences of smoking, alcohol use, and illicit drugs at five worksites”. *Public Health Reports*, vol. 110 (Oxford University Press, New York, 1995), pp. 593-599.

14. W. E. K. Lehman and D. D. Simpson, “Patterns of drug use in a large metropolitan workforce”, *Drugs in the Workplace: Research and Evaluation Data*, vol. II ..., pp. 45-62.

15. B. Butler, “Employee opinion: workplace and other drug use” (undated paper).

16. E. Single and others, *International Guidelines for Estimating the Economic Costs of Substance Abuse* (Ottawa, Canadian Centre on Substance Abuse, 1995).

Attributable fractions for alcohol and illicit drugs in relation to crime in Canada: conceptualization, methods and internal consistency of estimates

K. PERNANEN

National Institute for Alcohol and Drug Research, Oslo, Norway and Uppsala University, Uppsala, Sweden

S. BROCHU

International Centre for Comparative Criminology, University of Montreal, Montreal, Canada

M.-M. COUSINEAU

International Centre for Comparative Criminology, University of Montreal, Montreal, Canada

L.-G. COURNOYER

Université du Québec à Hull, Hull, Canada

FU SUN

International Centre for Comparative Criminology, University of Montreal, Montreal, Canada

ABSTRACT

A research programme in Canada is aimed at estimating attributable fractions for the use of alcohol and illicit drugs in relation to crime. Analyses from two studies of new inmates in federal penitentiaries are presented, the first based on a computer-driven questionnaire completed by 8,598 inmates and the second on interviews with 477 inmates. One method used in the estimation combined the following three models linking psychoactive substances to criminal behaviour: the intoxication model, the economic model and the systemic model. Data pertaining to the first two models were used to illustrate this method. Consistency checks showed that crime events attributed to illicit drugs or alcohol were concordant with the inmate being addicted to a substance, and with the inmates' overall assessments of drugs or alcohol on their criminality. Issues discussed include validity, the extent to which findings can be generalized and the advantages and drawbacks of basing attributable fraction estimates on data from self-reports on individual crime events.

Introduction

The present article outlines the steps taken to arrive at estimates of attributable fractions for drugs and alcohol in relation to crime in Canada. In particular, the article looks at a method that will enable estimates to be made for important subgroups of the population and different types of drugs and types of crimes.

Some of the empirical studies needed for making overall estimates are still at the fieldwork stage, and therefore no final estimates can be made. However, the following is presented: *(a)* the method; *(b)* its conceptual background; *(c)* some findings that will lead up to the estimates; and *(d)* some methods used for checking the robustness (the internal consistency) of the estimates. Some preliminary calculations that give an indication of the range of attributable fractions for federal inmates in Canada are also provided.

The empirical material available for the estimates in the present article is limited to information on inmates in federal prisons in Canada. In all probability, that population incurs costs to society far beyond its numerical size. For a full cost accounting, however, data on drug and alcohol use and criminality patterns in other populations is also needed, in particular with regard to individuals arrested for a variety of crimes and inmates in provincial prisons, who generally have committed less serious crimes than federal inmates have.

Conceptual background

Any estimation procedure for the causal contribution of psychoactive substances on crime is based on key conceptual assumptions. These assumptions are based on findings from past empirical research in the field. Applying the conceptual frames to social reality requires data that allow assessments regarding how applicable alternative causal models are. Because such conceptually relevant data have been missing, researchers have sometimes refrained from making any estimates on the crime component in the social costs of alcohol and illicit drugs. This has been the case for recent social cost estimates in Canada. In other studies, estimates have been based on questionable conceptual and empirical assumptions.

The starting point for the present calculations is to use the following three models, which assign different causal roles to drugs and alcohol in relation to crime: *(a)* the pharmacological or intoxication model; *(b)* the economic means model; and *(c)* the illegal system model. The models used are a tripartite collection that Goldstein used for classifying drug-related violence. In addition, in the fourth model in the series, some crimes are alcohol-related or drug-related by legal definition. However, they are solidly based on empirical findings from research on a variety of crimes.

The intoxication model attributes a direct causal role to a substance used at the time of a crime. The assumption is that intoxication made, or helped make, the individual commit an illegal act that he or she would not otherwise have committed. In the study of the effects of alcohol, this model is often referred to as a “disinhibition” model. It has been used frequently in various estimations of the role of alcohol on crime, specifically in calculating attributable fractions linked to violent crime. Since no other information has been available regarding the causal role of alcohol, it has been assumed that all crimes in which the perpetrator had been drinking were caused by drinking, that is, that if the perpetrator of the violent crime had not been intoxicated at the time, he or she would not have committed the crime. With one important modification, this model is also used as part of the four model conceptual frame for estimating attributable fractions for drugs and alcohol on crime.

The second causal model used, the economic means model (or, in Goldstein's terms, the economic-compulsive model) pertains mainly to the role of drugs and to a much lesser extent to alcohol, as motivators in predominantly acquisitive crimes. Psychoactive substances serve as incentives for a person to commit a crime so that they will get money or other means for acquiring drugs or alcohol.

The third causal model, the systemic model, concerns crimes that were committed, for example, in the course of selling drugs, collecting drug debts and conflicts over drug territory. If it can be assumed that the individual would not have committed these crimes had he or she not been involved in the illegal economy, the crime can be considered to be caused by the drug being present as a commodity within a system of illegal transactions and enforcement methods.

The fourth, the substance-defined model, is not a causal one, but represents a tautological connection with alcohol and drug use. The crimes in this category are included on the basis of laws regulating alcohol and drugs in society. Drinking and driving is by far the most common of the alcohol-defined crimes. Several drug-related offences, such as the manufacture, smuggling and trafficking of drugs, are included in the category of drug-defined crimes. Possession and use of most illicit drugs are also defined as criminal acts in many countries and would be covered by this model. (The present estimates are limited to the population of federal inmates, therefore few cases of minor drug crimes will appear in the estimates.) These will be added as the fourth stage of the construction of the composite attributable fraction model.

Some overlap can be expected between the positive cases in the four models in any population. A certain proportion of individuals who committed a crime under drug intoxication were also driven by the motive to get more drugs for personal use, for instance (so as to prevent their supply from running out). In a similar way, some individuals who used violence to collect a drug debt for themselves or for someone else in the distribution chain did so in order to get drugs or the means to buy drugs for personal use. If there is a great deal of overlap between the positive cases from the three models, it can be inferred that any one of the models would have provided a good approximation of the attributable fraction. More importantly, however, a great deal of overlap is also an internal confirmation of the validity of the combined measure.

Individual-level versus aggregate-level data

As shown below, it is possible with the right kind of data to use aggregate-level data to calculate attributable fractions for some factors on crime. Time series analyses have been used for this purpose. However, they require relatively valid data that has been collected over lengthy periods of time. Such data exist for alcohol but do not exist for illegal drugs. For illicit drugs, aggregate-level analyses cannot therefore be used for calculating attributable fractions.

Aggregate-level data have certain weaknesses for the purpose of causal attribution. They are dependent on the availability of data on other potential causative factors to be used as controls. Even for alcohol consumption, no statistical series exist on alcohol use in different subgroups. Using this method, it is therefore not possible to calculate attributable fractions for alcohol on, for example, crimes committed by those under the age of 30 and over, or for men versus women.

Using individual-level data for estimation makes it possible to distinguish individual cases from non-cases on attributable fraction variables. This means that attributable fractions can very easily be arrived at, for example, for different types of crimes or for different subgroups of offenders. Given a large enough sample of perpetrators or crime events, estimates can be obtained of what proportion of violent crimes among perpetrators under 30 years of age were attributable to alcohol or drugs.

The greatest drawback of individual-level data is that they must be collected by special studies. In addition, much of the information has to be based on self-reports, and it is known that self-reports may be unreliable. The questions generally concern past behaviour with a risk for memory lapses. In the case of sensitive information, social desirability may affect the validity of the information given.

The present data are based on information about inmates in federal penitentiaries. All information (all the sample units) represents cases on the dependent variable: there are no individuals in the sample who did not commit a crime and who could serve as controls for analyses assigning an explanatory value to independent variables, such as alcohol or drug use, or risk calculations linked to alcohol and drugs. This places restrictions on the type of analyses that can be made.

The ideal type of study, which would allow a flawless estimation of attributable fractions, may be easily specified in theory but, at the present time at least, is impossible to conduct in practice. The choice is between methods that are all lacking in some respects. The choice depends primarily on the data available for the purpose and the modelling preferences of the researchers.

The empirical studies in the research programme

Two studies are used for estimation in the present article: the Computerized Lifestyle Assessment Instrument (CLAI) data made available by the Correctional Service of Canada (CSC) and interviews with 477 male inmates in federal prisons in Ontario and Quebec (the Pathway study). The latter study was specifically conducted for the estimations of the project. In some respects, the studies complement each other and in other respects they support alternative estimation methods. If the alternative estimates are within a tolerable range of differences, confidence in their robustness is strengthened.

Data collection has been completed in a third study relevant to assigning an attributable fraction for alcohol and illicit drugs in relation to crime in Canada: information on arrests made in 24 locations in Canada during a one-month period (1-31 May 2000). These data were collected by police officers and are based on information available at the time of the arrest. Most sites were selected on the basis of a stratification of communities in Canada according to population size: *(a)* two megacities (with populations of over 1 million); *(b)* three large cities (with populations of between 500,000 and 1 million); *(c)* three medium-sized cities (with populations of between 250,000 and 500,000); *(d)* six small cities (with populations of between 100,000 and 250,000); and *(e)* ten other communities of interest. Within categories *(b)*, *(c)* and *(d)*, sites have been selected on the basis of statistical information on their overall crime rate: one characterized by a relatively high crime rate, one with a medium rate and one with a low crime rate. (Such stratification was not possible in the “megacity” category.) Data from the study of arrestees add information on another population of relevance to the connections between drugs, alcohol and crime.

Three smaller-scale studies were later added to the research programme for the purpose of extending descriptive and causal objectives to other key populations: interviews with 100 male inmates in provincial prisons; interviews with 100 male provincial probationers; and interviews with 100 female inmates in provincial prisons.

Information on the crime and substance use patterns of provincial inmates and probationers is important because criminality differs between these two populations and the federal inmates who are the subject of the present article. Convicted criminals with a sentence of at least two years' imprisonment serve their sentences in federal penitentiaries, those with a lesser prison sentence serve their time in provincial custody.

The Computerized Lifestyle Assessment Instrument and the Pathway studies

The Computerized Lifestyle Assessment Instrument study

The Computerized Lifestyle Assessment Instrument (CLAI) is both a diagnostic tool and a survey instrument used by the Correctional Service of Canada. It is administered to all federal inmates upon admission to an assessment centre, prior to their being sent to an institution. CLAI helps in taking into account treatment and other individual needs of the inmate. There are detailed questions on alcohol and drug use and criminal activities. Such details make the database uniquely suitable for some of the purposes of the present article.

The data are collected by means of a computer-driven questionnaire: the inmates enter responses to questions that appear on a computer screen. It takes an average of two hours to fill out the questionnaire.

The data collection on the CLAI started in 1990. Over the years, an increasing number of penitentiaries have contributed information to the computer file. In the file to which the authors have access, there is information on close to 17,000 inmates. The best geographical coverage is for the period 1993-1995 and has been selected for the present analyses (N=8598). However, the differences between the total file and the subfile are generally negligible.

The Pathway study

In this study, 477 inmates were interviewed at regional reception centres in Ontario and Quebec. The data were collected between September 1999 and January 2000 in Ontario and between February and December 1999 in Quebec. The most central data collection instrument of the study was a calendar used in charting the 36 most recent months in the inmate's life prior to arrest. The focus was on several aspects of the relationship between drug and alcohol use and criminal behaviour. Much of this information was not available in the CLAI data.

The Pathway study also incorporated central questions from the CLAI on the inmate's drug and alcohol use and criminality. It included the same tests for dependence on alcohol (the Alcohol Dependence Scale) and drugs (the Drug Addiction Severity Test), enabling an aggregate-level reliability check of estimates made from the two studies.

Differences between the CLAI and the Pathway studies

A major difference between the studies is that the CLAI drew its population from all five regions of the Correctional Service in Canada (Atlantic, Ontario, Pacific, Prairie and Quebec), while the Pathway study was carried out on inmates in the Ontario and Quebec regions only. The decision to limit that study to the two regions was made for financial reasons. Analyses will be conducted to determine what effect the difference in geographical base of the inmate populations may have for the results and how any biases can be corrected.

The time periods of data collection also differ. As was pointed out above, information from 1993-1995 has been used to arrive at the CLAI estimates in the present article. The Pathway data, on the other hand, were collected an average of five to six years later. Analyses of the CLAI data from the Quebec region (which has been providing data regularly since the early 1990s) showed very small differences in drug-use patterns among inmates over a period of six years.

The selection procedure for federal inmates in the two studies was about the same, although the CLAI study was a census (with some attrition), while the Pathway study used a random sampling procedure to select incoming inmates to the study. A sampling procedure was necessary because the inflow of new prisoners was greater than what could be handled by two interviewers, which was the maximum number possible for logistical reasons. In both studies, the inmates participated in the interview about two weeks into their stay at the reception centres.

Reasons for attrition in the Pathway interviews at the Ontario and Quebec reception centres are shown in table 1. The response rate in the Ontario part of the study, calculated out of those who were contacted for an interview, was 84.8 per cent. The corresponding figure for the Quebec part of the study was 78.9 per cent.

Table 1. Pathway study: sample attrition and response rates in Ontario and Quebec

	Response rates	
Subjects	*Ontario*	*Quebec*
Total sampled	342	419
Not eligible	10	44
Not available	35	90
Approached	297	285
Refused	44	56
Interrupted	1	4
Interviewed	252	225
Response rate *(*percentage*)*	84.8	78.9

As was mentioned above, the main data collection instrument of the Pathway study is a 36-month calendar onto which a monthly record of drug-use patterns and criminality was recorded. Whereas the CLAI instrument only asked for crime-specific information on alcohol and drug use in connection with the most serious crime on the present sentence, the calendar instrument asked for this information on all self-reported crimes over a three-year period. Many of these crimes have remained undetected by authorities. The great number of crime episodes and the longer reference period, among several other features, adds to the power of the analyses that can be performed on the Pathway data.

While the CLAI data were entered by the inmate in response to questions and response alternatives appearing on a computer screen, the calendar data were filled in by the interviewer while she (all research assistants on the study were women), together with the respondent, consulted the calendar in order to place occurrences in the correct time period on the calendar. Other data collection instruments in the Pathway study were filled out as a write-in questionnaire by the inmate, although he was allowed to ask questions on the meaning of questions, for example. Most of the interview session was spent on the calendar.

The relationship between estimates and possible biases

Comparisons between the CLAI and the Pathway studies must take into account that they pertain to somewhat different geographical areas and to time periods that are five to six years apart on average. In the CLAI data from the period 1993-1995, 37 per cent of the new inmates had been admitted in the Ontario region and 40 per cent in the Quebec region. Thus, approximately 23 per cent of inmates in the CLAI data file are from outside the two regions. A correction factor will be used in generalizing findings from the Pathway study to the total federal inmate population in Canada.[*]

Three major limitations remain, even with the corrections outlined above:

(a) All the estimates presented in the present article pertain to inmates in federal prisons in Canada only. Generalizations to similar populations should be made with great caution. Separate studies will be carried out on provincial prisoners and individuals who have been arrested for a crime (as was mentioned above);

(b) All the estimates presented here pertain to male offenders only. The Pathway study was carried out on male inmates exclusively. The Correctional Service of Canada collects CLAI data on both male and female inmates. However, the CLAI file to which the authors have access currently does not contain data on female prisoners. Female inmates make up only about 2 per cent of the federal inmate population, and special characteristics of this group would have a negligible effect on the overall estimates pertaining to the federal inmate population;

(c) All the findings in the CLAI and the Pathway studies were based on information provided by the inmates themselves. Giving of false information cannot be ruled out, and memory lapses may also play a part. This, of course, is a shortcoming common to much of the social-survey-type research in sensitive areas of study.

Constructing the estimates

The intoxication model data

Some of the questions in the CLAI study were used in the Pathway study, partly for the purpose of replicating the questions with a different data collection method.

[*] Preliminary findings indicate that the attributable fraction on crime for alcohol and the combined use of drugs and alcohol is considerably higher in western Canada than in Ontario and Quebec, while the fraction for drugs only is very similar between those regions.

Crime-specific information on drugs and alcohol use was available for the most serious crime on the inmate's current sentence. The proportions of inmates who were under the influence of a substance when committing such a crime are as follows:

Associative fractions from intoxication model

	CSC-CLAI	*Pathway-CLAI*
Drugs	0.16	0.20
Alcohol	0.21	0.19
Drugs and alcohol	0.13	0.14
No substance	0.50	0.47
Total	1.00	1.00

These figures correspond to the one-model estimates of attributable fractions sometimes used in social cost calculations. If a restricted focus is maintained on the attributional fraction for alcohol on crime, an attributional fraction of 0.34 (0.21+0.13) would be arrived at, on the basis of the CSC study, and 0.33 on the basis of the Pathway study. For drugs, the fractions would be 0.29 and 0.34 respectively. It should be noted, however, that a large portion of the cases for drugs and alcohol overlap. The main difference in the estimates from the two studies is a 25 per cent higher attributable fraction for drugs in the Pathway study.

According to the CSC study, cocaine had been used prior to the crime by 9 per cent of the inmates, cannabis by 3 per cent and heroin by 2 per cent. In addition, cocaine in combination with alcohol had been used prior to the crime by 5 per cent, cannabis with alcohol by 3 per cent and other drugs with alcohol by 5 per cent of the inmates. (Heroin had not been used with alcohol.) These figures will later be used in constructing four-model attributable fractions for cocaine, cannabis and heroin.

Among the individuals who were under the influence of any of these substances, the following proportions of inmates said that they would not have committed the crime if they had not been under the influence:

Intoxication crimes attributed to alcohol and drugs by the perpetrators

(Percentage)

	CSC-CLAI	*Pathway-CLAI*
Drugs	77	66
Alcohol	79	70
Drugs and alcohol	86	74

By multiplication, the part of the attributable fraction contributed by the intoxication model is obtained as shown by the following figures:

Corrected associative fractions from intoxication model

	CSC-CLAI	*Pathway-CLAI*
Drugs	0.13	0.14
Alcohol	0.16	0.13
Drugs and alcohol	0.11	0.10
No substance	0.60	0.63
Total	1.00	1.00

These figures pertain to the most serious crime on the inmates' current sentence. They will later be adjusted based on available information regarding all the crimes on the inmates' current sentence.

The comparison between the two sets of estimates gives an indication of the robustness of the estimates in the face of different data collection methods (computer-driven questionnaire in the CLAI versus personal interview situation in the Pathway study). The estimates from the two studies are fairly similar, considering that they are based on partly different regional populations and different time periods.

The economic model data

In response to questions on the role of alcohol and drugs as motivators for the most serious crime, such as "Was this crime committed to get or while trying to get alcohol/drugs for your own personal use?", the following proportions of inmates reported that such was the case:

Associative fractions from economic model

	CSC-CLAI	*Pathway-CLAI*
Drugs	0.12	0.15
Alcohol	0.03	0.03
Drugs and alcohol	0.06	0.05
No substance	0.79	0.77
Total	1.00	1.00

The data show that illicit drugs are greater motivators for crime than is alcohol, as expected: 12 per cent of the inmates in the CSC study and 15 per cent in the Pathway study stated that they had committed the most serious crime on the current sentence in order to get drugs for their personal use, while the percentage for alcohol was 3 per cent. Combined use as a cause of crime was evident also in this context, with 6 per cent and 5 per cent of inmates stating that the crime was committed in order to obtain both alcohol and drugs for personal use. Expressed as a total share, alcohol was at least a partial motivator in 9 per cent in the CSC study (3 per cent + 6 per cent) (compared with 8 per cent in the Pathway study) of the most serious crimes committed, while this was true for double that share (6 per cent + 12 per cent) (compared with 20 per cent in the Pathway study) in the case of illicit drugs. There is considerable agreement between the estimates from the two studies, which adds confidence to the reliability of the estimates.

With a starting point in the intoxication model, the importance of the economic factor will depend on the number of cases it adds to those already identified by the former model. The findings presented in table 2 indicate that the additional contribution of the economic factor is rather modest.

Table 2. Perpetrators who committed the crime in order to get alcohol or drugs according to whether they were on alcohol or drugs at the time of the crime (CSC-CLAI)

(Percentage)

Subjects	*On alcohol*	*On alcohol and drugs*	*On drugs*	*On neither*
To get alcohol	10.6	3.6	0.3	0.2
To get drugs	1.8	12.8	56.1	1.9
To get drugs and alcohol	6.3	25.4	9.1	0.7
To get neither	81.4	58.3	34.5	97.2
Total	100.0	100.0	100.0	100.0

Combining the two models, the corrected intoxication model and the economic model gives the following estimates from the two sets of CLAI data:

Associative fractions from combined intoxication-economic model

	CSC-CLAI	*Pathway-CLAI*
Drugs	0.13	0.17
Alcohol	0.15	0.13
Drugs and alcohol	0.14	0.12
No substance	0.58	0.58
Total	1.00	1.00

The estimates based on the CSC study do not differ greatly from the estimates from the corrected intoxication model: the fraction for alcohol is 0.01 lower and that for the combination of drugs and alcohol is 0.03 higher. The differences in the Pathway data are approximately of the same magnitude.

Calculations show that 93 per cent of cases in the combined model from the CSC study were already included in the corrected intoxication model. The number of drug cases rose by 7 per cent compared with the corrected intoxication model, while 4 per cent were added to alcohol cases and 1 per cent to the combined category of drugs and alcohol.

The contribution from the systemic (illegal economy) model to the attributable fraction

The information for this segment of the conceptual model was available in the calendar part of the Pathway study. These analyses were not available at the time of writing. Including cases from the systemic model will not change the attributable fraction for alcohol from the two-model estimate because almost all cases in the systemic model will be attributable to drugs. Neither is it likely that the attributable fraction for the combined category of drugs and alcohol will change to any notable degree from the two-model estimate.

Substance-defined crimes

In the last stage of the model combinations, drug offences will be added as attributable cases to the drug category, and drinking and driving infractions as cases to the alcohol category.

Internal consistency checks

Consistency and reliability checks were especially needed in material that was based entirely on self-reports.

Consistency check number one

One internal consistency check relates the attributable fraction status of the crime to the addiction status of the inmate based on two validated and widely used dependency scales: the Alcohol Dependence Scale (ADS) and the Drug Addiction Severity Test (DAST). If there is a strong positive relationship, that is, if the cases on the attributable fraction variable to a large extent overlap with the perpetrator being defined as an addict, this is an additional indicator of the construct validity of the attributable fraction and it will increase confidence in the measure. Consistency check number one is demonstrated in the following figures:

Consistency check between two-model attribution of crime to drugs and alcohol and addiction status of perpetrator (CSC-CLAI)

(Percentage)

	Crime attributable to			
Inmate addiction	*Drugs*	*Alcohol*	*Alcohol and drugs*	*No substance*
Drugs	76	9	65	13
Alcohol	5	35	42	5

The figures show that 5 per cent of inmates whose most serious crime was attributable to drugs according to the two-model estimate were addicted to alcohol, while 76 per cent were addicted to drugs. A high proportion of drug-addicted individuals was also found among those whose crime was attributed to the combination of alcohol and drugs. Similarly, 35 per cent of those whose crime was attributable to alcohol were addicted to alcohol and 9 per cent to drugs. Addiction to drugs and to alcohol was relatively high among those whose crime was attributed to both substances. Among those perpetrators whose crime was not attributable to any substance, only 5 per cent were addicted to alcohol and 13 per cent to drugs, both proportions well below the level of addicted inmates in the total population (12 per cent and 29 per cent).

Consistency check number two

The inmates were asked to rate the effect that alcohol and drugs had on their involvement in crime by answering the question "What do you feel has been the overall effect of your drug use (alcohol use) on your involvement in crime?". This has been used as another internal consistency check, as demonstrated in the following figures:

Consistency check between two-model attribution of crime and the perpetrators' assessment of the influence of drugs and alcohol on their criminality (data from CSC-CLAI)

(Percentage)

	Crime attributable to			
Perpetrators	*Drugs*	*Alcohol*	*Alcohol and drugs*	*No substance*
Who felt that drugs had increased involvement in crime	85	28	77	21
Who felt that alcohol had increased involvement in crime	17	76	75	10

More than four fifths of those offenders whose most serious crime was attributable to drugs felt that their drug use generally had increased their involvement in crime. In comparison, less than one fifth felt that alcohol had increased their involvement. The same substance-specific pattern was evident with regard to the inmates whose crimes were attributable to alcohol. The proportion of offenders with a crime attributable to drugs who felt that alcohol had increased their involvement in crime (17 per cent) was relatively close to the proportion of offenders whose crime was not attributable to any substance and who felt that alcohol had increased their involvement in crime. The proportion of offenders with a crime attributable to alcohol (28 per cent) was also relatively close to the proportion of offenders whose crime was not attributable to any substance (22 per cent).

A comparison will be made later between two-, three- and four-model estimates as to how well they discriminate on these and other consistency checks. The aim is to obtain some insight into what was gained in accuracy between estimates of attributable fractions from different combinations of models, and whether it is justified to assign the crime or the individual as a case on the attributable fraction variable.

Another indication of the reliability of the measure was mentioned above. It is the extent of overlap in the cases that the different models bring to the attributable fraction. As shown, the cases included on the basis of the economic model greatly overlapped with the cases obtained from the (corrected) intoxication model. On the other hand, 76 per cent of the intoxication cases for the combined category of alcohol and drugs could not be found in the economic model, mainly because there were many more cases from the intoxication model than from the economic model.

Questioning the assumptions

Assessments about being under the influence of drugs or alcohol

Self-reports on the level of intoxication have been used in numerous surveys on alcohol use in general populations around the world. On the whole, the experience is that the data obtained are sufficiently valid. Self-reports on drinking among alcohol abusers are considered to be even more valid. Emergency-room studies also indicate that self-reports yield valid data on alcohol use in connection with the injury, whether from violence or accident. There is also strong evidence that drug-use surveys in general populations of adults, high school students etc. on the whole provide valid data on drug use.

A fair number of studies have been conducted on the validity of self-reports on substance use among drug users and abusers. Out of 54 reports examined by the present authors, 48 assessed that the self-reports were reasonably valid if certain conditions were met in the execution of such studies.

Many studies on violent crimes have used police or court records to ascertain whether the perpetrator and victim had been drinking at the time of the crime. Results from such studies are used to provide attributable fractions in calculating social costs of alcohol. However, there is often no definition of "being under the influence" in such studies, and circumstantial information (such as the crime having occurred inside or outside a bar) is sometimes used to classify a crime as being alcohol-related. In addition, information on alcohol and drug use in police and court records is to a large extent provided by individuals involved in the crime. It is therefore doubtful if studies using official records of crime events provide estimates that are appreciably more valid than data based on self-reports by inmates.

Whether the crime would have been committed in the absence of alcohol or drugs

The assumption that the inmate is in a position to assess whether he would have committed the crime had he not been under the influence of alcohol or drugs can be seriously questioned. As discussed in the CSC and Pathways studies, 77 per cent and 66 per cent of those under the influence of drugs, 79 per cent and 70 per cent of those intoxicated from alcohol and 86 per cent and 74 per cent of those intoxicated from both substances stated that they would not have committed their most serious crime had they not been intoxicated. The correction decreases the sizes of the attributable fractions by 14 per cent to 23 per cent in the CSC study, and 26 per cent to 34 per cent in the Pathway study. Despite severe doubts regarding the validity of such judgements, such a correction seems to provide an improvement compared with accepting the intoxication model without any downward adjustments.

The validity of judgements about one's own motivations for committing a crime or crimes

In this case, to get drugs or alcohol for personal use is perhaps easier to assess for the actor than the counterfactual scenario of the previous point.

Using two different methods of data collection

Another indication on the aggregate validity of the responses to the questions asked can be had from a comparison of the CLAI findings and the subset of identical questions that were included in the Pathway study. Such a comparison also, to some extent, serves as a check of the assumptions above, in the sense that it provides an indication as to the robustness of the estimates in the face of varying situational settings: an interview situation with a female research assistant from outside the prison setting versus responding to a computerized questionnaire by means of punches on a keyboard. It has been found that, although the figures from the two studies are not identical, they fall within a relatively narrow range of estimates.

Discussion

One important conceptual question concerns what to do with the combined category of crime associated with both alcohol and drugs. It is important to ask whether crimes attributable to this combination can be divided up between attributable fractions for alcohol and fractions for drugs. It may not make sense to ask an inmate or arrested person which of the substances was the most important in causing the crime. To some extent, the perpetrator's history of abuse, treatment and societal reactions may provide sufficiently valid information for assigning a prime role to either of the two substances.

Is it, however, necessary to separate the role of drugs and alcohol in the cases where both had been involved in causing the crime? It is a fact that they exist in the crime episodes simultaneously, either as pharmacologically causal, motivating or systemic factors. Perhaps it is time to take this into account in assigning attributable fractions. However, the advisability of using a combined attributable fraction will naturally depend on what purpose it is used for, and the framework of the costing process in particular.

Another question is whether something is being missed by using the four-component method by comparison, for instance, with the analyses of aggregated data that have been used to arrive at attributable fractions for alcohol on crime (almost exclusively violent crime). Different types of time-series analyses are potentially the most powerful methods available for aggregate-level estimates.

In this context, it is worth remembering that it is not at all possible to conduct time-series analyses for illicit drugs. Some type of individual-level data must be used. Information must be collected from individuals who participated in the crime episodes. Observers or informants can be used in some cases, but the most relevant and detailed data can only be provided by the actors themselves, through self-reports. An interesting question is how comparable attributable fractions for alcohol from, for instance, time-series analyses are to attributable fractions based on self-reports of the role of alcohol as a determinant in individual crime events. Such comparisons are possible for a number of countries or other jurisdictions, but for alcohol only.

It is possible that something further may be missed in the four-model method. There may be countries and cultures where additional models should be taken into account, in addition to the intoxication model, the economic and the systemic model and the substance-defined crimes.

There may be simpler ways of collecting data for the method of estimation used in the present article. The study from Canada presented here has six components. Conducting such studies of inmates in prisons and individuals who are arrested requires a sometimes lengthy process for obtaining access to relevant samples. Having several components also requires added effort and supervision, compared with conducting one larger-scale study. There may be a simpler way of getting the same information or sufficient information of some other kind that would make it possible to calculate reasonably valid attributable fractions.

The final aim of the project is to provide attributable fractions for alcohol, drugs, cocaine, cannabis and heroin on crime in Canada. However, for the purpose of calculating the social costs of keeping inmates in federal penitentiaries (a considerable sum), using attributable fractions for federal inmates is probably more accurate than using an overall fraction for all crimes in Canada. In the same way, estimates from the study of individuals arrested by the police may be more (but not exclusively) relevant for policing costs.

Validity issues in general population studies are in some respects identical to studies with inmates. Under-reporting may be more likely in general population studies because there is no incentive to tell the truth, in particular about drug use and other illegal activities. In some cases, such an incentive does exist for inmates if they want to be free of their drug problem or to spend their prison time in a treatment environment which, in many cases, is more pleasant than a standard prison setting. This of course opens up possibilities for over-reporting the role of alcohol and drugs.

The method outlined above provides easily accessible information on the share of crime contributed by different causal processes and their combinations. From the perspective of prevention, it will also be important to know which type of causal process predominates, what the overlaps between the determinant models are and what changes occur over time in these constellations.

Finally, using attributable fractions that are built up from individual-level causal models regarding the effects of alcohol and drug use on crime may bring the study and estimation of attributable fractions closer to the main body of research on drugs, alcohol and crime.

Bibliography

Adrian, M. Social costs of alcohol. *Canadian journal of public health* (Ottawa) 79: 316-322, 1988.

Drugs, alcohol and crime: patterns among Canadian federal inmates. *By* S. Brochu *and others*. *Bulletin on narcotics* 51:1-2: 57-73, 1999 (United Nations publication).

Goldstein, P. J. The drugs/violence nexus: a tripartite conceptual framework. *Journal of drug issues* 14:493-506, 1985.

International guidelines for estimating the costs of substance abuse. *By* E. Single *and others*. Ottawa, Canadian Centre on Substance Abuse, 1996.

Stinson, F. S., and S. F. De Bakey. Alcohol-related mortality in the United States, 1979-1988. *British journal of addiction*, 87:5:777-783, 1992.

Estimating the costs of substance abuse to state budgets in the United States of America

S. E. FOSTER

Vice President and Director of Policy Research and Analysis, National Center on Addiction and Substance Abuse at Columbia University, New York, United States of America

D. MODI

Research Associate, National Center on Addiction and Substance Abuse at Columbia University, New York, United States of America

ABSTRACT

The National Center on Addiction and Substance Abuse at Columbia University is conducting a study on the impact of substance abuse on the budgets of the governments of the 50 states in the United States of America, the District of Columbia and Puerto Rico. The goals of the study are to provide policy makers with a map of how substance abuse affects the costs of state governments; to document the total bill for substance abuse that states pay, itemizing expenditure on prevention, treatment, research and consequences; and to point the way to more cost-effective investments.

The present article documents the methodology used in the study by the National Center on Addiction and Substance Abuse to estimate state costs linked to substance abuse. The methodology relies on previously documented costs of illness research to estimate the costs imposed by substance abuse on the health-care system and on the state workforce. It also documents costs that are directly attributable, namely, prevention, treatment, research and regulation. Finally, the methodology establishes the pool of substance-involved individuals for whom prevention and treatment may be a necessary condition of reducing public spending. The study demonstrates that states devote tremendous resources to managing the effects of substance abuse, while only a small portion of state spending is aimed at reducing substance abuse through treatment and prevention programmes. By providing a map of state spending on substance abuse, the study has established a base against which policy makers can begin to compare the value of alternative policies concerning prevention, treatment, regulation and tax that might reduce the consequences of substance abuse and addiction. Understanding the enormous costs attached to current policy choices with regard to substance abuse should help policy makers think more strategically about how they might invest in interventions that would yield a better return.

Introduction

In the United States of America, spending on efforts to deal with substance abuse has, historically, been a minor blip on the radar of state budgeting. As the links between substance abuse and crime, child abuse and neglect, traffic accidents, disease and disability have been uncovered, state policy makers have begun to understand that substance abuse and addiction and their associated costs cut far deeper into state budgets than previously imagined.

To document just how large the toll is, the National Center on Addiction and Substance Abuse at Columbia University (CASA) is conducting a study of the impact of substance abuse on the budgets of the governments of the 50 states in the United States, the District of Columbia and Puerto Rico. The goals of the study are to provide policy makers with a spending map of the many ways that substance abuse affects the costs of state governments; to document the total bill for substance abuse that states pay, itemizing expenditures on prevention, treatment, research and consequences; and to point the way to more cost-effective state investments.

Better investments in the area of substance abuse are essential for three reasons. Firstly, current government spending on efforts to deal with substance abuse is bad public policy because it does not attempt to address the problem of untreated substance abuse and addiction, paying extraordinary costs for preventable consequences. Secondly, current thinking about government spending as inevitable annual outlay rather than as investment in better results neither encourages nor demands the development of more cost-effective prevention, treatment or intervention. Finally, the bad public policy of not addressing the problem of substance abuse and addiction is fast becoming bad politics because it involves profligate spending with no accountability.

The aim of the present article is to document the methodology used by CASA to estimate state costs linked to substance abuse and addiction for the purposes of inviting both domestic and international comment and refining the approach.

Previous attempts to document costs of substance abuse primarily have taken the form of cost-of-illness studies that estimate the overall economic costs to society of the abuse of drugs, alcohol and tobacco [1, 2]. Those studies have been compelling but they have not provided comprehensive estimates of costs to government. Other approaches have estimated the costs of substance abuse to selected government programmes such as health care [3, 4], federal entitlement programmes [5], prisons and jails [6] and child welfare [7]. Those estimates have been of value to states but their narrow focus has not provided policy makers with aggregate spending across budget categories. Both approaches have failed to capture the attention of a critical mass of state policy makers or persuade them of the economic value of allocating resources in a different manner.

Methodology

To gather data for the study, CASA administered a survey in September 1998 in the 50 states of the United States, the District of Columbia and Puerto Rico. The survey was completed by the authorities of 45 states, the District of Columbia and Puerto Rico.* The participating jurisdictions account for approximately 90 per cent of budget spending at the state level in the country.

*The states that did not participate in the survey were Indiana, Maine, New Hampshire, North Carolina and Texas.

To determine which state programmes to include in the study, CASA undertook the following:

(a) A wide range of literature on the consequences of substance abuse for government programmes was reviewed;

(b) State programmes designed to prevent or treat substance abuse or to deal with the consequences of substance abuse were identified. In the latter category, only those programmes were included that were large enough to be of any consequence in terms of affecting the total amount spent on efforts to deal with substance abuse;

(c) State budget and programme officials were consulted to understand how those programmes were financed and to determine the most efficient and effective way to gather the data on spending;

(d) Between March 1998 and August 1998, interviews were conducted in California, Florida, Minnesota, New Jersey and Vermont* to ascertain which government programmes were affected by substance abuse and to learn what, if anything, had already been done to record the costs to states of substance abuse. Some states, such as California, had already done extensive studies on the subject and had even tallied up direct state expenditures on the prevention and treatment of substance abuse [8].

CASA selected state budget officers as the appropriate target for data collection because those officers had the broadest view of and most expertise in the budget-related issues and because CASA was particularly interested in educating budget officers about the extent to which substance abuse affected the budget. A questionnaire was designed, dividing functions into broad sections, consistent with the way that most budget offices were organized. The programmes for which data were needed were grouped into nine clusters: human and social services; developmental disabilities and mental health; health; education; correctional services; public safety; judiciary; regulation and compliance; and capital spending. The instrument was designed in that way in order to make it easier for the budget office to distribute the survey questions among a variety of specialists in the budget office, requesting a manageable amount of data from each individual.

In order to capture as much of the spending associated with a particular programme as possible, CASA designed a survey instrument** to obtain data on:

(a) Fiscal year 1998 for each state, the state's own sources of general revenue, including spending of general funds and non-general funds but not federal or local funds;***

*Two former state budget directors were contracted to help choose states and set up interviews. The study team, consisting of staff and consultants, conducted over 40 interviews with state officials and their staff. The interviews were designed to identify ways to develop a cost base that was both complete and consistent with the way in which programmes were organized and administered in different states.

**The survey instrument was developed for and with CASA by the Fiscal Studies Program, a research unit within the Rockefeller Institute of Government (in Albany, New York). The Fiscal Studies Program was also responsible for collecting the budget data and conducting specific analyses.

***General funds refer to predominant funds for financing a state's operations. Revenues are received from broad-based state taxes; however, how specific functions are financed differs from state to state. Non-general funds refer to other state funds (expenditures from revenue sources that are restricted by law for particular governmental functions or activities), for example, a tax dedicated to a particular trust fund; and bonds (expenditures from the sale of bonds, generally for capital projects).

(b) Reported expenditures (not appropriations) from the executive budget presented in the winter or spring of 1998, since some states did not publish data on the adopted budget (differences between the proposed and adopted budgets were not expected to be large enough to skew the findings);

(c) All costs (programme administration, fringe benefits, service providers and capital).

To refine the programme categories, clarify instructions and get a sense of the kinds of questions state budget officers would have, CASA pre-tested the questionnaire in California, Florida and New York.

Linking expenditure to substance abuse

The data represented a combination of costs attributable to substance abuse and costs associated with substance abuse. Costs attributed directly to substance abuse and addiction fell into three main categories: *(a)* spending on prevention, treatment and research; *(b)* spending on the consequences, including health-care spending based on the probable causal link between substance abuse and addiction and a particular disease state, and spending on state worker absenteeism linked to substance abuse; and *(c)* expenditures for alcohol and tobacco regulation. For those categories, it was either self-evident that costs were attributable (prevention, treatment, research, regulation and compliance) or it was important to establish a causal link (health care and state workforce).

For other areas of spending, whether substance abuse caused the spending was of less concern than whether treatment or intervention would reduce the cost of the consequences associated with the problem. It is, however, a very important policy distinction. The cost-of-illness model has focused on increasing the precision of linking costs to causality. The operational question for a policy maker is not how many welfare recipients are receiving assistance because of their substance abuse, but how many welfare recipients will be impeded in their efforts to cease being on welfare and return to work because they abuse alcohol or drugs. Similarly, it is less important to establish the percentage of state inmates who committed crimes because of substance abuse than to determine the group of prisoners for whom treatment for substance abuse may be necessary to keep them from returning to prison.

It is necessary to establish the pool of substance-involved individuals, in other words, those who constitute the target group for policy intervention. Subsequent work should focus on matching the different forms of intervention with the different needs of individuals within that pool. Substance abuse often appears as one of a cluster of behaviours leading to increased costs to states; thus, solving the addiction problem is a necessary step towards eliminating those costs, but it will not always be sufficient.

Estimating the share of state spending on efforts to deal with substance abuse

CASA developed estimates of the share of spending for each programme designed to deal with problems that reasonably could be attributed to or associated with substance abuse by relying on an extensive review of the literature on the subject, including research conducted by CASA itself. Where possible, those estimates were based on peer-reviewed research. Where published research was lacking, CASA developed estimates and clearly documented the techniques used. Although using such a range of techniques lacked methodological purity, it provided national estimates where none were currently available.

The first step was to identify spending on programmes that were devoted to dealing with problems attributable entirely (100 per cent) to substance abuse. For the remaining programmes, the shares were adjusted for state differences in the prevalence of substance abuse.*

To obtain a national estimate for state spending on efforts to deal with substance abuse, CASA calculated average per capita substance abuse spending in each programme area for the total of the 47 responding jurisdictions. It multiplied those averages by the population of the non-responding states to estimate their substance abuse spending. Estimated spending for both responding and non-responding jurisdictions were summed to estimate spending levels for the nation as a whole. Although 47 jurisdictions participated in the project, in some instances they did not complete certain sections of the survey. In those cases, CASA also estimated aggregate spending for the non-responding states in all categories except research because of the low response rate in that section of the survey (only five responses).

Substance abuse prevention, treatment and research

CASA asked states to report all spending for programmes aimed at reducing alcohol, drug and tobacco use and abuse and for programmes providing treatment for tobacco use and for alcohol and illicit drug abuse. Examples included statewide media campaigns, local prevention networks, inter-agency coordination of prevention programmes, prevention education, treatment facilities, outpatient care programmes, research and capital spending for treatment facilities. All such programmes were devoted to dealing with problems attributable entirely to substance abuse.

Health care

Health-care spending was the second largest component of state budgets, after primary and secondary education. States spent about $60.4 billion (about 9.7 per cent of total state expenditures) of their own funds to finance health care under the Medicaid programme, the federal and state programme of health insurance for the poor and medically needy [10]. Some states also financed health-care costs for people not qualifying for Medicaid.**

Substance abuse increased state spending on health care in three ways:

(a) Some people would become ill or injured as a result of their own substance abuse and would receive health care related to the illness. For example, lung cancer resulting from smoking led to a variety of health-care expenditures, such as hospital, physician and medication costs;

*Costs in those categories were linked primarily to alcohol and illicit drugs. Because the sample size was not sufficient to estimate the prevalence of illicit drug abuse by state and because most research showed the primary drug of abuse to be alcohol with a high prevalence of alcohol abuse occurring in conjunction with illicit drug use, the associated fractions were scaled using alcohol prevalence of binge drinking in the state in the past month [9].

**Examples of other state-funded health-care programmes include programmes for patients infected with the human immunodeficiency virus (HIV) who receive general assistance, child health-care programmes and prenatal care programmes.

(b) Substance abuse injured innocent parties. Mothers who smoked during pregnancy would generally have lower birthweight babies, thus increasing state-financed costs from the time of the child's birth (and possibly increasing state-financed health expenditure throughout the child's life);

(c) People who smoked or abused alcohol or drugs often would have a lower level of health in general and would have more frequent, longer and more severe illnesses. For example, bouts of influenza tended to last longer for smokers than for non-smokers. Because the available data were inadequate, the CASA analysis did not include such costs.

CASA calculated the health-care portion of the analysis without taking into account the cost data from the state survey for two reasons. The basis for the estimates of health-related spending was epidemiological research showing a link between substance abuse and illness. Some states might have explored such literature in greater depth than others and some might have interpreted the research differently. It made sense to interpret and apply the epidemiological research uniformly across states. In addition, many states did not have ready access to data that described the illnesses of or the health care received by their residents on Medicaid and that were related specifically to substance abuse or that identified substance abusers; that fact made it difficult for patterns of Medicaid utilization to be analysed. Although in some cases such data existed, they were located in massive databases that included confidential patient information; thus, it would be difficult for states to use the databases for research purposes.

Although obtaining data on state-level spending by type of illness would have been ideal, it was not possible. Instead, state-level data on spending by type of provider were obtained. CASA devised the following two-step methodology to estimate the share of state health expenditures accounted for by efforts to deal with substance abuse, taking advantage of as much state-specific data as possible:

(a) National-level attributable fractions by substance and provider type were estimated. An attributable fraction is an estimate of the share of spending in a given programme that is caused by smoking, alcohol or drug abuse. For example, if 12 per cent is the "smoking-attributable fraction" for Medicaid-financed physicians' services, it means that, on average, about 12 per cent of Medicaid payments to physicians is attributable to smoking;

(b) Those attributable fractions were multiplied by state-specific health spending to arrive at state estimates of spending attributable to substance abuse.

Step one: national attributable fractions

National-level attributable fractions were developed for each major form of substance abuse (smoking, alcohol and drugs), for each major type of medical provider (such as hospitals, physicians and home providers). A total of 24 different attributable fractions were developed: three substance types by eight provider types.*

To estimate attributable fractions, population-attributable risk (PAR) values were used, either estimated directly or as reported in epidemiological research. A

*The provider types refer to hospital overnight, emergency room, outpatient, medical-provider visit, home-provider visit, medical supply purchase, prescription drugs and dental.

PAR value is an estimate of the probability that a given episode of disease is attributable to (or caused by) a factor such as substance abuse. It reflects both the relative risk of getting the disease and the prevalence of substance abuse. For example, if 19 per cent is the alcohol-related PAR value for liver cancer, it means that 19 per cent of the incidence of liver cancer is the result of alcohol abuse.

For alcohol, the PAR values used were developed by the National Institute on Alcohol Abuse and Alcoholism of the United States for specific diagnosis (malignant neoplasm of the esophagus, alcohol-related injury involving motor vehicle traffic etc.). For illicit drugs, CASA developed its own PAR values based on a thorough review of the epidemiological research. In the case of smoking, CASA used state-specific smoking attributable fraction developed by Miller and others [11].

Those PAR values were applied to the latest available public-use medical care databases, relying on the coding system of the *International Classification of Disease* [12]. The database of the 1987 National Medical Care Expenditure Survey, conducted by the National Center for Health Services Research, was used for hospital inpatients, hospital outpatients, physician services, prescription drugs and miscellaneous services. It was assumed that nursing home expenditures would have the same attributable fractions as that of elderly hospital patients. The process undertaken to calculate hospital spending attributable to alcohol was as follows:

(a) The amount of hospital spending financed by Medicaid for each patient with a given disease was multiplied by the PAR value of that disease. The number of all patients with that disease was calculated in order to develop an estimate of the hospital spending financed by Medicaid for that disease that was attributable to alcohol abuse. The process was repeated for each illness represented in the data file;

(b) The total amount of hospital spending financed by Medicaid that was attributable to alcohol across all diseases was calculated in order to estimate the total hospital spending by Medicaid that was attributable to alcohol;

(c) The amount of hospital spending financed by Medicaid that was attributable to alcohol was divided by the total hospital spending financed by Medicaid to arrive at the “alcohol-attributable fraction”: the share of hospital spending financed by Medicaid that was attributable to alcohol abuse;

(d) The process was repeated for other provider types and other substance types. The result was an attributable fraction for each type of provider and each type of substance.

Step two: applying attributable fractions to state health spending

To develop state-by-state estimates of Medicaid and other health spending attributable to substance abuse, CASA multiplied the attributable fractions by provider type (derived in step one) by the 1998 spending by provider type obtained from the Health Care Financing Administration. Several states did not provide a sufficiently detailed breakdown of spending by provider type. For those states, CASA used a two-step process. First, it calculated average attributable fractions by substance type using state-specific data for 1997 from the Health Care Financing Administration, effectively weighting the national provider-type attributable fractions by the state’s spending by provider type. CASA then multiplied those state-specific weighted-average attributable fractions by 1998 total state spending on health programmes to arrive at substance abuse attributable spending.

State workforce

Several studies have focused on documenting and quantifying the adverse effects of alcohol, tobacco and illicit drug use on the workforce [13-15]. Some have been studies of one organization only, others of large firms and still others of particular regions, making the comparison of results difficult. A further complicating factor is the variation in definitions of the quantity and frequency of substance use.

Drug and alcohol use has been associated with employee absenteeism, lower productivity, increased turnover, workplace accidents and higher costs of health insurance [14]. Because of severe data limitations, the CASA study has focused only on analysing the absenteeism of substance abusers (comparing it with that of persons not abusing substances); by sex and substance type.*

CASA conducted a logistical regression using data from the National Household Survey of Drug Abuse (for 1994) and the National Longitudinal Survey of Youth (for the periods 1984-1988 and 1992-1994). The National Longitudinal Survey of Youth allowed CASA to control a large number of relevant demographic and socio-economic variables and to focus on absenteeism. That methodology was used to pinpoint a probable causal relationship between employee substance abuse and absenteeism. The sample was restricted to individuals who had paying jobs in the private sector, who had occupations in any area except the military, farming, fishing or forestry and who worked in any industry but not in the military, public administration or agriculture. From that analysis, CASA identified prevalence rates and extra days absent as a result of substance abuse, for men and women, by substance type.**

Following that, the prevalence of substance abuse (by gender and type of substance) was multiplied by the number of persons in the state's workforce (broken down by gender) to obtain the estimated number of substance abusers in the workforce (broken down by gender and type of substance). Those subtotals were multiplied by gender-specific and substance-specific extra days of absences per person per year to obtain the total number of days lost per year. Once those subtotals were combined, that number was divided by the expected number of days of work per year (the number of persons in the workforce multiplied by 230 days) to obtain a proportion spent on substance abuse.

In the state workforce section of the survey, CASA requested payroll figures for state government employees, total spending on fringe benefits and the share of spending on efforts to deal with substance abuse in employee assistance programmes. The share of spending attributable to substance abuse, adjusted for state differences in prevalence of binge-drinking and illicit drug use, was applied to the payroll and fringe benefits. That total was added to 100 per cent of the share of spending on substance abuse in employee assistance programmes to obtain the figure for total spending attributable to substance abuse in the state workforce sector.

*CASA adopted the methodology in its study entitled "Substance Abuse and American Business" (publication forthcoming).

**The following definitions were used:

(*a*) Smoker: an employee who smoked more than 16 cigarettes per day in the previous month;

(*b*) Heavy drinker: a male employee who drank more than 5 drinks, five or more times in the previous month or a female employee who drank more than 3 drinks, five or more times in the previous month;

(*c*) Current drug user: an employee who used cannabis and/or cocaine in the previous month;

(*d*) Absent: an indicator for worker absence at any time during the previous month (National Household Survey of Drug Abuse) or week (National Longitudinal Survey of Youth).

Criminal justice

For most states, spending on prisons has been the fastest growing part of the budget, jumping by 28 per cent between 1995 and 1996 [6]. In its report on the adult prison population, CASA documented the enormous impact of substance abuse on state spending for correctional services [6]. To estimate the percentage of the inmate population involved in substance abuse, CASA used the following categories for such inmates:

(a) Ever used illegal drugs regularly;

(b) Convicted of a drug law violation;

(c) Convicted of an offence involving driving under the influence;

(d) Under the influence of drugs and/or alcohol while committing the crime that led to incarceration;

(e) Committed an offence to get money for drugs;

(f) Had a history of alcohol abuse (defined as ever having received treatment for alcohol abuse).

Using such a definition, it was estimated that 81 per cent of state prison inmates were involved in substance use.

To arrive at total state costs of adult correctional services associated with substance abuse, CASA combined state expenditures for such services in the following areas:

(a) Costs of running and maintaining adult correctional facilities, associated administrative and staffing costs;

(b) Costs of special programmes such as mental health, education or religious services provided to adult inmates;

(c) Parole and early-release programmes;

(d) Adult probation;

(e) State categorical aid to localities for adult correctional services;

(f) Capital spending on prisons.

CASA applied the 81 per cent share, adjusted for state differences in prevalence of alcohol and illicit drug use, to the costs of the state offender rehabilitation programmes and added 100 per cent of the costs of treatment programmes for alcohol and drug abusers provided by state correctional departments. It was assumed that a similar share (81 per cent) of adult probationers and parolees were involved in substance abuse and that local spending of state aid for correctional services would follow the same pattern.

Juvenile justice

Data on the prevalence of drug and alcohol abuse in the juvenile justice system were found to be scarce.* In one study, it was shown that 70 per cent of juvenile offenders had a serious alcohol or illicit drug problem [16]. In a more recent study

*CASA is conducting a study of substance abuse and the juvenile justice system.

in New Jersey, 67 per cent of male juvenile offenders reported having used cannabis in the previous 30 days while 57 per cent of such offenders reported having used alcohol in the previous 30 days [17]. In a study in Maricopa, Arizona, 56 per cent of male juvenile offenders tested positive for drugs [18].

In the absence of recent national estimates of the extent of substance abuse in the juvenile justice system, CASA conducted an analysis of data for 1997 gathered through the Arrestee Drug Abuse Monitoring Program by the National Institute of Justice of the United States Department of Justice. Variables were chosen to mirror those in the CASA report on the adult prison population [6]. Juveniles* in correctional facilities who had been involved in drugs were categorized as follows:

(a) Tested positive for drugs;

(b) Reported having used alcohol in the previous 72 hours;

(c) Under the influence of or in need of alcohol or drugs;

(d) Received treatment in the past;

(e) Currently receiving or could use treatment for abuse of alcohol or various illicit drugs. It was found that 66.3 per cent of youth in the juvenile justice system were involved in substance abuse.

To arrive at total state costs for juvenile justice attributable to substance abuse, CASA combined state expenditures in the following areas:

(a) Juvenile correctional facilities, including residential centres, boot camps and work-study camps;

(b) Diversion programmes;

(c) Capital costs of juvenile correctional facilities.

CASA applied the 66.3 per cent share, adjusted for state differences in prevalence of alcohol and illicit drug use, to the costs of such juvenile justice and added 100 per cent of the costs of treatment programmes for alcohol and drug abusers provided by state departments of juvenile justice.

Judiciary

The judiciary system has several branches: criminal, juvenile, family and civil and drug courts (which may be further differentiated into family drug court or juvenile drug court). CASA did not identify any studies documenting the full impact of substance abuse on courts, although several studies documented the prevalence and characteristics of drug law offenders (drug possession and trafficking) in both juvenile and adult courts [19, 20]. To develop a more comprehensive picture of the impact of substance abuse on the courts, CASA employed the methodology described below:

Criminal courts

CASA analysed the involvement of arrestees in substance abuse, using 1997 data from the Arrestee Drug Abuse Monitoring Program, to estimate the proportion

*All the juveniles in the sample were male.

of substance abusers entering the judiciary system. The following categories were used for those involved in substance abuse:

(a) Tested positive for drugs;

(b) Reported having used alcohol in the previous 72 hours;

(c) Under the influence of, or in need of, alcohol or drugs;

(d) Received treatment in the past, currently receiving treatment or could use treatment for abuse of alcohol or illicit drugs.

It was estimated that 83.8 per cent of criminal court costs were substance-linked.

Family courts

Previous CASA research had shown that 70 per cent of child welfare cases were involved in substance abuse [7]; that is, in those cases, the problem had been caused or exacerbated by substance abuse and addiction. In some states, juvenile justice cases may be represented in this category as well. Seventy per cent of those costs were assumed to be linked to substance abuse.

Civil courts

No share for spending on substance abuse was developed for civil courts because it was not possible to link costs of tort, property rights, and estate or small claim cases to substance abuse and addiction.

Drug courts

Any state spending on drug courts, including family dependency drug courts, was considered 100 per cent attributable to substance abuse.

In order to estimate costs relating to substance abuse that were linked to the courts, state authorities were asked to identify all court costs for state programmes by type of court, including court personnel, contracted services, supplies and the cost of state programme administrators and policy analysts who spent the majority of their time on the programme. The share of spending attributable to substance abuse, adjusted for state differences in prevalence of alcohol and illicit drug use, were applied to the total spending by type of court. The values of substance-linked spending by type of court were added together to obtain a total for courts.

Child welfare programmes

The link between substance abuse and child neglect and abuse is well documented. CASA research found that substance abuse and addiction cause or exacerbate 70 per cent of child welfare cases in the United States. In other studies, the rate of substance abuse among parents of children in child protective services was found to be between 40 and 90 per cent [22-25]. For the present study, 70 per cent was used as the share of child welfare spending attributable to substance abuse.

To determine spending on child welfare, states were asked to identify all programme costs, including grants to individuals and families, the cost of caseworkers or service providers and other programme costs. They were also asked to include the following costs: adoption assistance; foster care; independent living; family preservation and other programmes to prevent out-of-home placements, promote reunification of families or provide a safe environment for children; child abuse and neglect intake and assessment; and administrative and staffing costs to run such programmes.

The 70 per cent share of spending attributable to substance abuse, adjusted for state differences in alcohol prevalence, was added to total state spending on child welfare, after the cost of any child welfare programmes specifically targeting substance abuse was removed. Both categories of costs were combined to obtain the total cost of substance abuse to the child welfare system.

Income support programmes

Substance abuse may be the primary reason that people need income assistance or it may impede a person's ability to become self-supporting. The income support programmes included in the present study were Temporary Assistance to Needy Families, General Assistance and state supplements to the Supplemental Security Income programme.

Temporary Assistance to Needy Families and General Assistance

In the majority of national and state prevalence studies, it was estimated that between 7 and 37 per cent of welfare recipients had a substance abuse problem [26]. Two previous studies by CASA estimated the prevalence of female recipients of Aid to Families with Dependent Children with substance abuse problems to be between 20 and 27 per cent [5, 27]. For the purposes of the present study, a more conservative figure of 20 per cent was used as the share of spending attributable to substance abuse for recipients of Temporary Assistance to Needy Families.

There were limited data available on the percentage of the population receiving General Assistance that is involved in substance abuse. In one study of a county in California, it was estimated that at least 43.3 per cent of the population receiving General Assistance had a substance abuse problem that was linked to the receipt of assistance [28]. In the absence of national data, CASA used 20 per cent as the proportion of substance-linked spending on the Temporary Assistance to Needy Families programme, recognizing that it was probably a very conservative estimate.

Supplemental Security Income

Federal legislation passed in 1996 ended payments to individuals who were receiving Supplemental Security Income because of drug addiction and alcoholism. When benefits were terminated on 1 January 1997, 2.6 per cent of all beneficiaries were removed from the rolls. About one third (34 per cent) of those people retained or re-established eligibility in December 1997 on the basis of a condition other than substance abuse [29]. Therefore, only 1 per cent of people receiving Supplemental Security Income were originally certified by virtue of drug or alcohol addiction. CASA could find no studies documenting the extent to which individuals qualifying for Supplemental Security Income under another condition also had drug and alcohol problems and, if so, what percentage of them might be capable of supporting themselves if their addiction problems were addressed. For that reason, for Supplemental Security Income, 1 per cent was used as the associated share of spending.

To estimate costs linked to substance abuse for those programmes, states were asked to identify costs for cash assistance, emergency assistance, employment and training services for the Temporary Assistance to Needy Families or for those receiving General Assistance, income maintenance to the aged, blind and disabled, and administrative costs to run those programmes. Shares of spending associated with substance abuse, adjusted for state differences in alcohol prevalence, were combined with total costs in each area to obtain a figure for total spending for income support programmes.

Mental health

According to data from a nationally representative population sample of non-institutionalized civilians in the United States, about one half (50.9 per cent) of those with a lifetime mental disorder also have a lifetime addictive disorder, otherwise described as alcohol and illicit drug abuse and dependence [30]. That may be a conservative estimate of the occurrence of a co-morbid addictive disorder in the population receiving mental health treatment through the state, since the institutionalized population was not surveyed and people with more severe mental health problems often receive residential care.

Mental health costs included in the present study refer to expenditures for administration, community contracts, housing programmes, institutionalization and capital costs for building and maintaining facilities. The share of spending associated with substance abuse was 50.9 per cent, which was applied to the total of those costs, after adjusting for differences in state prevalence of alcohol use.

Developmental disabilities

To estimate the share of state costs for the developmentally disabled that was caused or exacerbated by tobacco, alcohol or drugs, CASA used as a basis *The Economic Costs of Alcohol and Drug Abuse in the United States, 1992* [1], in which the number of individuals with fetal alcohol syndrome who were receiving care in 1992 was estimated at 38,884,* or approximately 9 per cent of the total developmentally disabled population of 434,657 who were served in 1992 in institutional and residential care in the United States [31]. While CASA considered the figure of 9 per cent to be conservative, since it was based solely on fetal alcohol syndrome, that figure was used to calculate the share of state spending for the developmentally disabled that was attributable to substance abuse. That share, adjusted for state differences in prevalence of alcohol and illicit drug use, was applied to total state costs for developmental disabilities, such as administration, community contracts, housing programmes, institutionalization and capital cost to build and maintain facilities, to develop state totals of associated costs.

Public safety

There were limited data available for estimating costs to the state for public safety other than for criminal and juvenile justice and courts. CASA asked states to report costs for special drug law enforcement programmes, highway safety and accident prevention programmes, state highway patrol and local law enforcement programmes.

*That figure includes moderately retarded persons with fetal alcohol syndrome aged 22-65 in the developmentally disabled systems and severely retarded persons with fetal alcohol syndrome aged 5-65 in those systems.

The main area where some data were available was highway safety, that is, the proportion of motor vehicle accidents that involved alcohol. There were no data available on the number of drug-related motor vehicle accidents. Using data collected by the National Highway Traffic Safety Administration [32], CASA made an estimate of the proportion of reported motor vehicle accidents involving alcohol, based on the following:*

(a) A calculation of the number of motor vehicle accidents where alcohol was involved for each type of accident (such as property damage, injury and fatality): crashes involving alcohol accounted for 16.7 per cent of accidents involving property damage, 20.4 per cent of accidents involving injuries and 40.8 per cent of accidents involving fatalities;

(b) A calculation of the percentage of total motor vehicle accidents involving alcohol for each accident type: property damage involving alcohol represented 78 per cent of all traffic accidents involving alcohol; injuries represented 21 per cent and fatalities represented 0.003 per cent;

(c) A calculation of an average for total accidents involving alcohol.

Using that approach, CASA estimated that 17.6 per cent of highway traffic accidents involved alcohol. That percentage was also applied to accident prevention programmes, state highway patrol and local law enforcement programmes not specifically targeting alcohol or drug abuse. Costs were adjusted for state differences in prevalence of alcohol use. The total cost of programmes targeting alcohol or drug abuse was included.

Capital costs

As mentioned in other categories, CASA included in its analysis funds expended (not budgeted amounts) by the state for new construction, capital improvements and equipment for adult and juvenile correctional facilities, substance abuse treatment facilities, mental health facilities and facilities for the developmentally disabled. Included were funds paid for out of current general taxes or dedicated taxes, capital spending from bond proceeds and interest paid out for bonds already issued. The share of each category spent on efforts to deal with substance abuse (for example, 81 per cent for capital spending on adult correctional facilities and 50.9 per cent for capital spending on mental health), adjusted for prevalence of alcohol and illicit drug use, was used to estimate the portion of capital spending linked to substance abuse. Capital spending associated with substance abuse was added to other costs in each category.

Education

States spent almost one quarter of total state funds (22 per cent) on primary and secondary education in the fiscal year 1998 [10]. In that area of the budget, it was difficult to establish the share of state spending attributable to substance abuse for three major reasons:

(a) State governments allocated most education funds in lump sums to local school districts;

*The estimation was made with the guidance of the author of the report presenting the data collected by the National Highway Traffic Safety Administration [32].

(b) There was a bias against labelling children; therefore, it was difficult for researchers to determine which children were exposed to substances in utero or in the home and which children were abusing substances;

(c) It was difficult to find literature or research linking costs in the education system to substance abuse.

Using the *International Guidelines for Estimating the Costs of Substance Abuse* [33] as a benchmark, there is neither a matrix of costs nor any delineation of the theoretical issues to help lead to agreement on how to measure those costs in the case of public education. Nonetheless, there is a broad consensus that the costs are potentially significant.*

CASA has identified three ways that substance abuse affects schools:

(a) Parental use affects the capacity and readiness of children to learn;

(b) Faculty and staff use affects the learning environment;

(c) Student use affects the interest and capacity to learn, and school security.

All of those factors might affect the costs of education. For example, maternal alcohol use during pregnancy could result in increased special education costs for students with fetal alcohol syndrome. Parental substance abuse might result in programmes for at-risk youth, staff-intensive compensatory education programmes, after-school programmes, summer school and other programmes. Substance abuse by students might necessitate increased staff in support and health-care roles or might result in class disruption; violence associated with such substance abuse might require increased school security costs for security personnel and equipment and increased expenses for insurance and workers' compensation and for repairs and replacement of vandalized or stolen material. Substance abuse by faculty members might involve increased workforce costs and lost productivity.

Although few of those costs are reported to states in ways that can be linked to budgets, when those costs are combined, they represent a considerable amount of expenditure. To take the first steps towards developing an estimate of the costs of substance abuse to the education system, CASA identified cost areas that could be linked to substance abuse, including the following:

Lost productivity of staff and added costs for additional staffing
Programmes for at-risk children
Prevention programmes
Special education programmes
Delinquency
Administration
Property and liability
Staff health insurance
Legal expenses
Drug testing
Employee assistance programmes
Employee training, policy and staff development
Capital outlay

*Experts in the field of education, school finance and cost estimation of substance abuse formed a focus group, conducted by CASA on 19 July 1999, in Washington, D. C. The group reached the conclusion that the costs of substance abuse were potentially significant.

CASA estimated that those costs combined could account for between 10 and 22 per cent of annual state expenditures for primary and secondary education.

To review that approach and associated estimates of costs, CASA convened a group of experts in the area of school finance and substance abuse. The group was concerned by the unavailability of data for making more precise estimates, but after reviewing and refining the list of effects, it informally posited a range of 10-20 per cent for the estimated impact of substance abuse on the public education system. For the purposes of the present study, the lower end of the range, 10 per cent, was chosen as a conservative estimate of the share of education spending attributable to substance abuse.

CASA included that estimate as a placeholder for budget purposes for three reasons:

(a) State budgets were heavily dominated by education spending, and failure to recognize costs in that area would be a major oversight;

(b) According to experts in the field and qualitative literature, substance abuse had a significant impact on schools and on the achievement of their goals;

(c) Schools represented an important opportunity to intervene, since problems of substance abuse that started in primary and secondary school would show up later in other state systems such as correctional services, child welfare, mental health or welfare.

By including that budget estimate, CASA hoped to promote research into the question of the impact of substance abuse on schools and education spending.

Regulation and compliance

In its analysis, CASA total spending on state personnel responsible for collecting alcohol and tobacco taxes (including fringe benefits) and state funds budgeted for boards or governing bodies enforcing alcohol and tobacco regulations or issuing alcohol and tobacco licences. Those costs were 100 per cent attributable to substance abuse. CASA also estimated the total tax revenues received by states from alcohol and tobacco sales.

Policy implications

The present study documents the full dimensions of state spending linked to substance abuse and addiction. Specifically, the study reveals that states devote tremendous resources to managing the effects of substance abuse through incarceration, treatment in health and mental health systems, foster care, special and compensatory education and various other areas. Only a small portion of state spending is aimed at reducing substance abuse through treatment and prevention programmes.

By focusing on state budget costs, it can be shown more clearly which of those costs are under the control of state policy makers. It does not, however, provide a comprehensive picture of governmental spending on that problem. Similar analyses of federal and local spending are needed to complete the picture.

By providing a map of state spending on substance abuse, the present study establishes a base against which policy makers can begin to compare the value of alternative prevention, treatment, regulatory and tax policies that might reduce the consequences for the state budget of substance abuse and addiction. Understanding the enormous costs attached to current policy choices on substance abuse should help policy makers think more strategically about how they might invest in interventions that would yield a better return on their investments.

If, for example, 4 per cent of a state budget is linked to coping with the consequences of untreated addiction in state prisons, the costs and benefits of interventions to reduce those costs can be considered in terms of how they affect current spending. States might want to consider investments such as the following: the elimination of mandatory sentences for drug offences and the requirement of mandatory treatment; treatment-linked diversion programmes; prevention activities in the schools; treatment for parents who abuse substances and who neglect or abuse their children; increased taxes on alcohol; or any of a variety of other interventions that might reduce the consequences of substance abuse linked to crime.

As the present study shows, a policy response of reductions in prevention or treatment expenditures will have the effect of increasing rather than decreasing state costs. Furthermore, policy strategies that involve only civil or criminal justice sanctions without requiring treatment will, in the long term, raise rather than reduce state costs. By thinking about expenditures as investments, policy makers will be in a better position to demand specific results for their investments. An investment-based approach will help policy makers ensure accountability for expenditure of public funds by showing the return and the results.

CASA recognizes that the spending map presented is breaking new ground and hopes that it will help to shift the public debate from expenditures on the problem of substance abuse to investments in better results.

References

1. H. Harwood, D. Fountain and G. Livermore, *The Economic Costs of Alcohol and Drug Abuse in the United States, 1992* (Washington, D. C., National Institute on Drug Abuse and National Institute on Alcohol Abuse and Alcoholism, 1998).

2. D. P. Rice, "The economic cost of alcohol abuse and alcohol dependence: 1990", *Alcohol Health and Research World*, vol. 17, No. 1 (1993), p. 10.

3. National Center on Addiction and Substance Abuse at Columbia University, *The Cost of Substance Abuse to America's Health Care System; Report 1: Medicaid Hospital Costs* (New York, National Center on Addiction and Substance Abuse at Columbia University, 1993).

4. National Center on Addiction and Substance Abuse at Columbia University, *Cost of Substance Abuse to America's Health Care System; Report 2: Medicare Hospital Costs* (New York, National Center on Addiction and Substance Abuse at Columbia University, 1999).

5. National Center on Addiction and Substance Abuse at Columbia University, *Substance Abuse and Federal Entitlement Programs* (New York, National Center on Addiction and Substance Abuse at Columbia University, 1995).

6. National Center on Addiction and Substance Abuse at Columbia University, *Behind Bars: Substance Abuse and America's Prison Population* (New York, 1998).

7. National Center on Addiction and Substance Abuse at Columbia University, *No Safe Haven: Children of Substance-Abusing Parents* (New York, National Center on Addiction and Substance Abuse at Columbia University, 1999).

8. D. R. Gerstein and others, *Evaluating Recovery Services: the California Drug and Alcohol Treatment Assessment* (Sacramento, California, Department of Alcohol and Drug Programs, 1994).

9. Centers for Disease Control and Prevention, "1997 BRFSS summary prevalence report", National Center for Chronic Disease Prevention and Health Promotion, 1998.

10. National Association of State Budget Officers, *Expenditures Report, FY 1998* (Washington, D. C., National Association of State Budget Officers, 1999).

11. L. S. Miller and others, "State estimates of total medical expenditures attributable to cigarette smoking", *Public Health Reports*, vol. 113, No. 5 (1993), pp. 447-458.

12. *International Classification of Diseases, Ninth Revision: Clinical Modification (ICD-9-CM)*, 5th ed. (Los Angeles, Practice Management Information Corporation, 1997).

13. T. C. Blum, P. M. Roman and J. K. Martin, "Alcohol consumption and work performance", *Journal of Studies on Alcohol*, vol. 2, No. 1 (1993), p. 61.

14. J. P. Hoffmann, C. Larison and A. Sanderson, *An Analysis of Worker Drug Use and Workplace Policies and Programs* (Rockville, Maryland, United States Department of Health and Human Services, 1997).

15. M. T. French, G. A. Zarkin and L. J. Dunlap, "Illicit drug use, absenteeism, and earnings at six U.S. worksites", *Contemporary Economic Policy*, vol. 16, No. 3 (1998), p. 334.

16. D. Brenna, "Substance abuse services in juvenile justice: the Washington experience", C. G. Leukefeld and F. M. Tims, eds., *Drug Abuse Treatment in Prisons and Jails*, NIDA Monograph Series No. 118 (Rockville, Maryland, National Institute on Drug Abuse, 1992).

17. A. Kline and G. Rodriguez, *Substance Use and Dependency among New Jersey Juvenile Arrestees* (Trenton, New Jersey, Department of Health and Senior Services, 1996).

18. Arizona Department of Juvenile Corrections, "Maricopa County juvenile arrestee drug usage", 1997.

19. J. M. Brown, and P. A. Langan, *State Court Sentencing of Convicted Felons*, 1994 (Washington, D. C., United States Department of Justice, 1998).

20. A. T. Stahl and others, *Juvenile Court Statistics 1996* (Washington, D. C., United States Department of Justice, 1999).

21. B. J. Ostrom and N. B. Kander, *Examining the Work of State Courts, 1998: a National Perspective from the Court Statistics Project* (Williamsburg, Virginia, National Center for State Courts, 1999).

22. M. J. Bane and J. Semidei, *Families in the Child Welfare System: Foster Care and Preventive Services in the Nineties* (New York, Department of Social Services, 1992).

23. Connecticut, Department of Children and Families, *A Report to the General Assembly: Child Protective Services and Adult Substance Abuse Treatment* (Hartford, Connecticut, Department of Children and Families, 1997).

24. United States of America, General Accounting Office, *Foster Care: Parental Drug Abuse Has Alarming Impact on Young Children* (Washington, D. C., United States General Accounting Office, 1994).

25. Nevada, Washoe County, Department of Social Services, *Fiscal Year 1994-95 Annual Report* (Reno, Nevada, Washoe County Department of Social Services, 1995).

26. K. Olson and L. Pavetti, *Personal and Family Challenges to the Successful Transition from Welfare to Work* (Washington, D. C., United States Department of Health and Human Services, 1996).

27. National Center on Addiction and Substance Abuse at Columbia University, *Substance Abuse and Women on Welfare* (New York, National Center on Addiction and Substance Abuse at Columbia University, 1994).

28. L. Schmidt, C. Weisner and J. Wiley, "Substance abuse and the course of welfare dependency", *American Journal of Public Health*, vol. 88, No. 11 (1998), p. 1616.

29. D. C. Stapleton and others, *Policy Evaluation of the Effect of Legislation Prohibiting the Payment of Disability Benefits to Individuals Whose Disability is Based on Drug Addiction and Alcoholism* (Fairfax, Virginia, Social Security Administration, 1998).

30. R. C. Kessler and others, "The epidemiology of co-occurring addictive and mental disorders: implications for prevention and service utilization", *American Orthopsychiatric Association*, vol. 66, No. 1 (1996), pp. 17-31.

31. D. Braddock, ed., *The State of the States in Development Disabilities* (Washington, D. C., American Association on Mental Retardation, 1998).

32. L. J. Blincoe, "The economic cost of motor vehicle crashes, 1994" (United States Department of Transportation, National Highway Traffic Safety Administration, technical report, 1996).

33. E. Single and others, *International Guidelines for Estimating the Costs of Substance Abuse* (Ottawa, Canadian Centre on Substance Abuse, 1996).

Offprint from Bulletin on Narcotics, vol. LII, Nos. 1 and 2, 2000

Estimating the economic costs of drug abuse in Colombia

A. PÉREZ-GÓMEZ

Director, RUMBOS—Programa Presidencial para Abordar el Consumo de Drogas (Presidential Programme against Drug Abuse), Bogotá, Colombia

E. WILSON-CAICEDO

Adviser, Instituto Roosevelt de Ortopedia Infantil, Bogotá, Colombia

ABSTRACT

The development of reliable estimates of the economic costs of substance abuse can help to prioritize drug issues and provide useful information to policy makers. Nevertheless, only a few developed countries have so far attempted to carry out such studies.

The present paper reviews the current situation of drug abuse in a developing country widely recognized as a country struggling to overcome a problem of drug production. The data confirm the trend observed elsewhere: countries in which drugs are produced also tend to see rising levels of consumption. On the other hand, it is shown that the lack of systematic information in many areas makes it almost impossible to calculate the costs of drug abuse in countries such as Colombia. Various suggestions are made with a view to correcting some of the problems identified.

Introduction*

In assessing the economic costs of drug abuse in Colombia, the following background information on the economic situation of the country should be borne in mind:

(a) At the time of the last census, in 1993, the total population of Colombia was set at 37.7 million. Projected estimates for 2000 place the figure at slightly over 42.3 million, including 12.4 million (29.3 per cent) people between 10 and 24 years of age;

(b) In December 1999, unemployment in the 11 main urban areas of Colombia was estimated at 18.1 per cent. The figure reached 20.1 per cent by the end of March 2000, which meant that some 1.5 million people were out of jobs;

*Figures given in the present section were drawn from: Departamento Nacional de Planeación, *Indicadores de Coyuntura Económica Mensual—Febrero de 2000* (Bogotá, 2000); Departamento Administrativo Nacional de Estadística, *Proyecciones Anuales de Población 1985-2015* (Bogotá, 2000); and Departamento Administrativo Nacional de Estadística, *Encuesta Nacional de Hogares 1999* (Bogotá, 2000).

(c) The gross national product for 1998 was estimated at 99,357.6 million United States dollars ($). For 1999, it was expected to fall by almost 15 per cent to $84,742 million;

(d) The internal inflation rate for 1999 was 9.2 per cent, while devaluation of the Colombian peso against the dollar reached 22.2 per cent (estimates for 2000: 10 per cent and 17.5 per cent, respectively);

(e) Colombian exports for the period January-November 1999 totalled $10,325.7 million, while imports amounted to $9,627.2 million. Exports consisted mainly of oil, coffee, chemical products for industrial use, coal, bananas and flowers. Imports, on the other hand, originated mostly in the United States of America, the European Community and the countries known as the Andean group (Bolivia, Ecuador, Peru and Venezuela);

(f) As estimated by a research team from a leading local university in a recent study,* drug production in Colombia may have led to illegal exports worth $2,229 million per year on average over the period 1982-1998. If accurate, the figure would surpass that of coffee, thus making the production of illicit crops the country's most important source of income based on agriculture.

Drug production estimates for 1995

From the early 1970s, an illegal drug industry gradually evolved in Colombia. Narcotrafficking, the term often used to describe the illicit activity, comprises the growth and trade (export) of marijuana, coca leaf and poppy flower. Only a handful of studies are available that attempt to analyse the economics of drug production. To date, no studies have been published regarding the economics of drug use.

According to one study on drug production [1], crops of marijuana, coca leaf and poppy flower covered some 110,000 hectares (ha) in 1994, that is, an area equivalent to 1.4 per cent of the land devoted to agricultural use in Colombia. On the basis of the same study, a brief review of the economics of each type of drug is given below.

Marijuana

Although cannabis was introduced at the time of the Spanish conquest, it was only by the 1970s that the illegal production and trade of marijuana began in Colombia as a direct result of the eradication of crops in Mexico. By the end of the same decade, production on the northern coast of Colombia was at a high point. Local government efforts to control the illicit exports, combined with the increasing substitution of imports by the United States, led to a decline in marijuana production and a displacement of the illicit industry towards the production of cocaine and heroin instead. Seemingly, these were regarded as easier products to smuggle, the crops being more difficult to detect and the produce yielding a higher price-to-volume ratio. Nevertheless, by 1994, the production and export of marijuana from Colombia to the United States market was second only to that of Mexico.

*Study carried out at the Universidad de los Andes, Centro de Estudios de Desarrollo Económico, quoted in *El Tiempo*, 5 August 2000 (Bogotá), pp. 2-7.

Most of the production of marijuana in Colombia takes place near the Atlantic coast in the northern part of the country around Sierra Nevada de Santa Marta in Magdalena province and Serranía del Perijá in Cesar province. Crops are also to be found in the southern provinces of Cauca and Nariño. Though there was a sharp decline in marijuana exports during the 1980s (their estimated worth is thought to have fallen from $132 million to $20 million), production seems to have somewhat recovered in recent years, possibly as a result of an apparent stabilization of the risks involved in smuggling it into the consumer markets. That may have something to do with a relaxation of legal sanctions against use in those markets, as they currently regard it as a soft drug.

By 1994, the total area devoted to growing marijuana in Colombia was estimated at 6,112 ha (down from an average of 10,062 ha in 1982-1987), and the corresponding exports were valued at some $250 million.

Coca leaf

Traditionally, the quality of local produce has been regarded as lower than that of Bolivia and Peru. As a consequence, until the earlier part of the 1980s, most of the base used in the cocaine production process in Colombia seems to have been imported from Bolivia and Peru, regarded, in that order, as the world's largest producers. Since then, however, the crop area of coca leaf has increased steadily (mainly in the eastern provinces of Meta, Caquetá and Guaviare and in southern Putumayo). In 1994, it was estimated at 45,000 ha, according to one source [2], while other estimates have been as high as 83,000 ha [3]. Depending on the figure of choice, Colombia would have become the third (or second) largest producer of coca leaf in the world.

Poppy flower

Poppy flower was the last drug-producing crop to be introduced into Colombia during the late 1980s. Since then, the area under poppy flower cultivation has rapidly expanded. In 1994, an estimated 20,405 ha were devoted to growing poppy flower, mostly in the central provinces of Tolima and Huila and in eastern Caquetá.

Little more is known about the economics of poppy flower cultivation in Colombia. From the flower, a latex is extracted from which heroin and morphine are obtained. The world's main producers are Myanmar and Afghanistan, followed by Colombia in third place.

Substance abuse in Colombia, 1992-1996

Major studies on substance abuse were conducted in Colombia for the years 1992 [4] and 1996 [5]. They had been commissioned by the Dirección Nacional de Estupefacientes, the national agency in charge of coordinating efforts to prevent and control illicit drug abuse. The same methodology was applied in both studies, although the size of the sample was about twice as large in 1996. The findings presented in the report on the 1996 study [6] are briefly summarized below.

Abuse of illicit drugs in Colombia was estimated to occur, at some time, among 6.5 per cent of the population, or 1,676,924 individuals. Prevalence of abuse was approximately four times higher among males than among females. On the whole, the extent of abuse increased with the level of instruction. The northwestern province of Antioquia had the highest rate of drug abuse (12.3 per cent). Higher levels occurred in urban areas.

In the year prior to the survey, 1.6 per cent of the population, or 400,768 individuals reported the abuse of at least one illicit drug.

In 1996, as compared with 1992, there was an increase in the abuse of drugs. It was a matter of concern that the estimated prevalence of abuse had doubled in just four years. The main differences were found in women, in 12- to 17-year-olds, and in the unemployed. Geographically, only the Pacific coast region showed a significant change.

The number of new abusers of any illicit drug was estimated at 117,453. For the most part, they were between 12 and 17 years of age, currently attended school, and lived in densely populated urban areas in the coffee-growing region (locally known as eje cafetero), and in or around the main cities (Bogotá, Cali and Medellín).

Marijuana had been, and remained, the illicit drug of choice for the majority of abusers. At least once in their lifetime, 5.4 per cent of the population had used it. Abusers of cocaine and basuco were estimated at 1.6 per cent and 1.5 per cent of the population, respectively. Heroin, on the other hand, showed a low prevalence of abuse. It was estimated that 12,566 individuals had used heroin at least once. A comparison of the abuse of illicit drugs by type during the year previous to the 1996 survey with the results of the 1992 survey showed that an increase in the abuse of marijuana (from 0.6 to 1.1 per cent) accounted for most of the difference.

An increase in the abuse of illicit drugs had important implications for treatment and prevention. In that regard, there was an obvious need to focus immediate attention on children and adolescents. In the long run, an increasing number of young abusers of illicit drugs would probably push up demand for treatment in the years to come, since many new abusers gradually sank into addiction.

During the month prior to the 1996 survey, 18.5 per cent of the population under study had used tobacco, and 21.4 per cent had done so within the last year. Proportions varied greatly according to gender and increased with age. The 25- to 44-year-old age group accounted for most of the abusers. A high prevalence of abuse was found among individuals with a low level of instruction, as well as among those with a high level. Prevalence was higher among the employed than among the unemployed. Geographically, drug abuse tended to increase with urbanization. Antioquia and Bogotá showed the highest rates of abuse within one month of, and a year prior to, the survey. In comparison with those of 1992, levels of drug abuse in the two cases were 3 and 4 per cent lower, respectively, the difference being greater among individuals over 25 years of age living in Bogotá or on the Atlantic coast. Students and unemployed individuals increased their use of tobacco.

The proportion of the population using alcoholic beverages of any kind during the year prior to, and within one month of, the survey was estimated at 59.8 and 35 per cent, respectively. Higher levels were to be found among 18- to 44-year-old males with a university education and currently employed. Geographically, a higher prevalence was to be found in Bogotá and the eastern region of Colombia. Use of alcohol tended to be higher in both low- and high-population areas.

The index known as the CAGE index of alcoholism* shows that an estimated 15.8 per cent of the population between 12 and 60 years of age was at risk of

*Based on D. Mayfield, G. McLeod and P. Hall, “The CAGE questionnaire: validation of a new alcoholism instrument”, *American Journal of Psychiatry*, vol. 131 (1974), pp. 1121-1123.

becoming alcoholic, while among the group of alcohol users within the last year, the proportion went up to 25.6 per cent. Prevalence was higher among 25- to 44-year-old males currently employed and living in or around Bogotá and in the eastern region of Colombia.

The number of people taking psychoactive pills of any kind during the year prior to the survey was estimated at 182,000, or 0.7 per cent of the population between 12 and 60 years of age. A higher level of abuse was associated with 45- to 60-year-old females (women were found to outpace men by more than two to one), most of them housekeepers living in urban areas. There was an observable decrease in the use of drugs of this type in 1996 as compared with 1992, although methodological difficulties were reported concerning the estimation of lifetime prevalence.

The proportion of the population that abused inhalable substances (such as gasoline, paint, thinner and glue) was estimated at 6.7 per cent, with a higher rate among 25- to 44-year-old males. The Pacific coast and the eastern region of Colombia showed a higher prevalence.

Substance abuse among Colombian youth in 1999

On 20 October 1999, the Colombian Programa Presidencial para Abordar el Consumo de Drogas (Presidential Programme against Drug Abuse), known as RUMBOS, conducted a national survey on drug consumption among 10- to 24-year-olds living in the provincial capital cities of Colombia. A very simple instrument in the form of a six-question survey was devised to overcome the universal aversion of young people to long questionnaires. A multimedia campaign had been conducted nationwide during the previous September, inviting youth to participate, and regional and local authorities as well as public and private civic organizations accepted the invitation to become involved. The project was a complete success. Out of 420,000 forms prepared, a record 305,869 (or 72.8 per cent) were submitted in response.

Differences in both methodology and target population make it impossible to compare the results of the survey with those of the two major studies on substance abuse referred to above. Moreover, the limited amount of information to be derived from such a simplified questionnaire could raise doubts about the exercise. Such an approach, however, has the following advantages:

(a) The size of the sample and the quality of the inferences based on the observations made are far greater than what can be achieved through the alternative method of household surveys, in which the number of responses were, by comparison, 8,975 in 1992 and 18,770 in 1996;

(b) By their very nature, household surveys cannot account for drug abuse in environments other than homes. Thus, institutionalized abusers (for example, those in foster homes, rehabilitation or detention facilities) go unnoticed who probably would have made a significant impact on the results. Very much the same could be said of children and youth living on the streets (a sad and depressing reality in developing countries) who are probably much more frequent drug abusers than their counterparts in the sanctuary of the home;

(c) The target population of youth between 10 and 24 years of age consists precisely of those who, statistically, are both at higher risk of becoming abusers

(requiring preventive measures) and, at the same time, more likely to be rescued from the inferno of addiction (through treatment);

(d) Simple and straightforward questions do away with interviewers and, at the same time, eliminate or reduce the possibility of denial or conscious alteration of facts;

(e) Lastly, the economics of the approach make a big difference, especially in countries such as Colombia, where the resources available to meet social needs are so limited.

Globally, the main findings of the RUMBOS survey may be summarized as follows:

(a) Among licit psychoactive substances, alcohol has, by far, the highest prevalence in terms of lifetime use (72.9 per cent) and use in the last month before the survey (47.6 per cent). Tobacco comes second (35.9 and 20.0 per cent for lifetime and last-month use, respectively). Use of tranquillizers and inhalables is much less frequent (approximately 2 and 1 per cent, respectively);

(b) The overall results for illicit drugs are shown in the following table:

Level of drug abuse

(Percentage of total responses)

	Period of use	
Drug	*Lifetime*	*Last month*
Marijuana	9.2	3.6
Cocaine	3.6	1.2
Basuco	2.1	0.9
Ecstasy	1.8	0.6
Hallucinogenic fungi	1.3	0.3
Acid	0.6	0.2
Mandrax	0.3	0.1
Amphetamines	0.7	0.2
Heroin	0.8	0.4

(c) In terms of gender, males equal females in their use of licit substances, except in the case of inhalables (where the proportion is 2:1). With illicit drugs, males exceed females in the following proportions: 2:1 (marijuana, ecstasy, acid and amphetamines); 3:1 (cocaine, fungi and heroin); and 4:1 (mandrax);

(d) The highest levels of abuse, whether of licit or illicit substances, are to be found among 20- to 24-year-olds, with adolescents aged 15 to 19 closely behind;

(e) Education does seem to have an effect on patterns of abuse. Youth at university level tend to indulge more in alcohol and tobacco. In the abuse of illicit drugs, that group comes second after youth with little or no education. Basuco and inhalables abuse is more frequent among the less educated;

(f) Youth without an activity (non-students or the unemployed) are prone to indulge in substance abuse as measured by the last-month prevalence ratio. In the long run, however, as measured by lifetime prevalence, the situation reverses.

Towards a study of the cost of illness associated with drug abuse

A three-day seminar on the estimation of the economic cost of drug abuse in Colombia, sponsored by the Embassy of Canada in cooperation with the Canadian Centre on Substance Abuse (CCSA), was held at RUMBOS in Bogotá, from 27 to 29 March 2000.

During the seminar, a comprehensive review was presented on the cost-of-illness methodology as applied to economic cost estimates of drug abuse. It soon became evident that the proposed CCSA international guidelines for estimating the costs of substance abuse might provide a useful framework for a similar study to be conducted in Colombia. There are, however, important differences between developed consumer countries and developing producer/consumer countries, differences extending well beyond the mere notion of economic development and calling for a somewhat varied approach to cost estimation and the selection of alternative analytical tools.

The present section provides a preliminary inventory of differences and key issues that need to be addressed in order to formulate alternative guidelines on cost estimation for cases such as that of Colombia.

General inadequacy of statistical data

Failure to maintain up-to-date records makes it very difficult to carry out detailed studies on specific issues. Colombia has been traditionally averse to record-keeping, a fact that has made research often nearly impossible, be it because of a total lack of information, random discontinuity in a time series, or an unexplained inconsistency among sources or between different periods. It is therefore not unusual, in Colombia, to find extreme discrepancies in estimates from what seem to be the simplest economic data. In such circumstances, studies that might have appeared easy when first undertaken often become impossible, and the task of research becomes endless and inconclusive.

Types of cost not incurred in developing producer/consumer countries

Certain costs, however normal in developed countries and regardless of their economic or social justification, appear to be either not currently found in developing countries, or of such limited importance as to be irrelevant for all practical purposes. Those costs include the following:

(a) Welfare costs arising from:

- *(i)* Treatment of addictions (such costs, in developing producer/consumer countries, being sometimes paid by the drug abusers themselves or by their families, although, more often than not, the addiction goes untreated);
- *(ii)* Unemployment benefits (food stamps or similar programmes);

(b) Costs of prevention and research.

Types of cost not incurred in developed consumer countries

Certain costs arising from drug-related activities, although part and parcel of day-to-day reality in developing producer/consumer countries, are largely unknown in developed consumer countries. Those costs include the following:

(a) Widespread corruption brought about by huge amounts of cash that make it possible to buy everything and everyone;

(b) Damage to the environment at every stage of the illegal activity (destruction of tropical forest to allow agriculture, and subsequent destruction of illicit crops by means of chemical agents that render the land useless for many years or for good);

(c) Economic effects of money (laundered or otherwise) derived from the manufacture of and trafficking in narcotic drugs;

(d) Economic effects of guerrilla warfare associated with the manufacture of and trafficking in narcotic drugs and its sequel (casualties and injuries; actual or threatened kidnappings with demands for ransom payments; private security arrangements; forced recruitment of youth/children; dislocation of communities; and property losses due to the impossibility of finding buyers);

(e) Law enforcement costs relating to the manufacture of and trafficking in narcotic drugs, including both police and military costs (whether or not borne by other countries);

(f) General institutional instability/destabilization.

Types of cost requiring alternative evaluation techniques

Certain costs, while similar in both developed and developing countries, should probably receive different treatment when attempting to place a value on them. They include the cost of human life in the presence of: chronically high levels of unemployment; a state of civil war; and measures to promote social cleansing.

Final comments on attempts to estimate costs arising from consumption

Efforts to estimate costs arising from the consumption of drugs may be hindered by certain limitations, ranging from severe to insurmountable. On the basis of a tentative list prepared for the Bogotá seminar, those limitations are briefly reviewed below.

Productivity losses from morbidity

Apart from the lack of precise record-keeping by Colombian hospitals, which makes it exceedingly difficult to accurately measure morbidity in general, there is an additional problem. Due to regulations governing the provision of health care, doctors and hospitals are frequently compelled not to record and report the real causes of a medical condition, lest the patient, with other options closed, go untreated.

Law enforcement costs arising from drug possession and crime related to drug use

There is no tradition of record-keeping in regard to law enforcement costs arising from drug possession and crime related to drug use, which may have made little difference, since most of the cases never reach the courts, or even the police station. Impunity has thus become the law of the land in Colombia, with general corruption as its unfortunate result.

Health-care costs

Because of a chronic lack of resources in the public welfare system, most health-care costs are borne by the private sector, including health maintenance organizations, insurance companies, families and individuals. There are no information systems or agencies that automatically collect the relevant data.

Prevention and research

The costs associated with prevention and research should be less difficult to assess, given the small sums involved. Prevention is mainly the responsibility of government agencies, always short of funds. Research, apart from the efforts of a few private institutions, is not actively promoted in Latin American countries.

Other costs

Other costs associated with substance abuse, although difficult to assess in economic terms, include those entailed by: children dropping out of school because of substance abuse; accidents, injuries or crimes caused by individuals while under the influence of drugs; losses in the workplace caused by accidents, absenteeism or psychological disorders brought about or worsened by drug abuse; and loss of value of real estate located in or around consumption areas.

References

1. Ricardo Rocha García, “Aspectos económicos de las drogas illegales”, in *Drogas Ilícitas en Colombia, Su Impacto Económico, Político y Social*, Francisco E. Thoumi and others, eds. (Bogotá, United Nations Development Programme/Dirección Nacional de Estupefacientes/Planeta Colombiana Editorial, 1997), pp. 141-151.

2. United States Department of State, Bureau for International Narcotics and Law Enforcement Affairs, *International Narcotics Control Strategy Report* (Washington, D.C., 1995).

3. D. Uribe, “Los cultivos ilícitos en Colombia”, in *Drogas Ilícitas en Colombia: su Impacto Económico, Político y Social*, F. E. Thoumi and others (Bogotá, Ediciones Ariel, 1997), pp. 35-136.

4. Dirección Nacional de Estupefacientes/Escuela Colombiana de Medicina/Fundación Santafé de Bogotá, *Estudio Nacional sobre Consumo de Sustancias Psicoactivas en Colombia, 1992* (Bogotá, 1993).

5. Dirección Nacional de Estupefacientes/Fundación Santafé de Bogotá, *Consumo de Sustancias Psicoactivas en Colombia, 1996* (Bogotá, 1997).

6. Ibid., pp. 99-101.

Offprint from Bulletin on Narcotics, vol. LII, Nos. 1 and 2, 2000

Illegal activities and the generation of value added: size, causes and measurement of shadow economies

F. SCHNEIDER*

Professor of Economics, Department of Economics, Johannes Kepler University of Linz, Linz-Auhof, Austria

ABSTRACT

Various methods of measurement are used to establish and present estimates of the size of the shadow economy in 76 developing countries, countries with economies in transition and States members of the Organization for Economic Cooperation and Development (OECD). From 1989 to 1993, the average size of the shadow economy as a percentage of gross domestic product (GDP) was 39 per cent in developing countries, 23 per cent in transition countries, and 12 per cent in OECD countries. An increasing burden of taxation and social security contributions combined with rising State regulatory activities are the driving forces behind the growth of the shadow economy. According to some findings, a growing shadow economy has a negative impact on official GDP growth, although other studies show the opposite effect.

Introduction

As crime and other underground economic activities (including the shadow economy) are a fact of life around the world, most societies attempt to control such activities by measures such as punishment, prosecution, economic growth or education. Gathering statistics about who is engaged in underground (or criminal) activities and about the frequency and scale of such activities is crucial for sound decision-making on the allocation of resources in this area. Unfortunately, it is very difficult to get accurate information about underground, or shadow-economy, activities, because those involved are careful to conceal their identities. Estimating shadow-economy activities can therefore be likened to a scientific passion for knowing the unknown.

Although a large literature** is devoted to specific aspects of the hidden economy, a comprehensive survey has just been carried out by Schneider and Enste

*Email address: friedrich.schneider@jk.uni-linz.ac.at.

**The literature on the “shadow”, “underground”, “informal”, “second”, “cash” or “parallel” economy is expanding rapidly. Various topics, including its measurement, its causes and its impact on the official economy, have been analysed. See, for example, Tanzi [2, 3], Frey and Pommerehne [4], Feige [5], Thomas [6], Loayza [7], Pozo [8], Lippert and Walker [9], Schneider [10-13], Johnson, Kaufmann and Shleifer [14], Johnson, Kaufmann and Zoido-Lobatón [15].

[1]. The subject remains controversial [16], and there are disagreements about the definition of shadow-economy activities, measurement procedures and the use of estimates in economic analysis and addressing policy issues.* Nevertheless, around the world, there are strong indications of a growing shadow economy. The size, the causes and the consequences of the shadow economy are different in different countries, but comparisons are possible and they may be of interest to social scientists and the public at large and useful to politicians, who will have to deal with this phenomenon sooner or later. By their very nature, shadow-economy activities are difficult to measure, and there is a wide divergence of opinion among academics, experts in the public sector, policy or economic analysts and politicians, as to what the shadow economy is all about or how big it is.

Nevertheless, the shadow economy is an area of growing concern, and there are several important reasons why politicians and public sector officials should be worried about its size and growth, including the following:

(a) If an increase in the shadow economy is caused mainly by a heavier tax and social security burden, it could lead to an erosion of the tax base and social security contributions, resulting in a decrease in tax revenue and a rising budget deficit, triggering even higher tax rates with a consequent growth in the shadow economy as the cycle continues. A growing shadow economy can therefore be seen as the result of decisions taken by individuals who feel overwhelmed by the demands of the State;

(b) A growing shadow economy implies that economic policy is based on erroneous or unreliable official indicators (such as unemployment rates, labour force statistics and levels of income and consumption). In such a situation, a prospering shadow economy may cause severe difficulties for politicians, because the unreliable official indicators may be used as the basis for questionable policy measures;

(c) On the one hand, a growing shadow economy may provide strong incentives to lure domestic and foreign workers and other resources away from the official economy. On the other hand, two thirds of the income earned in the shadow economy is spent in, and strongly stimulates the growth of, the official economy.**

The growing concerns referred to above and the complex issues raised by the underground economy inspired the present study, which involved the challenging task of collecting all available data on the shadow economy, in an effort to gain insight into its main causes and its effects on the official economy. The matters covered in the sections below are as follows: first, an attempt is made to define the shadow economy; secondly, empirical results are presented on the size of the shadow economy in 76 countries throughout the world; thirdly, the main causes of the shadow economy are considered; fourthly, the interactions of the official and unofficial economies are analysed; fifthly, the various methods used to estimate the size of the shadow economy are presented; and finally, a summary is provided and various conclusions are drawn.

*Compare the views of Tanzi [17], Thomas [18] and Giles [19, 20].

**This figure has been derived from polls of the German and Austrian populations about the effects of the shadow economy. For further information, see Schneider [21]. The polls also show that two thirds of the value added accounted for by the shadow economy would not be produced in the official economy if the shadow economy did not exist.

Definition of shadow economy

Most attempts to measure the shadow economy are hampered by the difficulty of defining the term. According to one commonly used working definition, the shadow economy consists of all currently unregistered economic activities that contribute to the officially calculated (or observed) gross national product (GNP).* Smith [24] defines it as "market-based production of goods and services, whether legal or illegal, that escapes detection in the official estimates of gross domestic product". As such definitions leave many open questions, table 1 might shed some light on the content of a possible consensus definition of the legal and illegal underground or shadow economy.

Table 1. Taxonomy of types of underground economic activity

Type of activity involved	*Monetary transactions and tax issues*		*Non-monetary transactions and tax issues*	
Illegal	Trade in stolen goods; drug manufacture and trafficking; prostitution; gambling; smuggling and fraud		Barter of drugs and stolen goods, smuggling etc.; production or theft of drugs for own use	
	Involving tax evasion	*Involving tax avoidance*	*Involving tax evasion*	*Involving tax avoidance*
Legal	Unreported income from self-employment; wages, salaries and assets from unreported work related to legal services and goods	Employee discounts and fringe benefits	Barter of legal services and goods	Do-it-yourself work and help given to neighbours

Note: Structure of table based on Lippert and Walker [9], p. 5.

Table 1 shows that the shadow economy includes unreported income from the production of legal goods and services, involving either monetary or barter transactions, hence all economic activities that would generally be taxable were they reported to the State tax authorities. A precise definition of the term seems quite difficult, if not impossible, as "the shadow economy develops all the time according to the 'principle of running water': it adjusts to changes in taxes, to sanctions from the tax authorities and to general moral attitudes, etc." (Mogensen and others [25], p. 5).** The subject of tax evasion or tax compliance, already the focus of considerable research [27], is beyond the scope of the present paper.

Size of the shadow economy in 76 countries

For single countries and sometimes for a group of countries, such as the States members of the Organization for Economic Cooperation and Development (OECD)

*This definition is used, for example, by Feige [5, 22] Schneider [10], Frey and Pommerehne [4], Lubell [23].

**For a detailed discussion, see Frey and Pommerehne [4], Feige [5], Thomas [6], Schneider [10, 13, 26].

or countries with economies in transition, research has been undertaken to estimate the size of the shadow economy (see Pozo [8], Loayza [7], Lippert and Walker [9], Schneider [13], Lackó [28]) using various methods and different time periods. In tables 2 to 4, an attempt is made to compare estimates of the size of the shadow economy in various countries over a fixed time period, using measurement methods described later in the present paper, in the section entitled "Methods used to estimate the size of the shadow economy",* with findings reported on the shadow economy in 76 countries throughout the world for the periods 1989-1990 and 1990-1993.**

For Central and South American countries, one estimate is made using the physical input method (Lackó [29]) and one using the MIMIC approach (Loayza [7]). For some countries, such as Brazil, Guatemala and Venezuela, estimates of the size of the shadow economy are quite similar; for others, such as Mexico, Panama and Peru, there are great differences. A ranking of the South American countries using the MIMIC approach shows that the biggest shadow economies, as a percentage of GDP, can be found in Bolivia, at 65.6 per cent of GDP, Panama, at 62.1 per cent, Peru, at 57.4 per cent, and Guatemala, at 50.4 per cent. Ranked lowest are the shadow economies of Costa Rica, at 23.2 per cent of GDP, Argentina, at 21.8 per cent, and Chile, at 18.2 per cent (all estimated for the period 1990-1993). In Asia, the shadow economy of Thailand is the largest, at 71 per cent of GDP, followed by the Philippines, at 50 per cent, and Sri Lanka, at 40 per cent. Hong Kong Special Administrative Region (SAR) of China and Singapore have the smallest shadow economies, each estimated at 13 per cent of GDP. The large size of the shadow economy in some developing countries suggests that it is more a "parallel" or second economy that has not been adequately captured by official statistics.

Transition countries

The physical input (electricity) method has been applied to the transition countries in central and eastern Europe and to the States that emerged from the breakup of the former Soviet Union. The results, shown in table 3, cover the periods 1989-1990, 1990-1993 and 1994-1995.*** According to the estimates based on the physical input method applied by Johnson and others [14] (shown together with values based on Lackó [28] for the States that constituted the former Soviet Union, during the period 1990-1993, Georgia had the largest shadow economy, at 43.6 (50.8) per cent of GDP, followed by Azerbaijan, at 33.8 (41) per cent, and the Republic of Moldova, at 29.1 per cent. The Russian Federation was in the middle range, with a shadow economy estimated at 27 (36.9) per cent of GDP. On the basis of Johnson and others, the shadow economies of Belarus, estimated at 14 per cent of GDP, and

*The physical input (electricity) and the currency demand methods are comparable because both assume an excessive use of a source (electricity and cash, respectively) for shadow-economy activities, and, in both, a "potential GNP" is calculated. The two methods are similarly used by Lackó [29-31], Portes [32], Johnson, Kaufmann and Zoido-Lobatón [15, 33], who have applied them to measure a series of shadow economies in a cross section of countries.

**It should be borne in mind that such country comparisons give only a very rough picture of the relative size of the shadow economy in different countries, because each method has its shortcomings. See, for example, Thomas [6, 18], Tanzi [17]. In the comparison presented, the same time periods (either 1989-1990 or 1990-1993) are used for all countries. If possible, the values were calculated as averages for each period.

***The results for the period 1989-1990, which was marked by the collapse of the communist regimes, can only be seen as rough approximations and are therefore not discussed in detail in the present paper.

Table 2. Size of the shadow economy in developing countries

(Percentage of GDP)

	Size of the shadow economy		
Developing countries	*Physical input (electricity) method (average 1989-1990)*	*Currency demand approach (average 1989-1990)*	*MIMIC[a] approach (average 1990-1993)*
Africa			
Botswana	27.0	..	..
Egypt	68.0	..	..
Mauritius	20.0	..	..
Morocco	39.0	..	..
Nigeria	76.0	..	..
South Africa	..	9.0[b]	..
Tunisia	45.0	..	..
United Republic of Tanzania	..	31.0[c]	..
Central and South America			
Argentina	..	..	21.8
Bolivia	..	..	65.6
Brazil	29.0	..	37.8
Chile	37.0	..	18.2
Colombia	25.0	..	35.1
Costa Rica	34.0	..	23.2
Ecuador	..	..	31.2
Guatemala	61.0	..	50.4
Honduras	..	..	46.7
Mexico	49.0	33.0[d]	27.1 (35.1)[d]
Panama	40.0	..	62.1
Paraguay	27.0	..	..
Peru	44.0	..	57.4
Uruguay	35.2	..	..
Venezuela	30.0	..	30.8
Asia			
China			
Hong Kong Special Administrative Region	13.0	..	..
Taiwan Province	..	..	16.5[e]
Cyprus	21.0	..	..
India	..	22.4[f]	..
Israel	29.0	..	..
Malaysia	39.0	..	..
Philippines	50.0	..	..
Republic of Korea	38.0	..	20.3[e]
Singapore	13.0	..	..
Sri Lanka	40.0	..	..
Thailand	71.0	..	..

Sources: Calculations based on Lackó ([29], table 18) for developing countries in Africa and Asia, and on Loayza [7] for those in Central and South America.

[a] Multiple-indicator multiple-cause.

[b] Based on Van der Berg [34], Hartzenburg and Leimann [35], who both used the currency demand approach.

[c] Based on Bagachwa and Naho ([36], p. 1394), who used the currency demand approach.

[d] Using the currency demand approach, Pozo [8] estimates the size of the shadow economy in Mexico, as a percentage of GDP, at 33 per cent in 1989-1990 and 35.1 per cent in 1990-1993.

[e] Estimated by Yoo and Hyun [37] using the income discrepancy method.

[f] Calculations based on Bhattacharyya [38].

Table 3. Size of the shadow economy in transition countries

(Percentage of GDP)

	Average size based on the physical input (electricity) method					
Transition countries	*1989-1990*		*1990-1993*		*1994-1995*	
States formerly part of the Soviet Union[a]						
Azerbaijan	21.9	(..)	33.8	(41.0)	59.3	(49.1)
Belarus	15.4	(..)	14.0	(31.7)	19.1	(45.4)
Estonia	19.9	(19.5)	23.9	(35.9)	18.5	(37.0)
Georgia	24.9	(..)	43.6	(50.8)	63.0	(62.1)
Kazakhstan	17.0	(13.0)	22.2	(29.8)	34.2	(38.2)
Kyrgyzstan	..	(13.9)	..	(27.1)	..	(35.7)
Latvia	12.8	(18.4)	24.3	(32.2)	34.8	(43.4)
Lithuania	11.3	(19.0)	26.0	(38.1)	25.2	(47.0)
Republic of Moldava	18.1	(..)	29.1	(..)	37.7	(..)
Russian Federation	14.7	(..)	27.0	(36.9)	41.0	(39.2)
Ukraine	16.3	(..)	28.4	(37.5)	47.3	(53.7)
Uzbekistan	11.4	(13.9)	10.3	(23.3)	8.0	(29.5)
Average for the States formerly part of the Soviet Union	16.7	(16.2)	25.7	(34.9)	35.3	(43.6)
Central and Eastern Europe						
Bulgaria	24.0	(26.1)	26.3	(32.7)	32.7	(35.0)
Croatia	22.8[b]	(..)	23.5[b]	(39.0)	28.5[b]	(38.2)
Czech Republic	6.4	(23.0)	13.4	(28.7)	14.5	(23.2)
Hungary	27.5	(25.1)	30.7	(30.9)	28.4	(30.5)
The former Yugoslav Republic of Macedonia	..	(..)	..	(40.4)	..	(46.5)
Poland	17.7	(27.2)	20.3	(31.8)	13.9	(25.9)
Romania	18.0	(20.9)	16.0	(29.0)	18.3	(31.3)
Slovakia	6.9	(23.0)	14.2	(30.6)	10.2	(30.2)
Slovenia	..	(26.8)	..	(28.5)	..	(24.0)
Average for the States of central and eastern Europe	17.6	(17.6)	20.6	(32.4)	20.9	(31.6)

Sources: Calculations based on Johnson and others ([14], table 1, pp. 182-183), Johnson and others ([15], p. 351) and, for data within parentheses, Lackó ([28], table 8).

[a] For the States formerly part of the Soviet Union, values calculated for 1990 were used as average values for 1989-1990, since no data for 1990 were available from Johnson and others [14].

[b] See Madzarevic and Milkulic ([41], table 9, p. 17), who used the discrepancy method.

Uzbekistan, at 10.3 per cent were the smallest. With the sole exception of the estimate for Uzbekistan based on Johnson and others, the shadow economy in all the other States born of the former Soviet Union experienced a strong increase from an average of 25.7 (34.9) per cent based on Lackó) for 1990-1993 to 35.3 (43.6) per cent based on Lackó) for 1994-1995. With regard to the transition countries of central and eastern Europe, the estimates based on Johnson and others for the period 1990-1993 show that Hungary has the largest shadow economy, at 30.7 per cent of GDP, followed by Bulgaria, at 26.3 per cent. The smallest are those of the Czech Republic, at 13.4 per cent of GDP, and Slovakia, at 14.2 per cent. On the basis of

Table 4. Size of the shadow economy in OECD countries

(Percentage of GDP)

	Size of the shadow economy			
	Physical input (electricity) method	*Currency demand method used by Schneider*		*Currency demand method used by Johnson and others*
OECD countries	*(1990)*	*(average 1989-1990)*	*(average 1990-1993)*	*(average 1990-1993)*
Australia	15.3	10.1	13.0	13.1
Austria	15.5	5.1	6.1	5.8
Belgium	19.8	19.3	20.8	15.3
Canada	11.7	12.8	13.5	10.0
Denmark	16.9	10.8	15.0	9.4
Finland	13.3	..	..	..
France	12.3	9.0	13.8	10.4
Germany	14.6	11.8	12.5	10.5
United Kingdom of Great Britain and Northern Ireland	13.1	9.6	11.2	7.2
Greece	21.8	..	..	27.2
Ireland	20.6	11.0	14.2	7.8
Italy	19.6	22.8	24.0	20.4
Japan	13.2	..	..	8.5
Netherlands	13.4	11.9	12.7	11.8
New Zealand[a]	..	9.2	9.0	9.0
Norway	9.3	14.8	16.7	5.9
Portugal	16.8	..	..	15.6
Spain[b]	22.9	16.1	17.3	16.1
Sweden	11.0	15.8	17.0	10.6
Switzerland	10.2	6.7	6.9	6.9
United States of America	10.5	6.7	8.2	13.9
Average for 21 OECD countries	15.1	11.9	13.5	11.3

Sources: Calculations using the physical input method (Lackó [28-31]) and the currency demand approach (Schneider [10, 13], Johnson and others [15, 33] and Williams and Windebank [40]).

[a] Calculations using the MIMIC method and the currency demand approach (Giles [20]).

[b] Calculations based on Mauleon [42].

the Lackó estimates, the former Yugoslav Republic of Macedonia has the largest shadow economy, at 40.4 per cent of GDP, followed by Croatia, at 39 per cent. The Lackó estimates show that the smallest shadow economies were those of Slovenia, at 28.5 per cent of GDP, and the Czech Republic, at 28.7 per cent. Whereas a strong increase was observed in the shadow economy of States that were part of the former Soviet Union for the periods 1990-1993 and 1994-1995, the average size of the shadow economy in States of central and eastern Europe was almost stable during those periods. The estimates of Johnson and others show that the shadow economy in States of central and eastern Europe averaged 20.6 per cent of GDP (32.4 per cent according to Lackó) over the period 1990-1993 and 20.9 per cent (31.6 per cent according to Lackó) over the period 1994-1995.

States members of the Organisation for Economic Cooperation and Development

Either the currency demand method or the physical input method was applied to each of the 21 States members of the Organisation for Economic Cooperation and Development (OECD). Two series of figures are based on the currency demand method, one from Schneider [10, 13] and one from Johnson and others [15, 33].* The series by Johnson and others, involving estimates of the shadow economy in most OECD countries (20 out of the 21 countries investigated) over the period 1990-1993, shows that the southern European countries had the largest shadow economies as a percentage of GDP: Greece (27.2 per cent), Italy (20.4 per cent), Spain (16.1 per cent) and Portugal (15.6 per cent). A similar result can be found by using estimates from Schneider and, to a much lesser extent, those provided by Lackó [31] through the physical input (electricity) method. Ranked at the lower end by Johnson and others are Switzerland (6.9 per cent), Norway (5.9 per cent) and Austria (5.8 per cent); whereas Schneider finds the United States of America (8.2 per cent), Switzerland (6.9 per cent) and Austria (6.1 per cent) at the bottom. The ranking of the size of the shadow economy in OECD countries by Schneider is supported by other studies.**

In table 5, OECD averages are shown for 1994-1995 and for 1996-1997. In principle, the ranking of shadow economies by size is similar to that of table 4. However, the size of the shadow economy in all OECD countries has increased compared to the results for 1990-1993. Whereas the average size of the shadow economy in the OECD countries studied was 13.5 per cent of GDP in 1990-1993, it had increased to 16 per cent of GDP in 1994-1995. It increased further to 16.9 per cent in 1996-1997. The findings clearly show that even in the late 1990s the shadow economy was still growing in most OECD countries.

Average size of the shadow economy in developing countries, transition countries and OECD countries

A comparison of the average size of shadow economies in the three major country groupings yields the results shown in table 6.

Only a crude comparison can be made of the size of the shadow economy in the various countries and country groupings because, in the studies conducted: *(a)* different independent variables (such as tax variables) and different specifications for the dependent variable and the relevant equations were used; *(b)* different assumptions about the velocity of currency circulation were made; and *(c)* other factors affecting electricity consumption were taken into account. As can be seen from table 6, the average size of the shadow economy in developing countries is by far the largest, at between 35 and 44 per cent of GDP, followed by the transition countries, at between 20.7 per cent and 34.9 per cent, and finally the OECD countries, estimated at 15.1 per cent using the electricity method and at 11.9 per cent

*The main difference between the two series is that, given a monetary approach, Johnson and others use average values, coming from different sources, of the size of the shadow economy of a country, whereas in Schneider, the currency demand approach and only one value for a given year (or an average over a time period) are used. The problem with using averages from various sources is that: the time period is greater (1985-1995); and the monetary approaches specified by different authors may be quite different.

**Similar rankings are established by Frey and Pommerehne [4], Frey and Weck-Hannemann [39], Williams and Windebank [40], Thomas [6] and Lippert and Walker [9].

Table 5. Size of the shadow economy in OECD countries, 1994-1997

(Percentage of GDP)

OECD countries	*Size of the shadow economy based on the currency demand approach* *Average 1994-1995*[a]	*Average 1996-1997*[a]
Australia	13.8	13.9
Austria	7.0	8.6
Belgium	21.5	22.2
Canada	14.8	14.9
Denmark	17.8	18.2
France	14.5	14.8
Germany	13.5	14.8
United Kingdom of Great Britain and Northern Ireland	12.5	13.0
Greece	29.6	30.1
Ireland	15.4	16.0
Italy	26.0	27.2
Japan	10.6	11.3
Netherlands	13.7	13.8
New Zealand	11.31[a]	..
Norway	18.2	19.4
Portugal	22.1	22.8
Spain	22.4	23.0
Sweden	18.6	19.5
Switzerland	6.7	7.8
United States of America	9.2	8.8
Average for 20 OECD countries	16.0	16.9

Sources: Calculations based on Schneider [13] and Schneider and Pöll [43].

[a]Calculated only for 1994, based on Giles [20].

Table 6. Average size of the shadow economy in developing countries, transition countries and OECD countries

Country grouping and measurement method used	*Average for 1989-1993 (percentage of GDP)*	
Developing countries (electricity method used)		
Africa	43.9	(39.4)[a]
Asia	35.0	
Central and South America	38.9	
Transition countries (*electricity method used*)		
Central and Eastern Europe	20.7	(32.4)[b]
States formerly part of the Soviet Union	25.7	(34.9)[b]
OECD countries		
Electricity method used	15.1	
Currency demand method used	11.9	

Source: Calculations using tables 2-4 above.

[a]Including South Africa.

[b]Based on values from Lackó [28] for 1990-1993.

using the currency demand method. But such a comparison, as noted above, can only be indicative, since the methods, statistical approaches and specifications used differ widely in the various studies.

Some remarks on the labour force in a shadow economy

After the review given above of the size and growth of the shadow economy in terms of value added over time, the focus on the present section will be on the "shadow labour market", since within the official labour market there is particularly close contact between those who are active in the shadow economy.* Moreover, by definition, every shadow economic activity to some extent involves a shadow labour market. Hence, the shadow labour market includes all cases where the employees or employers, or both, occupy a position in the shadow economy. Why do people work in the shadow economy? In the official labour market, the costs that firms (and individuals) have to pay when officially hiring someone are greatly increased by the burden of taxation and social contributions linked to wages, as well as by legal and administrative regulations to control economic activity.** In various OECD countries, such costs are greater than the wage effectively earned by the worker, thus providing a strong incentive to work in the shadow economy. The underground use of labour may involve a second job after (or even during) regular working hours. Another form of work in the shadow economy is carried out by individuals who do not participate in the official labour market. A third component consists in the employment of people (for example, clandestine or illegal immigrants) who are not allowed to work in the official economy.

Studying the labour market in the shadow economy is even more difficult than studying value added in the shadow economy, because very little is known about how many hours a shadow economy worker actually works on average (from full time to only a few hours). Empirical facts are therefore not easy to come by. The few estimates available are shown in table 7 for OECD countries.*** The figures in table 7 give a rough idea of the size of the shadow labour market. For example, the estimates for Denmark show that the population of adult Danes engaged in the shadow economy ranged from 8.3 per cent of the total labour force in 1980 to 15.4 per cent in 1994. In Germany, the figure rose from 8 to 12 per cent during the period 1974-1982, to 22 per cent in 1997-1998.

That is a very strong increase for both countries. The size of the labour force in the shadow economy is also quite large in other countries: in Italy, 30-48 per cent (1997-1998); in Spain, 11.5-32.3 per cent (1997-1998); in Sweden, 19.8 per cent (1997-1998); and in France, 6-12 per cent (1997-1998). In the European Union, at least 10 million people are engaged in shadow-economy activities, and in the OECD countries, about 16 million work on an illicit, irregular or unofficial basis. Those figures demonstrate that the labour market in the shadow economy is thriving, and may explain, for example, why there is such high and persistent unemployment in Germany.

*Pioneering work in this area has been done by L. Frey [44-47], Cappiello [48], Lubell [23], Pozo [8], Bartlett [49] and Tanzi [17].

**This is especially true in Europe (for example, in Germany and Austria), where the total tax and social security burden adds up to 100 per cent of the wage effectively earned. See the section below on the increase in the burden of taxation and social security contributions.

***For developing countries, the literature about the shadow labour market includes Dallago [50], Pozo [8], Loayza [7] and especially Chickering and Salahdine [51].

Table 7. Size of the labour force in the shadow economy of selected OECD countries, 1974-1998

(percentage of GDP)

Country or economic grouping	*Year*	*Participants*[a] *(thousands)*	*Participants*[b] *(percentage of labour force)*	*Size of the shadow economy*[c] *(percentage of GDP)*	*Source of estimates of the level of participation*
Austria	1990-1991	300	9.6	5.47	Schneider [13, 21]
	1997-1998	500	16.0	8.93	
Denmark	1980	..	8.3	8.6	Mogensen and others [25]
	1986	..	13.0	..	
	1991	..	14.3	11.2	
	1994	..	15.4	17.6	
France	1975-1982	800-1 500	3.0-6.0	6.9	Raffaele de Grazia, Le travail clandestin: situation dans les pays industrialisés à économie de marché (Geneva, International Labour Organization, 1983) and author's calculations
	1997-1998	1 400-3 200	6.0-12.0	14.7	
Germany	1974-1982	2 000-3 000	8.0-12.0	10.6	de Grazia, op. cit. and Schneider [21]
	1997-1998	5 000	22.0	14.7	
Italy	1979	4 000-7 000	20.0-35.0	16.7	D. Gaitani and G. d'Aragona, "I. Somersi", *Nord e Sud*, vol. 26, No. 7 (1979), pp. 26-46 and author's calculations
	1997	6 600-11 400	30.0-48.0	27.3	
Spain	1979-1980	1 250-3 500	9.6-26.5	19.0	Author's calculations
	1997-1998	1 500-4 200	11.5-32.3	23.1	
Sweden	1978	750	13.0-14.0	13.0	de Grazia, op. cit. and author's calculations
	1997	1 150	19.8	19.8	
European Union	1978	10 000	..	14.5	de Grazia, op. cit. and author's calculations
	1997-1998	20 000			
OECD	1978	16 000	..	15.0	de Grazia, op. cit. and author's calculations
	1997-1998	35 000			

[a] Estimated number of persons holding full-time jobs, including unregistered workers, illegal immigrants and those holding second jobs.

[b] Population aged 20-69; in Denmark, those heavily engaged in shadow economy activities.

[c] Based on the currency demand method. Source of data on the size of the shadow economy: Friedrich Schneider [10, 12] and "Ist Schwarzarbeit ein Volkssport geworden? Ein internationaler Vergleich des Ausmaßes der Schwarzarbeit von 1970-97", in *Der Sozialstaat zwischen Markt und Hedonismus*, Siegfried Lamnek and Jens Luedtke, eds. (Obladen, Germany, Leske und Budrich, 1999), pp. 293-318.

More detailed information on the labour supply in the underground economy is given by Lemieux and others [52] using data from a survey conducted in the city of Quebec, Canada. In particular, their study provides some economic insight into the size of the distortion caused by income taxes and the welfare system. The findings of the study suggest that hours worked in the shadow economy are responsive to changes in net wages in the regular (official) sector. The study also provides some support for the existence of a Laffer curve. The Laffer curve suggests that an increase in the marginal tax rate leads to a decrease in tax revenue when the tax rate is too high. According to the empirical findings of the study, that is attributable to a misallocation of work from the official to the informal sector, where it is not taxed. In that case, the exchange of labour market activities between the two sectors is high. The empirical findings clearly indicate that "participation rates and hours worked in the underground sector also tend to be inversely related to the number of hours worked in the regular sector" (Lemieux and others, [52], p. 235). The findings demonstrate a large negative elasticity of hours worked in the shadow economy with respect to the wage rate in the regular sector and also a high mobility between the sectors.

Main causes of the growth of the shadow economy

Increase in the burden of taxation and social security contributions

In almost all studies,* it has been found that an increase in taxes and social security contributions is one of the main causes of the growth of the shadow economy. Since taxes affect choices of labour and leisure and also stimulate the labour supply in the shadow economy, or the untaxed sector of the economy, the distortion of that choice is a major concern of economists. The bigger the difference between the total cost of labour in the official economy and the after-tax earnings from work, the greater is the incentive to avoid the loss by working in the shadow economy. Since the difference depends broadly on the social security system and the overall tax burden, they are key features of the existence and growth of the shadow economy. But even tax reforms with major deductions in the tax rate deductions will not lead to a substantial decrease in the size of the shadow economy. They will only be able to stabilize the size of the shadow economy and avoid a further increase. Social networks and personal relationships, the high profit from irregular activities and associated investments in real and human capital are strong ties that prevent people from transferring to the official economy. For Canada, Spiro [54] expected similar reactions of people facing an increase in indirect taxes (value added tax (VAT) and goods and services tax (GST)). After the introduction of GST in 1991, in the midst of a recession, people suffering economic hardship because of the recession turned to the shadow economy, which led to a substantial loss in tax revenue. "Unfortunately, once this habit is developed, it is unlikely that it will be abandoned merely because economic growth resumes." (Spiro [54], p. 255). They may not return to the formal sector, even in the long run. That makes it even more difficult for politicians to carry out major reforms, because they may not gain much from them.**

*See, for example, Thomas [6], Lippert and Walker [9], Schneider [10-13, 21, 53], Johnson and others [15, 33], Tanzi [17] and Giles [19].

**See Schneider [11, 21] for similar findings on the effects of a major tax reform in Austria on the shadow economy. Schneider shows that a major reduction in the direct tax burden did not lead to a major reduction in the shadow economy. Because legal tax avoidance was abolished and other factors, like regulations, were not changed, for many taxpayers the actual tax and regulation burden remained unchanged.

The most important factor in neoclassical models is the marginal tax rate. The higher the marginal tax rate, the greater is the substitution effect and the bigger the distortion of the decision between labour and leisure. Especially when taking into account that the individual can also receive income in the shadow economy, the substitution effect is definitely larger than the income effect* and, hence, the individual works less in the official sector. The overall efficiency of the economy is therefore, ceteris paribus, lower, and the distortion leads to a welfare loss (based on official GNP and taxation). But the welfare might also be viewed as increasing, if the welfare of those who are working in the shadow economy were taken into account, too (Thomas [6], pp. 134-137).

Empirical results of the influence of the tax burden on the shadow economy are provided in the studies of Schneider [11, 53] and Johnson and others [15, 33]. They all found strong evidence for the general influence of taxation on the shadow economy. This strong influence of indirect and direct taxation on the shadow economy will be further demonstrated by showing empirical results for Austria and the Scandinavian countries. In the case of Austria, Schneider [11] estimates a currency demand function including as driving forces for the shadow economy the following four types of variables:

(a) The burden of total direct taxation;

(b) The burden of indirect taxation;

(c) The complexity of the tax system;

(d) The intensity of government regulation.

The results of the application of the currency demand function are shown in table 8.

All coefficients of the independent variables have the theoretically expected sign and, with the exception of the indirect tax burden, are statistically significant at the 95 per cent confidence level. The other test statistics show satisfactory results, In particular, the “true ex-post” forecast of currency demand for the period 1985-1991 indicates that the major independent factors in the currency demand function are included. The driving force for the shadow economy activities is the direct tax burden (including social security payments), which has the biggest impact, followed by the intensity of regulation and the complexity of the tax system. A similar result has been achieved by Schneider [26] for Scandinavian countries (Denmark, Norway and Sweden). In all three countries, various tax variables (average direct tax rate and average total tax rate (consisting of indirect and direct tax rates)) and marginal tax rates have the expected positive sign (on currency demand) and are highly significant statistically. Similar results were reached by Kirchgaessner [55, 56] for Germany and by Kloveland [57] for Norway and Sweden.

Several other recent studies provide further evidence of the influence of income tax rates on the shadow economy. Cebula [58], using Feige data for the shadow economy, found evidence of the impact of government income tax rates, Internal Revenue Service (IRS) audit probabilities and IRS penalty policies on the relative size of the shadow economy in the United States. Cebula concludes that refraining from any further increase of the top marginal income tax rate may at least curb a further increase in the shadow economy, while increased IRS audits and penalties

*If leisure is assumed to be a normal good.

Table 8. Results of the application of the currency demand function to Austria

	Dependent variable: real currency per capita, ln (CUR_t/POP_t) estimation period,	
Independent variables	*1956-1991*	*1956-1985*
Lagged dependent variable ln (CUR_{t-1} /POP_{t-1})	0.534** (8.91)	0.551** (9.43)
Real consumption per capita ln (C_t /POP_t)	0.703** (5.49)	0.724** (5.99)
Number of Eurocheque systems per capita ln (ES_{t-1} /POP_{t-1})	-0.213* (-2.51)	-0.174* (-2.09)
Real interest rate on bonds ln (IR_t)	-0.123* (-2.51)	-0.139* (-2.65)
Direct tax burden (including social security payments) ln ($DIRT_t$)	0.173** (3.09)	0.182* (2.86)
Indirect tax burden ln ($INDT_t$)	0.117(*) (1.88)	0.123(*) (1.92)
Complexity of the tax system ln ($VIST_t$)	0.154** (2.77)	0.147** (2.86)
Intensity of regulation ln (REG_t)	0.166** (2.94)	0.159** (2.72)
Constant term	-2.24(*) (-1.80)	-2.39(*) (-1.74)
Test statistics		
R^2	0.992	0.990
S.E.	0.014	0.015
Durbin's h	1.06	1.16
rho (1)	0.18	0.20
D.F.	27	21
Ex-post Forecast 1985-1991		
RMSE	-	1.51
Theil's U 1	-	0.42

Note: All equations are estimated by an ordinary least-squares procedure using annual data. R^2 is the coefficient of determination (corrected for the degrees of freedom); S.E. shows the standard error of the estimation. Durbin's h is Durbin's h-test against auto-correlation when lagged dependent variables are used as regressors. Rho (1) is the auto-correlation coefficient of first order. D.F. stands for the "degrees of freedom". RMSE is the root mean squared error and Theil's U 1 stands for Theil's inequality coefficient. The term "ln" indicates that these variables have been transformed to natural logarithms. Numbers in parentheses below coefficient estimates are t-values. (*),* and** indicate significance at the 90, 95 and 99 per cent confidence level, respectively.

might reduce the size of the shadow economy. His findings indicate that there is generally a strong influence of State activities on the size of the shadow economy. For example, if the marginal federal personal income tax rate increases by one percentage point, ceteris paribus, the shadow economy rises by 1.4 percentage points. In another investigation, Hill and Kabir [59] found empirical evidence that marginal tax rates are more relevant than average tax rates, and that a substitution of direct taxes by indirect taxes seems unlikely to improve tax compliance. Further evidence of the effects of taxation on the shadow economy is presented by Johnson and others [33], who come to the conclusion that it is not higher tax rates per se that increase the size of the shadow economy, but the ineffective and discretionary application of the tax system and government regulations. Their finding that there is a negative correlation* between the size of the unofficial economy and the top (marginal) tax rates might be unexpected. But since other factors like tax deductibility, tax relief, tax exemptions, the choice between different tax systems and various other options for legal tax avoidance were not taken into account, it is no great surprise.** On the other side, Johnson and others [33] find a positive correlation between the size of the shadow economy and the corporate tax burden. They come to the overall conclusion that there is a large difference between the impact of either direct taxes or the corporate tax burden. Institutional aspects, like the efficiency of the administration, the extent of control rights held by politicians and bureaucrats, and the amount of bribery and especially corruption, therefore play a major role in the bargaining game between the Government and the taxpayers.

Intensity of regulations

The increase of the intensity of regulation (often measured by the number of laws and regulations, such as licence requirements) is another important factor that reduces the freedom of choice for individuals engaged in the official economy.*** Labour market regulations, trade barriers and labour restrictions for foreigners are relevant examples. Johnson and others [33] find overall significant empirical evidence of the influence of (labour) regulations on the shadow economy, and the impact is clearly described and theoretically derived in other studies, for example, the study on Germany carried out by the Deregulation Commission in 1990-1991. Regulations lead to a substantial increase in labour costs in the official economy. But since most of the costs can be shifted on to the employees, they provide another incentive to work in the shadow economy, where they can be avoided. Empirical evidence supporting the model of Johnson and others [14], which predicts, inter alia, that countries with more general regulation of their economies tend to have a higher share of the unofficial economy in total GDP, is found in their empirical analysis. A one-point increase of the regulation index (ranging from 1 to 5, with 5 corresponding to the most regulation in a country), ceteris paribus, is associated with an 8.1 percentage point increase in the share of the shadow economy, when controlled for GDP per capita (Johnson and others [33], p. 18). They conclude that it is the enforcement of regulations, which is the key factor in the burden levied on firms and individuals, and not the overall extent of regulations, mostly not enforced, that drives firms into the shadow economy. Friedman and others [60] reach a similar result. In their study, every available measure of regulation is significantly correlated with the share of the unofficial economy and the sign of the relationship is unambiguous: more regulation is correlated with a larger shadow economy. A one-point increase in an index of regulation (ranging from 1 to 5) is associated with a 10 per cent increase in the shadow economy for 76 developing countries, transition countries and developed countries.

*The higher the top marginal tax rate, the lower the size of the shadow economy.

**Friedman and others [60] found a similar result in a cross-country analysis showing that higher tax rates are associated with less official activity as a percentage of GDP. They argue that entrepreneurs go underground not to avoid official taxes but to reduce the burden of bureaucracy and corruption. However, considering their empirical (regression) results, the finding that higher tax rates are correlated with a lower share of the unofficial economy is not very robust, and in most cases, using different tax rates, they do not find a statistically significant result.

***For a (social) psychological and theoretical foundation of this feature, see Brehm [61, 62], and for a first application to the shadow economy, see Pelzmann [63].

The above-mentioned findings demonstrate that Governments should put more emphasis on improving enforcement of laws and regulations, rather than increasing their number. Some Governments, however, prefer this policy option (more regulations and laws) when trying to reduce the shadow economy, mostly because it leads to an increase in power of the bureaucrats and to a higher rate of employment in the public sector.*

Social transfers

The social welfare system leads to strong negative incentives for beneficiaries to work in the official economy since their marginal tax rate often equals or nearly reaches 100 per cent. That can be derived either from the neoclassical leisure-income model or from empirical results.** Such a system provides major disincentives for individuals who are getting welfare payments to even search for work in the official economy, since their overall income is much higher when they are still receiving the transfers, while possibly working in the underground economy.

Public sector services

An increase in the size of the shadow economy leads to reduced State revenues, which in turn reduces the quality and quantity of publicly provided goods and services. Ultimately, this can lead to an increase in the tax rates for firms and individuals in the official sector, quite often combined with a deterioration in the quality of the public goods (such as the public infrastructure) and of the administration, with the consequence of even stronger incentives to participate in the shadow economy. Johnson and others [33] present a simple model of this relationship. Their findings show that smaller shadow economies appear in countries with higher tax revenues, if achieved by lower tax rates, fewer laws and regulations and less bribery facing enterprises. Countries with a better rule of the law that is financed by tax revenues also have smaller shadow economies. Transition countries have higher levels of regulation leading to a significantly higher incidence of bribery, higher effective taxes on official activities and a large discretionary framework of regulations, and consequently to a higher shadow economy. The overall conclusion is that "wealthier countries of the OECD, as well as some in eastern Europe, find themselves in the 'good equilibrium' of relatively low tax and regulatory burden, sizeable revenue mobilization, good rule of law and corruption control, and [relatively] small unofficial economy. By contrast, a number of countries in Latin America and the former Soviet Union exhibit characteristics consistent with a 'bad equilibrium': tax and regulatory discretion and burden on the firm is high, the rule of law is weak, and there is a high incidence of bribery and a relatively high share of activities in the unofficial economy" ([15], p. 388).

Effects of the shadow economy on the official economy

In order to study the effects of the shadow economy on the official economy, several studies integrate underground economies into macroeconomic models.***

*See, for example, Frey [64] for a first application of the public choice theory to the shadow economy.

**See, for example, Lemieux and others [52].

***For Austria, this was done by Schneider and others [68] and Neck and others [69]. For further discussion of this aspect, see Quirk [70] and Giles [19].

Houston [65] develops a theoretical macroeconomic model of the business cycle as well as tax and monetary policy linkages with the shadow economy. He concludes from his investigation of the growth of the shadow economy that, on the one hand, its effect should be taken into account in setting tax and regulatory policies, and, on the other, the existence of a shadow economy could lead to an overstatement of the inflationary effects of fiscal or monetary stimuli. Adam and Ginsburgh [66] focus on the implications of the shadow economy on official growth in their study for Belgium. They find a positive relationship between the growth of the shadow economy and the official one, and under certain assumptions (that is, very low costs of entry into the shadow economy due to a low probability of enforcement), they conclude that an expansionary fiscal policy has a positive stimulus for both the formal and informal economies. A study for the United States by Fichtenbaum [67] argues that the United States productivity slowdown over the period 1970 to 1989 was vastly overstated, as the underreporting of income due to the more rapid growth of the United States shadow economy during that period was not taken into account.*

Another hypothesis is that a substantial reduction of the shadow economy leads to a significant increase in tax revenues, and therefore to a greater quantity and quality of public goods and services, which ultimately can stimulate economic growth. Some authors found evidence for that hypothesis. A recent study by Loayza [7] presents a simple macroeconomic endogenous growth model whose production technology depends on congestable public services. The determinants and effects of the informal sector are studied, where excessive taxes and regulations are imposed by Governments and where the capability to enforce compliance is low. The model leads to the conclusion that in economies where the statutory tax burden is larger than the optimal tax burden, and where the enforcement of compliance is too weak, the increase in the relative size of the informal economy generates a reduction of economic growth. The linkage is due to the strongly negative correlation between the informal sector and public infrastructure indices, with public infrastructure being the key element for economic growth. For example, Loayza finds empirical evidence that, in Latin American countries, if the shadow economy increases by one percentage point of GDP, ceteris paribus, the growth rate of official real GDP per capita decreases by 1.22 percentage points. This negative impact of informal sector activities on economic growth is not broadly accepted.** For example, the key feature of the model has been criticized, because the model is based on the assumption that the production technology essentially depends on tax-financed public services, which are subject to congestion. In addition, the informal sector is not paying any taxes, but must pay penalties that are not used to finance public services. The negative correlation between the size of the informal sector and economic growth is therefore not very surprising.

Depending on the prevailing view of the informal sector, the opposite conclusion might be reached. In the neoclassical view, the underground economy is optimal in the sense that it responds to the demand of the economic environment for urban services and small-scale manufacturing. From this point of view, the informal sector provides the economy with a dynamic and entrepreneurial spirit and can lead to more competition, higher efficiency and strong boundaries and limits for government activities. The informal sector may offer great contributions “to the creation of markets, increase financial resources, enhance entrepreneurship, and transform the legal, social, and economic institutions necessary for accumulation” (Asea [72],

*Compare also the findings of Pommerehne and Schneider [71], who come to similar conclusions.

**See Asea [72] for a more detailed criticism of the Loayza model.

p. 166). The voluntary self-selection between the formal and informal sectors, as described above in microeconomic models, may provide a higher potential for economic growth and, hence, a positive correlation between an increase in the informal sector and economic growth. The effects of an increase in the size of the shadow economy on economic growth therefore remain highly ambiguous.

The empirical evidence of the hypotheses is also not clear. On the one hand, since many Latin American countries had or still have a tradition of excessive regulations and weak government institutions, Loayza [7] finds some evidence of the implications of his growth model during the early 1990s in those countries. The increase in the size of the shadow economy negatively affects official GDP growth by reducing the availability of public services for everyone in the economy, and by causing the existing public services to be used less efficiently, or not at all. On the other hand, the positive effects of shadow economy activities should also be considered. Empirical findings of Schneider [21] show clearly that over 66 per cent of the earnings in the shadow economy are immediately spent in the official sector. The positive effects of this expenditure for economic growth and for the indirect tax revenues must be taken into account as well. Bhattacharyya [38] found clear evidence that, in the United Kingdom (1960-1984), the hidden economy had a significant positive effect on consumer expenditures in the official economy. He points out that the hidden economy has a positive effect on consumer expenditure for non-durable goods and services, and an even stronger positive effect on consumer expenditure for durable goods and services.*

Methods used to estimate the size of the shadow economy

As has already been mentioned above in the section on the definition of the shadow economy, the attempt to measure the size of a shadow economy is a difficult and challenging task. In the present section, a comprehensive overview is given of the current knowledge of the various procedures for estimating the size of the shadow economy. To measure the size and development of the shadow economy, three different types of methods are most widely used.** They are briefly discussed in the next three sections.

Direct approaches

There are micro-approaches that employ either well-designed surveys and samples based on voluntary replies or tax auditing and other compliance methods. Sample surveys designed for estimating the size of the shadow economy are widely used in a number of countries.*** The main disadvantage of such a method is that it presents the flaws of all surveys: the precision of averages and results depend greatly on the willingness of respondents to cooperate. It is difficult to assess the rise of undeclared work from a direct questionnaire. Most interviewees hesitate to confess fraudulent behaviour and responses are seldom reliable, so that it is difficult to

*A close interaction between official and unofficial economies is also emphasized in Giles [19] and in Tanzi [17].

**The discussion below closely follows Schneider and Enste [1]. See also Frey and Pommerehne [4], Feige [5], Thomas [6, 18] and Schneider [10, 13, 26].

***The direct method of voluntary sample surveys has been extensively used for Norway by Isachsen and others [73] and Isachsen and Strom [74]. For Denmark, this method is used by Mogensen and others [25], who report estimates of the shadow economy at 2.7 per cent of GDP for 1989, 4.2 per cent for 1991, 3 per cent for 1993 and 3.1 per cent for 1994.

calculate a true estimate, in monetary terms, of the amount of undeclared work. The main advantage of the method lies in the detailed information that it provides about the structure of the shadow economy, but the results from such surveys are highly sensitive to the way in which the questionnaire is formulated.*

Estimates of the shadow economy can also be based on the discrepancy between income declared for tax purposes and that measured by selective checks. Fiscal auditing programmes have been particularly effective in that regard. Designed to measure the amount of undeclared taxable income, they have been used to calculate the size of the shadow economy in several countries.** A number of difficulties beset this approach. First, using tax compliance data is equivalent to using a possibly biased sample of the population. However, since a selection of taxpayers for auditing is generally not random, but based on properties of submitted tax returns that indicate a certain likelihood of fraud, such a sample is not a random one of the whole population. That factor is likely to bias compliance-based estimates of the shadow economy. Secondly, estimates based on tax audits reflect the portion of shadow economy income that the authorities succeeded in discovering, and are likely to be only a fraction of hidden income.

A further disadvantage of the two direct methods (surveys and tax auditing) is that they lead to only point estimates. Moreover, since it is unlikely that they capture all shadow economy activities, they can be seen as providing estimates at the lower end of the scale. They are currently unable to provide estimates of the development and growth of the shadow economy over a longer period of time. As already noted, however, they have at least one considerable advantage: they can provide detailed information about the structure of shadow economy activities and about those who work in the shadow economy.

Indirect approaches

Indirect approaches, which are also called “indicator” approaches, are mostly macroeconomic measures that use various economic and other indicators containing information about the development of the shadow economy over time. Currently, there are five indicators that leave some traces of the development of the shadow economy, as described below.

Discrepancy between national expenditure and income statistics

One approach is based on discrepancies between income and expenditure statistics. In national accounting, the income measure of GNP should be equal to the expenditure measure of GNP. Thus, if an independent estimate of the expenditure site of the national accounts is available, the gap between the expenditure measure and the income measure can be used as an indicator of the scale of the shadow economy.*** However, since national accounts statisticians will be anxious to minimize this discrepancy, the initial discrepancy or first estimate, rather than the pub-

*The advantages and disadvantages of this method are extensively dealt with by Mogensen and others [25].

**For the United States, see IRS [75, 76], Simon and Witte [77], Witte [78], Clotefelter [79] and Feige [80]. For a more detailed discussion, see Dallago [50] and Thomas [6].

***See, for example, Franz [81], for Austria; MacAfee [82], O’Higgins [83] and Smith [24], for the United Kingdom; Petersen [84] and Del Boca [85], for Germany; and Park [86], for the United States. For a survey and critical remarks, see Thomas [6].

lished discrepancy, should be employed for this purpose. If all the components of the expenditure site were measured without error, then this approach would indeed yield a good estimate of the scale of the shadow economy. Unfortunately, that is not the case, and the discrepancy therefore reflects all omissions and errors everywhere in the national accounts statistics as well as the shadow economy activity. The estimates may therefore be very crude and of questionable reliability.*

The discrepancy between the official and actual labour force

A decline in the participation of the labour force in the official economy can be seen as an indication of increased activity in the shadow economy. If the participation of the total labour force is assumed to be constant, a decreasing official rate of participation can be seen as an indicator of an increase in the activities of the shadow economy, ceteris paribus.** The weakness of this method is that differences in the rate of participation may also have other causes. Moreover, people can work in the shadow economy and have a job in the official economy. Such estimates may therefore be viewed as weak indicators of the size and development of the shadow economy.

Transactions approach

This approach has been developed by Feige.*** It assumes that there is a constant relation over time between the volume of transactions and official GNP. Feige's approach therefore starts from Fisher's quantity equation,

M*V = p*T where M = money, V = velocity, p = prices
and T = total transactions.

Assumptions have to be made about the velocity of money and about the relationships between the value of total transactions (p*T) and total (official + unofficial) nominal GNP. Relating total nominal GNP to total transactions, the GNP of the shadow economy can be calculated by subtracting the official GNP from total nominal GNP. However, to derive figures for the shadow economy, Feige has to assume a base year in which there is no shadow economy. Therefore, the ratio of p*T to total nominal (official = total) GNP was "normal" and would have been constant over time, if there had been no shadow economy. This method, too, has several weaknesses, for instance, the assumption of a base year with no shadow economy and the assumption of a normal ratio of transactions constant over time. Moreover, to obtain reliable estimates of the shadow economy, precise figures of the total volume of transactions should be available. This availability might be especially difficult to achieve for cash transactions, because they depend, among other factors, on the durability of bank notes, in terms of the quality of the paper on which they are printed.**** Also, in this approach, the assumption is made that all vari-

*A related approach is pursued by Pissarides and Weber [87], who use microdata from household budget surveys to estimate the extent of income understatement by the self-employed. In this micro-approach, more or less the same difficulties arise and the figures calculated for the shadow economies may be crude.

**Such studies have been made, for example, by Contini [88] and Del Boca [85], for Italy; and by O'Neill [89], for the United States. For a survey and critical remarks, see Thomas [6].

***For an extended description of this approach, see Feige [5, 90, 91]. For a further application for the Netherlands, see Boeschoten and Fase [92], and for Germany, see Langfeldt [93].

****For a detailed criticism of the transaction approach, see Boeschoten and Fase [92], Frey and Pommerehne [4], Kirchgaessner [56], Tanzi [2, 94], Dallago [50], Thomas [6, 18, 95] and Giles [19].

ations in the ratio between the total value of the transactions and the officially measured GNP are due to the shadow economy. This means that a considerable amount of data is required in order to eliminate financial transactions from “pure” cross payments, which are totally legal and have nothing to do with the shadow economy. In general, although this approach is theoretically attractive, the empirical requirements necessary to obtain reliable estimates are so difficult to fulfil that its application may lead to doubtful results.

Currency demand approach

The currency demand approach was first used by Cagan [96], who calculated a correlation of the currency demand and the tax pressure (as one cause of the shadow economy) for the United States over the period 1919 to 1955. Twenty years later, Gutmann [97] used the same approach, but did not use any statistical procedures. Instead, he only looked at the ratio between currency and demand deposits over the years 1937 to 1976.

Cagan’s approach was further developed by Tanzi [98, 99], who econometrically estimated a currency demand function for the United States for the period 1929 to 1980 in order to calculate the shadow economy. His approach assumes that shadow (or hidden) transactions are undertaken in the form of cash payments, so as to leave no observable traces for the authorities. An increase in the size of the shadow economy will therefore increase the demand for currency. To isolate the resulting excess demand for currency, an equation for currency demand is econometrically estimated over time. All possible conventional factors, such as the development of income, payment habits and interest rates, are subject to control.

Additionally, such variables as the direct and indirect tax burden, government regulations and the complexity of the tax system, which are assumed to be the major factors causing people to work in the shadow economy, are included in the estimation equation. The basic regression equation for the currency demand, proposed by Tanzi [99], is the following:

$$\ln (C / M_2)_t = \beta_0 + \beta_1 \ln (1 + TW)_t + \beta_2 \ln (WS / Y)_t + \beta_3 \ln Rt + \beta_4 \ln (Y / N)_t + u_t \text{ with } \beta_1 > 0,\ \beta_2 > 0,\ \beta_3 < 0,\ \beta_4 > 0$$

where

ln denotes natural logarithms,

C / M2 is the ratio of cash holdings to current and deposit accounts,

TW is a weighted average tax rate (to proxy changes in the size of the shadow economy),

WS / Y is a proportion of wages and salaries in national income (to capture changing payment and money holding patterns),

R is the interest paid on savings deposits (to capture the opportunity cost of holding cash) and

Y / N is the per capita income.*

*In table 8 of the present paper, the econometric estimation of such a currency demand function for Austria is shown. More causes of the shadow economy (regulations, different tax rates, complexity of the tax system) are also included. The application of such a currency demand equation has been criticized by Thomas [18], but part of this criticism has been considered by Giles [19-20] and Bhattacharyya [38], who both use the latest econometric techniques.

The excess increase in currency, which is the amount unexplained by the conventional or normal factors (mentioned above) is then attributed to the rising tax burden and the other reasons leading people to work in the shadow economy. Figures for the size and development of the shadow economy can be calculated in a first step by comparing the difference between the development of currency when the direct and indirect tax burden (and government regulations) are held at its lowest value, and the development of currency with the current (much higher) burden of taxation and government regulations. Assuming in a second step the same income velocity for currency used in the shadow economy as for legal M1 in the official economy, the size of the shadow economy can be computed and compared to the official GDP.

The currency demand approach is one of the most commonly used approaches. It has been applied to many OECD countries [12, 13, 15, 40], but has nevertheless been criticized on various grounds [6, 8, 18, 86, 95]. The most commonly raised objections to this method are as follows:

(a) Not all transactions in the shadow economy are paid in cash. Isachsen and Strom* used the survey method to find out that in Norway, in 1980, roughly 80 per cent of all transactions in the hidden sector were paid in cash. The size of the total shadow economy *(*including barter*)* may thus be even larger than previously estimated;

(b) Most studies consider only one particular factor, the tax burden, as a cause of the shadow economy. But other factors, such as the impact of regulations, taxpayers' attitudes toward the State and tax morality, are not considered, because reliable data for most countries are not available. If, as seems likely, these other factors also have an impact on the extent of the hidden economy, it might again be higher than reported in most studies;**

(c) A further weakness of this approach, at least when applied to the United States, is discussed by Garcia [101], Park [86] and Feige [91], who point out that increases in currency demand deposits are due largely to a slowdown in demand deposits rather than to an increase in currency caused by activities in the shadow economy;

(d) Blades [102] and Feige [80, 103] criticize Tanzi's studies on the grounds that the United States dollar is used as an international currency. Tanzi should have considered allowed United States dollars, which are used as an international currency and held in cash abroad.*** Moreover, Frey and Pommerehne [4]

*See [74] and Anne I. Isachsen and Steinar Strom, "The hidden economy, the labour market and tax evasion", *Scandinavian Journal of Economics*, vol. 82 (1980), pp. 304-311.

**One weak justification for the use of only the tax variable is that it has by far the strongest impact on the size of the shadow economy in the studies known to the authors. The only exception is the study by Frey and Weck-Hannemann [39], where the "tax immorality" variable has a quantitatively larger and statistically stronger influence than the direct tax share in the model approach. In the study of the United States by Pommerehne and Schneider [71], which covers various tax measures and provides data on regulations, tax immorality and minimum wage rates, the tax variable has a dominating influence and contributes roughly 60-70 per cent to the size of the shadow economy. See also Zilberfarb [100].

***In another study by Tanzi ([3], pp. 110-113), this criticism is explicitly dealt with. A careful investigation of the amount of United States currency used abroad and in the shadow economy and of traditional criminal activities has been undertaken by Rogoff [104], who concludes that bills of large denomination are the major driving force behind the growth of the shadow economy and traditional criminal activities because of reduced transaction costs.

and Thomas [6, 18 and 95] claim that Tanzi's parameter estimates are not very stable;*

(e) Another weak point of this procedure, in most studies, is the assumption of the same velocity of money in both types of economy. As Hill and Kabir [59] for Canada and Kloveland [57] for the Scandinavian countries argue, there is already considerable uncertainty about the velocity of money in the official economy; the velocity of money in the hidden sector is even more difficult to estimate. Without knowledge about the velocity of currency in the shadow economy, assumption of an equal money velocity in both sectors has to be accepted;

(f) Finally, the assumption of no shadow economy in a base year is open to criticism. Relaxing this assumption would again imply an upward adjustment of the figures attained in the bulk of the studies already undertaken.

Physical input (electricity consumption) method

*Kaufmann-Kaliberda method***

To measure overall (official and unofficial) economic activity in an economy, Kaufmann and Kaliberda [107] assume that the consumption of electric power is the single best physical indicator of overall economic activity. Overall (official and unofficial) economic activity and electricity consumption have been empirically observed throughout the world to move in lockstep with an electricity/GDP elasticity usually close to one. By having a proxy measurement for the overall economy and subtracting it from estimates of official GDP, Kaufmann and Kaliberda derive an estimate of unofficial GDP. Kaufmann and Kaliberda thus suggest that the growth of total electricity consumption is an indicator for representing growth of official and unofficial GDP. According to this approach, the difference between the gross rate of registered (official) GDP and the gross rate of total electricity consumption can be attributed to the growth of the shadow economy. This method is very simple and appealing; however, it can also be criticized on the following grounds:

(a) Not all shadow economy activities *(*such as personal services*)* require a considerable amount of electricity. Other energy sources can be used *(*gas, oil, coal etc.*)*, so that only a part of the shadow economy will be captured;

(b) Over time, there has been considerable technical progress. Both the production and use of electricity are more efficient than in the past, and that will apply in both official and unofficial uses;

(c) There may be considerable differences or changes in the elasticity of electricity/GDP across countries and over time.***

*However, in studies of European countries, Kirchgaessner [55, 56] and Schneider [26] reach the conclusion that the estimation results for Germany, Denmark, Norway and Sweden are quite robust when using the currency demand method. For Canada, Hill and Kabir find that the rise of the shadow economy varies with respect to the tax variable used and conclude that "when the theoretically best tax rates are selected and a range of plausible velocity values is used, this method estimates underground economic growth between 1964 and 1995 at between 3 and 11 per cent of GDP" ([59], p. 1553).

**This method was used earlier by Lizzeri [105] and Del Boca and Forte [106], and then much later by Portes [32], Kaufmann and Kaliberda [107] and Johnson, Kaufmann and Shleifer [14]. For a critique, see Lackó [29-31, 108].

***Johnson, Kaufmann and Shleifer [14] attempt to adjust for changes in the elasticity of electricity/GDP.

Lackó method

Lackó [28, 29, 108] assumes that a certain part of the shadow economy is associated with the household consumption of electricity. It includes, inter alia, so-called household production, do-it-yourself activities and other non-registered production and services. Lackó assumes that in countries where the section of the shadow economy associated with household electricity consumption is high, the rest of the hidden economy, that is, the part that Lackó cannot measure, will also be high. Lackó ([29], pp. 19 ff.) assumes that in each country a part of the household consumption of electricity is used in the shadow economy.

Lackó's approach ([108], p.133) can be described by the following two equations:

$$\ln E_i = \alpha_1 \ln Ci + \alpha_2 \ln PRi + \alpha_3 G_i + \alpha_4 Q_i + \alpha_5 H_i + u_i \qquad (1)$$
$$\text{with } \alpha_1 > 0,\ \alpha_2 < 0,\ \alpha_3 > 0,\ \alpha_4 < 0,\ \alpha_5 > 0$$

$$H_i = \beta_1 T_i + \beta_2 (S_i - T_i) + \beta_3 D_i \qquad (2)$$
$$\text{with } \beta_1 > 0,\ \beta_2 < 0,\ \beta_3 > 0$$

where

i is the number assigned to the country,

E_i is per capita household electricity consumption in country i in millions of tons,

C_i is per capita real consumption of households without the consumption of electricity in country i in United States dollars (at purchasing power parity),

PR_i is the real price of consumption of 1 kilowatt-hour of residential electricity in United States dollars (at purchasing power parity),

G_i is the relative frequency of months when heating is needed in homes in country i,

Q_i is the ratio of energy sources other than electric energy to all energy sources in household energy consumption,

H_i is the per capita output of the hidden economy,

T_i is the ratio of the sum of paid personal income, corporate profit and taxes on goods and services to GDP,

S_i is the ratio of public social welfare expenditures to GDP,

and D_i is the sum of number of dependants over 14 years and of inactive earners, both per 100 active earners.

In a cross-country study, Lackó econometrically estimates equation (1), substituting equation (2) for H_i. The econometric estimation results can then be used to establish an ordering of the countries with respect to electricity use in their shadow economies. For the calculation of the actual size (value added) of the shadow economy, Lackó should know how much GDP is produced by one unit of electricity in the shadow economy of each country. Since the data are not known, Lackó takes the results obtained from shadow economy estimations carried out for a market economy using another approach during the early 1990s, and applies the results to the other countries. Lackó used the shadow economy of the United States as such a base (the shadow economy value of 10.5 per cent of GDP taken from Morris [109]), and then calculated the size of the shadow economy for other countries. Lackó's method is also open to the following criticism:

(a) Not all shadow economy activities require a considerable amount of electricity and other energy sources can be used;

(b) Shadow economy activities do not take place only in the household sector;

(c) It is doubtful whether the ratio of social welfare expenditures can be used as the explanatory factor for the shadow economy, especially in transition countries and developing countries;

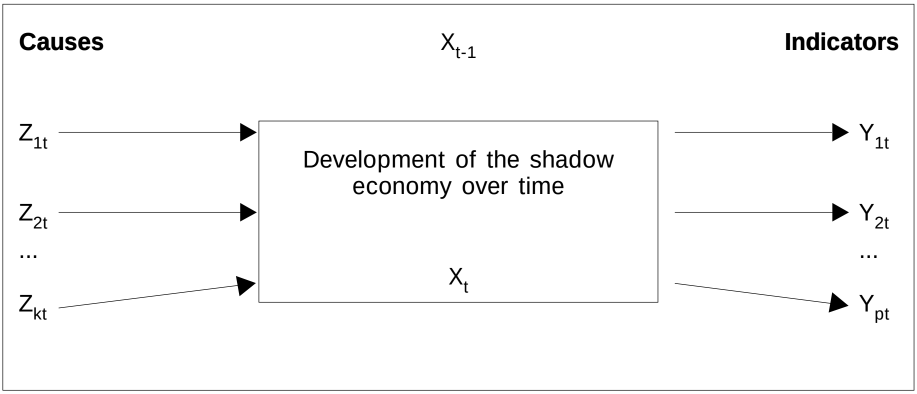
(d) It is not clear which base value of the shadow economy is the most reliable in calculating the size of the shadow economy for all other countries, especially the transition countries and developing countries.

Model approach*

All methods described so far that are designed to estimate the size and development of the shadow economy consider just one indicator that must capture all effects of the shadow economy. However, it is obvious that its effects show up simultaneously in the production, labour and money markets. An even more important critique is that the causes that determine the size of the hidden economy are taken into account only in some of the monetary approach studies that usually consider one cause, the burden of taxation. The model approach explicitly considers multiple causes leading to the existence and growth of the shadow economy over time, with the multiple effects that it entails. The empirical method used is quite different from those used so far. It is based on the statistical theory of unobserved variables, which considers multiple causes and multiple indicators of the phenomenon to be measured. For the estimation, a factor-analytic approach is used to measure the hidden economy as an unobserved variable over time. The unknown coefficients are estimated in a set of structural equations within which the unobserved variable cannot be measured directly. The DYMIMIC (dynamic multiple-indicators multiple-causes) model consists in general of two parts, while the measurement model links the unobserved variables to observed indicators. The structural equations model specifies causal relationships among the unobserved variables. In this case, there is one unobserved variable, the size of the shadow economy. It is assumed to be influenced by a set of indicators for the size of the shadow economy, thus capturing the structural dependence of the shadow economy on variables that may be useful in predicting its movement and size in the future. The interaction over time between the causes Z_{it} (i = 1, 2, ..., k), the size of the shadow economy X_t and the indicators Y_{jt} (j = 1, 2, ..., p) is shown in figure I.

*This part is a summarized version from a longer study by Aigner, Schneider and Ghosh ([110], p. 303), applying this approach for the United States over time. The pioneers of this approach are Weck [111] and Frey and Weck-Hannemann [39], who applied this approach to cross-section data from the 24 OECD countries for various years. Before turning to this approach, they developed the concept of "soft modelling" (Frey, Weck and Pommerehne [112] and Frey and Weck [113, 114]), an approach which has been used to provide a ranking of the relative size of the shadow economy in different countries.

Figure I. Development of the shadow economy over time



There is a large body of literature [6, 8, 10, 12, 15, 19, 20, 33] on the possible causes and indicators of the shadow economy. Causes of the following three types have been identified:

(a) The burden of direct and indirect taxation, both actual and perceived. A rising burden of taxation provides a strong incentive to work in the shadow economy;

(b) The burden of regulation as a proxy for all other State activities. It is assumed that increases in the burden of regulation give a strong incentive to enter the shadow economy;

(c) The tax morality (citizens' attitudes towards the State), which describes the readiness of individuals (at least partly) to leave their official occupations and enter the shadow economy. It is assumed that a declining tax morality tends to increase the size of the shadow economy.*

A change in the size of the shadow economy may be reflected in the following indicators:

(a) Development of monetary indicators. If activities in the shadow economy rise, additional monetary transactions are required;

(b) Development of the labour market. Increasing participation of workers in the hidden sector results in a decrease in participation in the official economy. Similarly, increased activities in the hidden sector may be expected to be reflected in shorter working hours in the official economy;

(c) Development of the production market. An increase in the shadow economy means that inputs *(*especially labour*)* move out of the official economy *(*at least partly*)*. Such displacement might have a depressing effect on the official growth rate of the economy.

*When applying this approach for European countries, Frey and Weck-Hannemann [39] had difficulty in obtaining reliable data for the cause series, besides the ones of direct and indirect tax burden. Hence, their study was criticized by Helberger and Knepel [115], who argue that the results were unstable with respect to changing variables in the model and over the years.

The latest use of the model approach has been undertaken by Giles [19, 20] and by Giles, Linsey and Gupsa [116]. They basically estimate a comprehensive (dynamic) MIMIC model to get a time-series index of the hidden/measured output of New Zealand or Canada, and then estimate a separate cash-demand model to obtain a benchmark for converting this index into percentage units. Unlike earlier empirical studies of the hidden economy, they paid proper attention to the non-stationary, and possible co-integration of time-series data in both models. This MIMIC model treats hidden output as a latent variable, and uses several (measurable) causal variables and indicator variables. The former include measures of the average and marginal tax rates, inflation, real income and the degree of regulation in the economy. The latter include changes in the (male) labour force participation rate and in the cash/money supply ratio. In their cash-demand equation they allow for different velocities of currency circulation in the hidden and recorded economies. Their cash-demand equation is not used as an input to determine the variation in the hidden economy over time. It is used only to obtain the long-run average value of hidden/measured output, so that the index for this ratio predicted by the MIMIC model can be used to calculate a level and the percentage units of the shadow economy. The latest combination by Giles of the currency demand method and the MIMIC approach clearly shows that some progress in the technique used in estimating the shadow economy has been achieved and a number of critical difficulties have been overcome.

Comparison of the results of the estimation of the shadow economy using different methods

As discussed above, there are nine different methods used to estimate the shadow economy. Table 9 shows the empirical results of the application of those methods to Canada, Germany, Italy, the United Kingdom and the United States.

The survey method used for all five countries provides lower estimates ranging from 1.5 per cent to 4.5 per cent for the period 1970-1980. The tax auditing method provides higher estimates of the shadow economy ranging from 2.9 per cent to 8.2 per cent for the period 1970-1980. Both methods also show that the shadow economy increases over time (for example, in the United States). The two discrepancy methods (expenditure versus income and official versus actual labour force) show no clear pattern. For some countries, they produce high shadow economy values (compared to the other methods used, as in the case of Germany); for some, the values are low (as in the case of Canada). Nor do they show a consistent time pattern. The physical input (electricity) method, for which only values for the period 1986-1990 are available for all five countries, shows values in the middle range for all countries (average value of 12.7 per cent over all countries and all periods). A comparison of the three monetary approaches (currency demand, cash-deposit ratio and transactions approach) reveals a clear pattern. The largest shadow economies for all five countries were achieved using the transactions approach (Feige method), ranging from 15 per cent to 35 per cent of GNP (average value of 21.9 per cent over all countries and periods). Somewhat lower results were achieved using the cash-deposit ratio approach (Gutmann method), ranging between 10 per cent and 30 per cent for all countries (average value of 15.5 per cent over all countries and all periods). Considerably lower values were achieved using the currency demand approach, ranging from 4 per cent to 20 per cent of GNP over the period 1970-1990 for all five countries (average value of 8.9 per cent over all countries and periods).

Table 9. Comparison of the average size, estimated by nine methods, of the shadow economy in five OECD countries over the period 1970-1990

(Percentage of GDP)

	Canada				*Germany*				*United Kingdom*				*Italy*				*United States*			
	Average over time period				*Average over time period*				*Average over time period*				*Average over time period*				*Average over time period*			
Method	*1970-1975*	*1976-1980*	*1981-1985*	*1986-1990*	*1970-1975*	*1976-1980*	*1981-1985*	*1986-1990*	*1970-1975*	*1976-1980*	*1981-1985*	*1986-1990*	*1970-1975*	*1976-1980*	*1981-1985*	*1986-1990*	*1970-1975*	*1976-1980*	*1981-1985*	*1986-1990*
Household surveys	-	-	1.3	1.4	3.6	-	-	-	1.5	-	-	-	-	-	-	-	3.7	4.5	5.6	-
Tax auditing	-	-	2.9	-	-	-	-	-	-	-	-	-	3.0	3.9	-	10.0	4.9	6.3	8.2	10.0
Discrepancy between expenditure and income	-	-	-	-	11.0	10.2	13.4	-	2.5	3.6	4.2	-	3.2	4.3	-	9.3	3.2	4.9	6.1	10.2
Discrepancy between official and actual employment	-	-	-	-	23.0	38.5	34.0	-	-	-	-	-	-	18.4	-	-	-	-	-	-
Physical input (electricity) method	-	-	-	11.2	-	-	-	14.5	-	-	-	13.2	-	-	-	19.3	-	-	7.8	9.9
Currency demand (Tanzi)	5.1	6.3	8.8	12.0	4.5	7.8	9.2	11.3	4.3	7.9	8.5	9.7	11.3	13.2	17.5	21.3	3.5	4.6	5.3	6.2
Cash deposit ratio (Gutmann)	13.8	15.9	11.2	18.4	-	-	-	-	14.0	7.2	6.2	-	23.4	27.2	29.3	-	8.8	11.2	14.6	-
Transactions approach (Feige)	-	26.5	15.4	21.2	17.2	22.3	29.3	31.4	17.2	12.6	15.9	-	19.5	26.4	34.3	-	17.3	24.9	21.2	19.4
MIMIC method (Frey and Weck-Hanneman)	-	8.7	-	-	5.8	6.1	8.2	-	-	8.0	-	-	-	10.5	-	-	-	8.2	-	-
Number of methods used	2	4	5	5	6	5	5	3	5	5	4	2	5	7	3	4	6	7	7	5

Note: The values were grouped (when possible, averaged) in the time periods 1970-1975, 1976-1980, 1981-1985 and 1986-1990, in order to make a rough comparison. The sources of the values are given by country.

Sources: Author's calculations using the following sources:
For Canada, Lippert and Walker [9], Thomas [6], Hill and Kabir [59], Schneider [12] and Bendelac and Clair [117].
For Germany, Lippert and Walker [9] and Schneider [10-12].
For the United Kingdom, Thomas [6], Lippert and Walker [9], Schneider [10-12] and Pozo [8].
For Italy, Thomas [6], Lippert and Walker [9], Pozo [8], Schneider [10-12] and Bendelac and Clair [117].
For the United States, Thomas [6], Lippert and Walker [9], Pozo [8], Schneider [10-12] and Bendelac and Clair [117].
Tanzi (1986), Feige (1986), Thomas (1986).

The currency demand approach shows a strongly rising shadow economy in all five countries, a result opposite to that given by the transactions and cash-deposit methods. The model approach shows values in the medium range, from 6.1 to 10.5 per cent for the period 1976-1980 (average value of 7.9 per cent for all countries over all periods). In general, these results demonstrate quite clearly that a huge range of estimates of the shadow economy for a country in a given time span is achievable using different calculation methods. Hence, there is a need for great caution when interpreting the size of the shadow economy of a country using only one method.

Summary and conclusions

There are many obstacles to be overcome to measure the size of the shadow economy and to analyse its impact on the official economy, although some progress has been made. The present paper has shown that while it is difficult to estimate the size of the shadow economy, it is not impossible. It has been demonstrated that with various methods, such as the currency demand method, the physical input method and the model approach, some insights can be provided into the size and development of the shadow economy of developing countries, transition countries and OECD countries. The results achieved through the use of these methods give the general impression that, in all the countries investigated, the shadow economy has become remarkably large.

In summary, there appears to be no best or commonly accepted method; each approach has its specific strengths and weaknesses as well as specific insights and results. Although the different methods provide a rather wide range of estimates, there is a common finding that the size of the shadow economy has been growing over the recent decade in most transition countries and in all the OECD countries studied. A similar finding has emerged for the shadow labour market, which is attracting growing attention because of high unemployment in European OECD countries. Furthermore, the results of the present survey show that an increasing burden of taxation and social security payments, combined with rising State regulatory activities, is the major driving force for the growth of the shadow economy. According to some studies, a growing shadow economy has a negative impact on official GDP growth, but other studies show a positive impact, hence, much more research is needed. Finally, shadow economies are a complex phenomenon, present to an important extent even in industrialized and developed economies. People engage in shadow economic activity for a variety of reasons, the most important of which include government actions, in particular, taxation and regulatory measures. With those two insights goes a third and no less important one: a Government that wants to decrease shadow economic activity has to first and foremost analyse the complex and frequently contradictory implications of its own policy decisions.

References

1. Friedrich Schneider and Dominik Enste, “Shadow economies: size, causes, and consequences”, *The Journal of Economic Literature* (Spring 2000).

2. Vito Tanzi, ed., *The Underground Economy in the United States and Abroad* (Lexington, Massachusetts, Lexington Books, 1982).

3. Vito Tanzi, “A second (and more sceptical) look at the underground economy in the United States” in *The Underground Economy in the United States and Abroad...*, pp. 38-56.

4. Bruno S. Frey and Werner W. Pommerehne, “The hidden economy: state and prospects for measurement”, *Review of Income and Wealth*, vol. 30, No. 1 (1984), pp. 1-23.

5. Edgar L. Feige, ed., *The Underground Economies: Tax Evasion and Information Distortion* (New York, Cambridge University Press, 1989).

6. Jim J. Thomas, *Informal Economic Activity*, London School of Economics, Handbooks in Economics (London, Harvester Wheatsheaf, 1992).

7. Norman V. Loayza, “The economics of the informal sector: a simple model and some empirical evidence from Latin America”, *Carnegie-Rochester Conference Series on Public Policy*, vol. 45 (1996), pp. 129-162.

8. Susan Pozo, ed., *Exploring the Underground Economy: Studies of Illegal and Unreported Activity* (Kalamazoo, Michigan, W. E. Upjohn Institute for Employment Research, 1996).

9. Owen Lippert and Michael Walker, eds., *The Underground Economy: Global Evidence of its Size and Impact* (Vancouver, British Columbia, The Fraser Institute, 1997).

10. Friedrich Schneider, “Measuring the size and development of the shadow economy. Can the causes be found and the obstacles be overcome?” in *Essays on Economic Psychology*, Hermann Brandstaetter and Werner Güth, eds. (Berlin, Springer Publishing Company, 1994), pp. 193-212.

11. Friedrich Schneider, “Can the shadow economy be reduced through major tax reforms? An empirical investigation for Austria”, *Supplement to Public Finance/Finances Publiques*, vol. 49 (1994), pp. 137-152.

12. Friedrich Schneider, “The shadow economies of Western Europe”, *Journal of the Institute of Economic Affairs*, vol. 17, No. 3 (1997), pp. 42-48.

13. Friedrich Schneider, “Further empirical results of the size of the shadow economy of 17 OECD-countries over time”, discussion paper to be presented at the fifty-fourth Congress of the International Institute of Public Finance, held in Cordoba, Argentina, in 1998 (Linz, Austria, University of Linz, Department of Economics, 1998).

14. Simon Johnson, Daniel Kaufmann and Andrei Shleifer, *The Unofficial Economy in Transition*, Brookings Papers on Economic Activity (Washington, D.C., 1997).

15. Simon Johnson, Daniel Kaufmann and Pablo Zoido-Lobatón, “Regulatory discretion and the unofficial economy”, *The American Economic Review*, vol. 88, No. 2 (1998), pp. 387-392.

16. “Controversy: on the hidden economy”, *Economic Journal*, vol. 109, No. 456 (June 1999).

17. Vito Tanzi, “Uses and abuses of estimates of the underground economy”, *The Economic Journal*, vol. 109, No. 456 (1999), pp. 338-340.

18. Jim J. Thomas, “Quantifying the black economy: ‘Measurement without theory’ yet again?” *The Economic Journal*, vol. 109, No. 456 (1999), pp. 381-389.

19. David Giles, “Measuring the hidden economy: implications for econometric modelling”, *The Economic Journal*, vol. 109, No. 456 (1999), pp. 370-380.

20. David Giles, “Modelling the hidden economy in the tax-gap in New Zealand”, working paper (Victoria, British Columbia, University of Victoria, Department of Economics, 1999).

21. Friedrich Schneider, “Stellt das Anwachsen der Schwarzarbeit eine wirtschaftspolitische Herausforderung dar? Einige Gedanken aus volkswirtschaftlicher Sicht”, *Mitteilungen des Instituts für angewandte Wirtschaftsforschung*, vol. I/98 (Linz, 1998), pp. 4-13.

22. Edgar L. Feige, “The underground economy and the currency enigma”, *Supplement to Public Finance/Finances Publiques*, vol. 49 (1994), pp. 119-136.

23. Herald Lubell, *The Informal Sector in the 1980’s and 1990’s* (Paris, Organisation for Economic Cooperation and Development, 1991).

24. J. D Smith, “Market motives in the informal economy”, in *The Economics of the Shadow Economy*, W. Gaertner and A. Wenig, eds. (Heidelberg, Springer, 1985), pp. 161-177.

25. Gunnar V. Mogensen, Hans K. Kvist, Eszter Körmendi and Soren Pedersen, *The Shadow Economy in Denmark 1994: Measurement and Results*, study No. 3 (Copenhagen, The Rockwool Foundation Research Unit, 1995).

26. Friedrich Schneider, “Estimating the size of the Danish shadow economy using the currency demand approach: an attempt”, *The Scandinavian Journal of Economics*, vol. 88, No. 4 (1986), pp. 643-668.

27. James Andreoni, Brian Erard and Jonathan Feinstein, “Tax compliance”, *Journal of Economic Literature*, vol. 36 (1998), pp. 818-860.

28. Mária Lackó, “Hidden economy an unknown quantity? Comparative analyses of hidden economies in transition countries in 1989-95”, working paper No. 9905 (Linz, Austria, University of Linz, Department of Economics, 1999).

29. Mária Lackó, “Hidden economy in East-European countries in international comparison”, working paper (Laxenburg, Austria, International Institute for Applied Systems Analysis, 1996).

30. Mária Lackó, “The hidden economies of Visegard countries in international comparison: a household electricity approach”, working paper (Budapest, Hungarian Academy of Sciences, Institute of Economics, 1997).

31. Mária Lackó, “Do power consumption data tell the story? (Electricity intensity and the hidden economy in post-socialist countries”, working paper (Laxenburg, International Institute for Applied Systems Analysis, 1997).

32. Alejandro Portes, “The informal economy”, in *Exploring the underground economy: Studies of Illegal and Unreported Activity*, Susan Pozo, ed. (Kalamazoo, Michigan, W. E. Upjohn Institute for Employment Research, 1996), pp. 147-165.

33. Simon Johnson, Daniel Kaufmann and Pablo Zoido-Lobatón, *Corruption, Public Finances and the Unofficial Economy*, discussion paper (Washington, D.C., World Bank, 1998).

34. Franz Van der Berg, “The shadow economy in South Africa: some new results”, unpublished manuscript (Cape Town, South Africa, University of South Africa, 1990).

35. G. M. Hartzenburg and A. Leimann, “The informal economy and its growth potential”, in *Economic Growth in South Africa*, E. Adebian and B. Standish, eds. (Oxford, Oxford University Press, 1992), pp. 187-214.

36. M.S.D. Bagachwa and A. Naho, “Estimating the second economy in Tanzania”, *World Development*, vol. 23, No. 8 (1995), pp. 1387-1399.

37. Tiho Yoo and Jin K. Hyun, “International comparison of the black economy: empirical evidence using micro-level data”, unpublished paper presented at the fifty-fourth Congress of the International Institute of Public Finance, held in Córdoba, Argentina, in 1998.

38. D. K. Bhattacharyya, “On the economic rationale of estimating the hidden economy”, *The Economic Journal*, vol. 109, No. 456 (1999), pp. 348-359.

39. Bruno S. Frey and Hannelore Weck-Hannemann, “The hidden economy as an ‘unobserved’ variable”, *European Economic Review*, vol. 26, No. 1 (1984), pp. 33-53.

40. Colin C. Williams and Jan Windebank, "Black market work in the European Community: Peripheral work for peripheral localities?", *International Journal of Urban and Regional Research*, vol. 19, No. 1 (1995), pp. 23-39.

41. Sanja Madzarevic and Davor Mikulic, "Measuring the unofficial economy by the system of national accounts", working paper (Zagreb, Institute of Public Finance, 1997).

42. Ignacio Mauleon "Quantitative estimation of the Spanish underground economy", discussion paper (Salamanca, Spain, University of Salamanca, Department of Economics and History, 1998).

43. Friedrich Schneider and Günther Pöll, "Schattenwirtschaft", in *Handbuch der Wirtschaftsethik*, Wilhelm Korff, ed. (Güthersloh, Germany, Gütersloher Publishing Company, 1999).

44. L. Frey, *Il lavoro a domicilio in Lombardia* (Milan, Giunta Regionale Lombarda, Assessorato al Lavoro, 1972).

45. L. Frey, "Il potenziale di lavoro in Italia", *Documenti ISVET*, No. 50 (1975).

46. L. Frey, "Il lavoro nero nel 1977 in Italia", *Tendenze della occupazione,* No. 6 (1978).

47. L. Frey, "Introduzione all' analisi economica del lavoro minorile", *Economia del Lavoro,* No. 1-2 (1980), pp. 5-16.

48. M. A Cappiello, "Proposita di bibliografia ragionata sull'economia sommersa nell'industria (Italia 1970-82)", in *L'altra metà dell'economia. La ricerca internazionale sull'economia informale,* A. Bagnasco, ed. (Naples, Liguori, 1986), pp. 307-345.

49. Bruce Bartlett, "Corruption, the underground economy, and taxation", unpublished manuscript (Washington, D.C., National Center for Policy Analysis, 1998).

50. Bruno Dallago, *The Irregular Economy: The "Underground Economy" and the "Black Labour Market"* (Aldershot, United Kingdom of Great Britain and Northern Ireland, Dartmouth Publishing Company, 1990).

51. Lawrence A. Chickering and Muhamed Salahdine, eds., *The Silent Revolution—The Informal Sector in Five Asian and Near-Eastern Countries,* an International Center for Economic Growth Publication (San Francisco, ICS Press, 1991).

52. Thomas Lemieux, Bernard Fortin and Pierre Fréchette, "The effect of taxes on labor supply in the underground economy", *The American Economic Review,* vol. 84, No. 1 (1994), pp. 231-254.

53. Friedrich Schneider, "The increase of the size of the shadow economy of 18 OECD-countries: some preliminary explanations", paper presented at the Annual Public Choice Meeting, held in Charleston, South Carolina, from 10 to 12 March 2000.

54. Peter S. Spiro, "Evidence of a post-GST increase in the underground economy", *Canadian Tax Journal*, vol. 41, No. 2 (1993), pp. 247-258.

55. Gebhard Kirchgaessner, "Size and development of the West German shadow economy, 1955-1980", *Zeitschrift für die gesamte Staatswissenschaft*, vol. 139, No. 2 (1983), pp. 197-214.

56. Gebhard Kirchgaessner, "Verfahren zur Erfassung des in der Schattenwirtschaft erarbeiteten Sozialprodukts", *Allgemeines Statistisches Archiv,* vol. 68, No. 4 (1984), pp. 378-405.

57. Jan Kloveland, "Tax evasion and the demand for currency in Norway and Sweden: Is there a hidden relationship?", *Scandinavian Journal of Economics,* vol. 86, No. 4 (1984), pp. 423-439.

58. Richard Cebula, "An empirical analysis of the impact of government tax and auditing policies on the size of the underground economy: the case of the United States 1993-1994", *American Journal of Economic Sociology,* vol. 56, No. 2 (1997), pp. 173-185.

59. Roderick Hill and Muhammed Kabir, “Tax rates, the tax mix, and the growth of the underground economy in Canada: What can we infer?” *Canadian Tax Journal/Revue Fiscale Canadienne*, vol. 44, No. 6 (1996), pp. 1552-1583.

60. E. Friedman, S. Johnson, D. Kaufmann and P. Zoido-Labton, “Dodging the grabbing hand: the determinants of unofficial activity in 69 countries”, discussion paper (Washington, D.C., World Bank, 1999).

61. J. W. Brehm, *A Theory of Psychological Reactance* (New York, Academic Press, 1966).

62. J. W. Brehm, *Responses to Loss of Freedom. A Theory of Psychological Reactance* (Morristown, General Learning Press, 1972).

63. Linde Pelzmann, *Wirtschaftspsychologie: Arbeitslosenforschung, Schattenwirtschaft, Steuerpsychologie* (Vienna, Springer, 1988).

64. Bruno S. Frey, “How large (or small) should the underground economy be?”, in *The underground economies: Tax Evasion and Information Distortion*, Edgar L. Feige, ed. (New York, Cambridge University Press, 1989), pp. 133-149.

65. John F. Houston, “Estimating the size and the implication of the underground economy”, Working Paper No. 87-89 (Philadelphia, Pennsylvania, Federal Reserve Bank, 1987).

66. Markus C. Adam and Viktor Ginsburgh, “The effects of irregular markets on macroeconomic policy: some estimates for Belgium”, *European Economic Review*, vol. 29, No. 1 (1985), pp. 15-33.

67. Ronald Fichtenbaum, “The productivity slowdown and the underground economy”, *Quarterly Journal of Business and Economies*, vol. 28, No. 3 (1989), pp. 78-90.

68. Friedrich Schneider, Markus F. Hofreither and Reinhard Neck, “The consequences of a changing shadow economy for the official economy: Some empirical results for Austria”, in *The Political Economy of Progressive Taxation*, Dieter Boes and Bernhard Felderer, eds. (Heidelberg, Springer, 1989), pp. 181-211.

69. Reinhard Neck, Markus Hofreither and Friedrich Schneider, “The consequences of progressive income taxation for the shadow economy: some theoretical considerations”, in *The Political Economy of Progressive Taxation*, Dieter Boes and Bernhard Felderer, eds. (Heidelberg, Springer Publishing Company, 1989), pp. 149-176.

70. Peter J. Quirk, “Macroeconomic implications of money laundering”, IMF working paper WP/96/66 (Washington, D.C., International Monetary Fund, 1996).

71. Werner W. Pommerehne and Friedrich Schneider, “The decline of productivity growth and the rise of the shadow economy in the United States”, working paper (Aarhus, Denmark, University of Aarhus, 1985).

72. Patrick K. Asea, “The informal sector: baby or bath water?”, *Carnegie-Rochester Conference Series on Public Policy*, 45 (1996), pp. 163-171.

73. Arne J. Isachsen, Jan Klovland and Strom Steinar, “The hidden economy in Norway”, in *The Underground Economy in the United States and Abroad*, Vito Tanzi, ed. (Lexington, Massachusetts, Heath, 1982), pp. 209-231.

74. Arne J. Isachsen and Strom Steinar, “The size and growth of the hidden economy in Norway”, *Review of Income and Wealth*, vol. 31, No. 1 (1985), pp. 21-38.

75. *Estimates of Income Unreported on Individual Tax Forms* (Washington, D.C., United States Department of the Treasury, Internal Revenue Service, 1979).

76. *Income Tax Compliance Research: Estimates for 1973-1981* (Washington, D.C., United States Department of the Treasury, Internal Revenue Service, 1983).

77. C. B. Simon and A. G. Witte, *Beating the System: The Underground Economy* (Boston, Massachusetts, Urban House, 1982).

78. A. D. Witte, “The nature and extent of unreported activity: a survey concentrating on recent United States research”, in *The Unofficial Economy: Consequences and Perspectives in Different Economic Systems*, S. Alessandrini and B. Dallago, eds. (Aldershot, United Kingdom of Great Britain and Northern Ireland, Gower, 1987).

79. Charles T. Clotefelter “Tax evasion and tax rates: an analysis of individual return”, *Review of Economic Statistics*, vol. 65, No. 3 (1983), pp. 363-373.

80. Edgar L. Feige, “A re-examination of the ‘underground economy’ in the United States”, *IMF Staff Papers*, vol. 33, No. 4 (Washington, D.C., 1986), pp. 768-781.

81. A. Franz, “Wie groß ist die ‘schwarze’ Wirtschaft?”, *Mitteilungsblatt der Österreichischen Statistischen Gesellschaft*, vol. 49, No. 1 (1983), pp. 1-6.

82. Kerrick MacAfee, “A glimpse of the hidden economy in the national accounts”, *Economic Trends*, vol. 136 (1980), pp. 81-87.

83. Michael O’Higgins, “Assessing the underground economy in the United Kingdom”, in *The Underground Economies: Tax Evasion and Information Distortion*, Edgard L. Feige, ed. (New York, Cambridge University Press, 1989), pp. 175-195.

84. Hans-Georg Petersen, “Size of the public sector, economic growth and the informal economy: development trends in the Federal Republic of Germany”, *Review of Income and Wealth*, vol. 28, No. 2 (1982), pp. 191-215.

85. Daniela Del Boca, “Parallel economy and allocation of time”, *Micros (Quarterly Journal of Microeconomics*), vol. 4, No. 2 (1981), pp. 13-18.

86. T. Park, “Reconciliation between personal income and taxable income”, mimeograph (Washington, D.C., Bureau of Economic Analysis, 1979), pp. 1947-1977.

87. C. Pissarides and G. Weber, “An expenditure-based estimate of Britain’s black economy”, Creative Learning Exchange working paper No. 104 (London, Creative Learning Exchange, 1988).

88. Bruno Contini, “Labour market segmentation and the development of the parallel economy—the Italian experience”, *Oxford Economic Papers*, vol. 33, No. 4 (1981), pp. 401-412.

89. David M. O’Neill, “Growth of the underground economy 1950-81: some evidence from the current population survey”, study prepared for the Joint Economic Committee of the United States Congress (Washington, D.C., Government Printing Office, 1983).

90. Edgar L. Feige, “How big is the irregular economy?”, *Challenge*, vol. 22, No. 1 (1979), pp. 5-13.

91. Edgar L. Feige, “Overseas holdings of United States currency and the underground economy”, in *Exploring the Underground Economy: Studies of Illegal and Unreported Activity*, Susan Pozo, ed. (Kalamazoo, Michigan, W. E. Upjohn Institute for Employment Research, 1996), pp. 5-62.

92. Werner C. Boeschoten and Marcel M. G. Fase, *The Volume of Payments and the Informal Economy in the Netherlands 1965-1982* (Dordrecht, M. Nijhoff, 1984).

93. Enno Langfeldt, “The unobserved economy in the Federal Republic of Germany”, in *The Unobserved Economy*, Edgar L. Feige, ed. (New York, Cambridge University Press., 1984), pp. 236-260.

94. Vito Tanzi, “The underground economy in the United States: reply to comments by Feige, Thomas, and Zilberfarb”, *IMF Staff Papers*, vol. 33, No. 4 (1986), pp. 799-811.

95. Jim J. Thomas, “The underground economy in the United States: a comment on Tanzi”, *IMF Staff Papers*, vol. 33, No. 4 (1986), pp. 782-789.

96. Phillip Cagan, “The demand for currency relative to the total money supply”, *Journal of Political Economy*, vol. 66, No. 3 (1958), pp. 302-328.

97. Pierre M. Gutman, “The subterranean economy”, *Financial Analysts Journal*, vol. 34, No. 1 (1977), pp. 24-27.

98. Vito Tanzi, “The underground economy in the United States: estimates and implications”, Banca Nazionale de Lavoro, vol. 135, No. 4 (1980), pp. 427-453.

99. Vito Tanzi, “The underground economy in the United States: annual estimates, 1930-1980”, *IMF Staff Papers,* vol. 30, No. 2 (Washington, D.C., International Monetary Fund, 1983), pp. 283-305.

100. Ben-Zion Zilberfarb, “Estimates of the underground economy in the United States, 1930-80”, *IMF Staff Papers*, vol. 33, No. 4 (1986), pp. 790-798.

101. Gillian Garcia, “The currency ratio and the subterranean economy”, *Financial Analysts Journal*, vol. 69, No. 1 (1978), pp. 64-66.

102. Derek Blades, “The hidden economy and the national accounts”, *OECD Occasional Studies* (Paris, Organization for Economic Cooperation and Development, 1982), pp. 28-44.

103. Edgar L. Feige, “Revised estimates of the underground economy: implications of U.S. currency held abroad”, in *The Underground Economy: Global Evidence of its Size and Impact*, Owen Lippert and Michael Walker, eds. (Vancouver, British Columbia, The Frazer Institute, 1997), pp. 151-208.

104. Kenneth Rogoff, “Blessing or curse? Foreign and underground demand for euro notes”, *Economic Policy*: *The European Forum,* No. 26 (1998), pp. 261-304.

105. C. Lizzeri, *Mezzogiorno in controluce* (Naples, Enel, 1979).

106. Daniela Del Boca and Francesco Forte, “Recent empirical surveys and theoretical interpretations of the parallel economy in Italy”, in *The Underground Economy in the United States and Abroad*, Vito Tanzi, ed. (Lexington, Massachusetts, Lexington Books, 1982), pp. 160-178.

107. Daniel Kaufmann and Aleksander Kaliberda, “Integrating the unofficial economy into the dynamics of post-socialist economies: a framework of analyses and evidence”, policy research working paper No. 1691 (Washington, D.C., World Bank, 1996).

108. Mária Lackó, “The hidden economies of Visegard countries in international comparison: a household electricity approach”, in *Hungary: towards a market economy,* L. Halpern and C. Wyplosz, eds. (Cambridge, Massachusetts, Harvard University Press, 1998), pp. 128-152.

109. B. Morris, *International Statistical Yearbook* (Budapest, 1993).

110. Dennis Aigner, Friedrich Schneider and Damayanti Ghosh, “Me and my shadow: estimating the size of the United States hidden economy from time series data”, in *Dynamic Econometric Modelling,* W. A. Barnett, E. R. Berndt and H. White, eds. (Cambridge, Massachusetts, Harvard University Press, 1988), pp. 224-243.

111. Hannelore Weck, *Schattenwirtschaft: Eine Möglichkeit zur Einschränkung der öffentlichen Verwaltung? Eine ökonomische Analyse,* Frankfurt, Verlag Lang, 1983.

112. Bruno S. Frey, Hannelore Weck and Werner W. Pommerehne, “Has the shadow economy grown in Germany? An exploratory study”, *Weltwirtschaftliches Archiv,* vol. 118, No. 4 (1982), pp. 499-524.

113. Bruno S. Frey and Hannelore Weck, “Bureaucracy and the shadow economy: a macro-approach”, in *Anatomy of Government Deficiences,* Horst Hanusch, ed. (Berlin), Springer (1983), pp. 89-109.

114. Bruno S. Frey and Hannelore Weck, “Estimating the shadow economy: a ‘naive’ approach”, *Oxford Economic Papers,* vol. 35 (1983), pp. 23-44.

115. Claus Helberger and Hans Knepel, “How big is the shadow economy? A re-analysis of the unobserved-variable approach of B. S. Frey and H. Weck-Hannemann”, *European Economic Journal*, vol. 32 (1988), pp. 965-976.

116. David Giles, Linsey Tedds and Gugsa Werkneh, “The Canadian underground and measured economies”, working paper (Victoria, British Columbia, University of Victoria, Department of Economics, 1999).

117. Jacques Bendelac and Pierre-Maurice Clair, “Macroeconomic measures of the hidden economy”, paper presented at the forty-ninth Congress of the International Institute of Public Finance, held in Berlin in 1993.