Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

International surveys: Motives and methodologies

by Mary E. Thompson

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

December 2008

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2008
Vol. 34, No. 2, pp. 131-141
Statistics Canada, Catalogue No. 12-001-X

International surveys: Motives and methodologies

Mary E. Thompson [1]

Abstract

The context of the discussion is the increasing incidence of international surveys, of which one is the International Tobacco Control (ITC) Policy Evaluation Project, which began in 2002. The ITC country surveys are longitudinal, and their aim is to evaluate the effects of policy measures being introduced in various countries under the WHO Framework Convention on Tobacco Control. The challenges of organization, data collection and analysis in international surveys are reviewed and illustrated. Analysis is an increasingly important part of the motivation for large scale cross-cultural surveys. The fundamental challenge for analysis is to discern the real response (or lack of response) to policy change, separating it from the effects of data collection mode, differential non-response, external events, time-in-sample, culture, and language. Two problems relevant to statistical analysis are discussed. The first problem is the question of when and how to analyze pooled data from several countries, in order to strengthen conclusions which might be generally valid. While in some cases this seems to be straightforward, there are differing opinions on the extent to which pooling is possible and reasonable. It is suggested that for formal comparisons, random effects models are of conceptual use. The second problem is to find models of measurement across cultures and data collection modes which will enable calibration of continuous, binary and ordinal responses, and produce comparisons from which extraneous effects have been removed. It is noted that hierarchical models provide a natural way of relaxing requirements of model invariance across groups.

Key Words: International surveys; Longitudinal surveys; Analysis of survey data; Random effects; Data collection mode effects; Hierarchical models; Measurement models.

1. Introduction

I have chosen the topic of international surveys since one such survey, the International Tobacco Control survey, has been a major part of my activity in the last few years, and there are some interesting intersections with the areas to which Joseph Waksberg gave his attention – particularly frames for telephone surveys, and the effects of stratification with widely varying sampling rates.

The paper will begin with some discussion of the motifs and motives of international surveys and some examples. It will touch on the challenges of organization, data collection and analysis. Finally, it will consider two problems to be addressed in analysis: (i) survey sampling theory and the pooling of data from several countries, and (ii) measurement across data collection modes and cultures.

The first large international survey was the World Fertility Survey, carried out in the 1970's through the International Statistical Institute, and funded by the U.S. Agency for International Development and other sponsors. It was a very ambitious one-time survey. The WFS eventually surveyed over 330,000 women in 61 countries, at a cost of about $50 million. It gave countries important comparison data on family sizes, and led to policy measures on population planning in several participating countries. It also produced analytic projects in the hundreds, including path-breaking methodological studies, and laid the foundation for international survey methodology, particularly in developing countries (Verma, Scott and O'Muircheartaigh 1980; Cleland and Verma 1989).

Another well known example is the Programme for International Student Assessment, a project of the Organization for Economic Co-operation and Development, beginning in 2000. PISA is a continuing survey, carried out every 3 years, with 15 year old youths in developed countries. It is growing in scope, with 67 countries expected to participate in 2009. The results allow countries to monitor the success of their education programs in providing verbal and quantitative literacy.

The Global Youth Tobacco Survey is a one-time survey which began in 2002, sponsored by the World Health Organization and the Centers for Disease Control and Prevention. The GYTS has focused on surveying youth aged 13 to 15 years in developing countries, and had carried out data collection in 129 countries by 2004. The objective is to measure tobacco use uptake among youth, and awareness of the associated health risks.

The European Social Survey (ESS 2008) is an "academically-driven social survey" in over 30 nations, funded by European and national agencies, and designed to "chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations".

Even as the use of local and national surveys is growing everywhere, so too is the incidence of international surveys, carried out by international agencies, non-governmental

[1] Mary E. Thompson, Department of Statistics and Actuarial Science, University of Waterloo. E-mail: methomps@uwaterloo.ca.

organizations and private sector firms. This burgeoning appears to be part of a trend toward global governance and concern for population health and well-being.

I have seen the purposes of international surveys classified as epidemiology, surveillance, monitoring and evaluation of the effects of policy. Evidently these classifications overlap. It can be argued that PISA, the GYTS and the ESS constitute surveillance and monitoring, because their data are related only indirectly to interventions. The WFS had a direct evaluation aspect, in countries that had introduced family planning programs. The International Tobacco Control (ITC) survey, to be discussed later in this section, is one of the few for which evaluation is the primary purpose.

Apart from scientific concerns, another important role for an international survey is to engage the governments of the countries; it provides a way for them to participate in global policy development even in the face of political and economic obstacles.

For the researcher, international surveys allow the comparison of the populations of countries, the possibility of interpretation of the differences, and sometimes even the possibility of shedding light on causes and effects – typically with the underlying aim of improving conditions and informing policy.

The International Tobacco Control Policy Evaluation Project (ITC Project) was initiated by Dr. Geoffrey T. Fong of Psychology at the University of Waterloo, with collaborators around the world (Fong, Cummings, Borland, Hastings, Hyland, Giovino, Hammond and Thompson 2006; Thompson, Fong, Hammond, Boudreau, Dreizen, Hyland, Borland, Cummings, Hastings, Siahpush, Mackintosh and Laux 2006). The impetus was the WHO Framework Convention on Tobacco Control (FCTC), which was passed in May 2003, and has been ratified by over 150 countries. By ratifying the treaty the participating countries pledge to introduce policy measures for tobacco control, such as strong health warning package labels, banning of cigarette advertising, and banning of smoking in public places. The necessity for national legislation has as a consequence that these measures are being introduced at various times and in various ways. For example, Canada in December 2000 introduced graphic warning labels, setting international precedents for the size of label (more than 50 % of the package) and vivid colour images. Since then a few other countries have adopted this same practice, while others have legislated prominent text warnings. For the current status of health warning regulations around the world, see ITC (2008). The MPower Report (WHO 2008) describes the global tobacco policy environment and six policies of focus for the FCTC.

The purpose of the ITC Project is to try to find out which measures are effective in reducing uptake of tobacco use, and in helping people already using tobacco to quit. Furthermore, it has the ambitious aim of trying to explain how those measures which are effective actually work. The investigating team includes social psychologists and specialists in social marketing, as well as epidemiologists and economists.

By September 2008 the ITC Project was carrying out surveys in 17 countries, with more likely to be added. The surveys began in 2002, in Canada, the US, the UK and Australia. That year, in each of the four countries, approximately 2000 adult smokerswere recruited by telephone using a geographically stratified random digit dial (RDD) frame, of which the science has origins in the famous Mitofsky-Waksberg method (Waksberg 1978). The recruited smokers were interviewed a week or two later, and have been followed up each year since then, regardless of whether they continued to smoke. Wave 6 for the ITC Four Country Survey was completed in February 2008.

Because sufficient sample size is needed to evaluate the effects of measures introduced between the waves, dropouts at each wave have been replaced with a cohort of new recruits. In the ITC Four Country Survey, new recruits in each wave have been selected using the same design as in Wave 1, without any attempt to match the characteristics of the dropouts. Weights construction at each wave is effectively carried out separately for each cohort, adjusting for differential attrition by region and by age-sex group. This design has helped us to discern "time-in-sample" effects, and time-in-sample is entertained as an explanatory variable in analytic models (Thompson, Boudreau and Driezen 2005).

The first national policy measures following 2002 were an advertising ban and enhanced warning labels in the UK, between Waves 1 and 2; graphic warning labels were introduced in Australia between Waves 4 and 5. In the ITC Four Country Survey we have what is sometimes called a natural experiment or quasi-experiment (Cook and Campbell 1979), where the non-policy countries serve as external controls; moreover, the longitudinal feature of the design provides internal control. The design has been replicated a number of times, with other sets of countries.

For example, it became clear early on that Ireland would be the first country to adopt national smoke-free legislation, coming into effect in March 2004. The ITC collaborators were able to put together parallel surveys in Ireland and the UK before the law came into effect, and to visit the same people a year later. The samples were again recruited nationally using a random digit dial (RDD) frame. There were 755 smokers in Ireland and 411 smokers in the UK who were interviewed at both waves. One interesting finding concerned support for a ban on smoking in pubs (Fong, Hyland, Borland, Hammond, Hastings, McNeill, Anderson, Cummings, Allwright, Mulcahy, Howell, Clancy, Thompson, Connolly and Driezen 2006). Figure 1 shows the proportions supporting or strongly supporting the ban in bars and pubs, in the two countries, by wave.

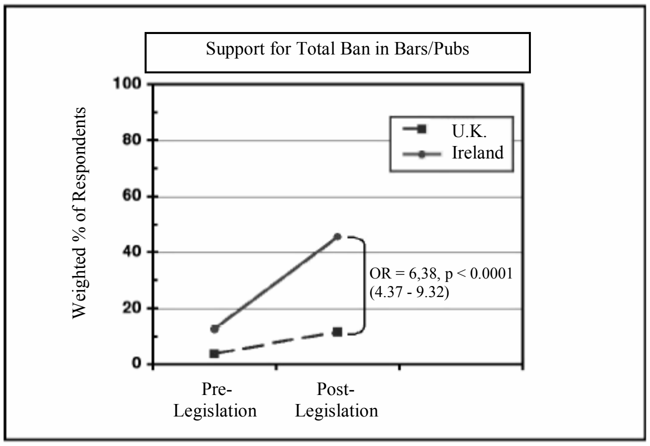

Figure 1 Support for smoke-free legislation in two waves of ITC Ireland/UK Survey

In the ITC sample of smokers the support increased between the two waves a little in the UK, and a great deal in Ireland. Moreover, the same survey showed no evidence that the reduction of smoking in public venues was associated with increased smoking in private venues. Showing broad acceptance of the smoke-free law by smokers, the ITC findings and others like them have helped bring about similar laws in Scotland, France, Germany, the rest of the UK, and the Netherlands. An ITC survey was carried out before and after the April 2006 implementation of the ban in Scotland, using the rest of the UK as the control, and the findings were replicated, except that by this time support in the rest of the UK had grown substantially (Hyland, Hassan, Higbee, Fong, Borland, Cummings, Thompson, Boudreau and Hastings 2008).

The model used for testing was simple: a GEE model, where Y is a binary measure of support for the ban, w is country, t is time, the wt term represents an interaction, and x is a vector of fixed individual level covariates:

$$\text{logit}[P(Y_t = 1 \mid w, x) = \alpha_0 + \alpha_1 w + \gamma t + \delta wt + x\beta,$$

$$\text{Corr}(Y_1,\ Y_2) = \rho.$$

The coefficient δ represents the difference in increase in support in the two countries, and we tested the hypothesis H_0: $\delta = 0$. There are other possible parametrizations, but this one has the advantage of matching the plot in Figure 1, which displays marginal proportions; the methodology is widely accepted, and supported by complex survey software.

2. Challenges

There are numerous challenges in carrying out an international survey. The WFS papers by Verma *et al.* (1980) and Cleland and Verma (1989) can be recommended for thoughtful discussions which are very little out of date. In this section we illustrate by describing some of the issues encountered by the ITC survey in organization and data collection.

Unlike the WFS, the ITC survey has been funded in the first instance by national granting programs, primarily the National Institutes of Health in the United States, and the Canadian Institutes for Health Research. The central infrastructure, led by Dr. Fong at the University of Waterloo and by Dr. K. Michael Cummings at Roswell Park Cancer Institute in Buffalo, works directly with investigating teams and agencies in the various countries. We have had to learn how to work with widely varying societies, political systems and cultures. Survey costs and budgets alone differ surprisingly from country to country. When governments contribute funding, they have their own requirements, and data ownership agreements must be negotiated. Since the amount of infrastructure and expertise can be quite different from place to place, the close coordination of the ITC Four Country Survey is difficult to replicate more widely.

For example, in the first half of 2008 the fieldwork was carried out for Wave 3 of a parallel survey (the ITC South-East Asia Survey) in Thailand and Malaysia, which are geographically close, and similar in some ways, but different in many dimensions. Thailand is ethnically quite homogeneous, while Malaysia has three major ethnic groups and many minor ones. More than half the population of Thailand lives in rural areas, but most of the Malaysian population is urban, and residential mobility is high. Thailand has extensive experience with surveys, including cohort studies, but when Wave 1 began in 2005, Malaysia was attempting this kind of cohort study for the first time. We tried to prescribe parallel sampling designs in the two countries, but had to make compromises. For example, it was found at the time of Wave 1 that the official sampling frames had different sized building blocks at the lowest level, consisting of clusters of households. This difference made the sample of households rather more dispersed in Malaysia, which had the smaller blocks. The greater dispersion meant greater work and costs. (Design effects are still larger for Malaysia than for Thailand, because of more heterogeneity at the level of the first stage units.)

An important aspect of the project is to try to build capacity for longitudinal health surveys in countries which are relatively new to this kind of work. We provide detailed protocols, training manuals, and data entry templates. We have learned to be more insistent on the identification of local expertise, particularly statistical expertise. Day-to-day communication is carried out by email and teleconferences. Final data cleaning and construction of survey weights normally occur at the University of Waterloo, but some country teams have been eager to participate in these parts of the operation.

We use telephone surveys, with recruitment by modified RDD, in the four original countries, as well as Ireland, South Korea, France and Germany; recruitment from the National Health Survey sample in New Zealand; and face to face surveys in Thailand, Malaysia Wave 1, China, Bangladesh, Mexico, and Uruguay.

In Malaysia Wave 2 we intended face-to-face data collection, but since both recontact and new recruitment proved to be very difficult because of a combination of factors, we moved to telephone interviewing where feasible in some areas. There was limited scope for comparison of modes, but in the large and mainly urban state of Selangor, 137 Wave 1 smokers (non-quitters) were reinterviewed face-to-face, and 63 were reinterviewed by telephone, making some tentative inferences possible. In Wave 3, an attempt was made to carry out both telephone and face-to-face interviewing in some of the same census districts, to enable a better assessment of data collection mode effects, and this study is in progress. At the same time, the proportion of the ITC Malaysia smoker sample interviewed face-to-face has decreased steadily, from 100% in Wave 1, to 63.5% in Wave 2, 44.4% in Wave 3. In Wave 4, we expect that telephone will be used exclusively in the mainland states.

The Wave 1 survey in the Netherlands has used parallel internet panel and RDD telephone interviewing, with sample sizes of about 400 and 1,800 respectively. This exercise will provide our best chance yet at being able to account for mode effects in modeling. These effects have been the subject of much research recently. For example, some studies have found that telephone respondents choose the extreme options of a Likert scale more of the time than web respondents do (Wichers and Zenderink 2006; Bronner and Kuijlen 2007).

The internet sample in the Netherlands consists of smokers randomly sampled from a large pre-recruited multi-purpose panel of about 200,000 people assembled by the firm TNS NIPO. The telephone sample, representing smokers accessible by land-line telephone, might well represent a different population of smokers. The low telephone response rates make clear that the public in the Netherlands is not as receptive to telephone surveys as the public in most of the other ITC countries. We requested that each group be asked about their accessibility by the other mode, so as to be able to use dual frame methods (Lohr and Rao 2000) to compute appropriate survey weights. We will also model propensity (Rosenbaum and Rubin 1984) for responding by telephone (say), given demographic variables and the accessibility variables, and control for propensity score in comparison of response patterns by mode.

Response rates vary a great deal, even within the ITC Four Country Survey, the response rates and retention rates being highest in Australia, and lowest in the United States. Certainly this jeopardizes the ability to compare across countries, in the sense that we can only compare the populations represented by the respondents – those in each country who would respond if approached under our protocol. The situation looks slightly better if we break response rates down into components. For example, we have seen from call attempt outcomes, and from our knowledge of the increased use of call filtering devices, that US adults are much harder to contact and recontact than adult residents of the other three countries. However, once contact is established, the US agreement or non-refusal rate is very similar, upwards of 80 %, to those of the other three countries.

We have measurement issues even for matters of fact, such as habits of tobacco purchase and use. In some countries like India, Bangladesh and Sudan, all under discussion for inclusion, there are many forms of tobacco in common use. For the developed countries, just keeping the list of cigarette brands current is a full time job. Compounding the difficulty is that whenever we ask about purchasing patterns or noticing advertisements we are asking people to remember what they have done over the previous two weeks, or some longer period. For the most part, we rely on self-report, and for a number of reasons self-report may not be accurate.

For attitudes and beliefs we have known all along that the questions must be suited to the language and the literacy level of the participants, but we were still surprised and sobered to find a high incidence of item non-response in outlying areas of one country, suggesting great difficulty with attitude and belief questions. In our pilot survey in India, the survey took an average of 1.5 hours per participant, despite having been shortened and simplified.

Psycho-social measurements need to be validated in each culture and language. For example, we have started to include a very short depression scale. Here is the version for the ITC Four Country Survey.

- During the last month, have you often been bothered by little interest or pleasure in doing things?
- During the last month, have you often been bothered by feeling down, depressed or hopeless?
- In the last year, have you been told by a doctor or other health care provider that you have depression?

And here is the version that we finally came to for ITC China Wave 2, on the advice of other researchers who reported having been able to validate a version of it.

> Below is a list of ways that you might have felt or behaved. Please tell me how often you have felt this way during the past week.
>
> 1. I did not feel like eating; my appetite was poor.
> 2. I felt hopeful about the future.
> 3. I felt sad.
> 4. I felt that people dislike me.

Ryder, Yang, Zhu, Yao, Yi, Heine and Bagby (2008) have released results of a very interesting comparative study of the expression of depression.

The catalogue of measurement issues goes on. Even if the question is supposed to be the same in two languages, it may be hard to find equivalences. We try to ensure a good quality translation using committee translation or comparison of independent translations, but must often fall short of perfection. For example, literal translations of English into French or German are a fair amount longer, and it takes considerable skill to make a translation that runs smoothly over the telephone. Thrasher, Quah, Borland, Awang, Sirirassamee, Boado, Miller, Watts and Dorantes (2008) describe a study in cognitive testing of some of the most important questions, across several countries.

There are more subtle cultural differences, particularly the degree to which respondents will give a socially desirable response. We have noticed what may be a higher tendency toward this among Mexicans and among anglophone Canadians. Johnson and Van de Vijver (2003) among others have discussed the possibility that cross-national differences in socially desirable responses may be related to "cultural value systems such as in the individualism and collectivism dimension" of Hofstede (1980).

In a longitudinal survey, we need to be concerned as well about the validity and reliability of repeated measures. As we have already indicated, it is common to observe what are called "time-in-sample effects", where the response proportions tend to drift upward or downward as the cohort proceeds, just because of the fact of being measured.

All these issues feed into the analytic challenges faced by researchers. Fundamentally, the aim of analysis must be to discern the real response (or lack of response) to policy change, separating it from the effects of data collection mode, differential non-response, external events, time-in-sample, culture, and language. This is a daunting task.

3. Pooling of data across countries

In the traditional survey analysis paradigm (Binder 1983; Godambe and Thompson 1986; Skinner 1989), there is a model for the responses y with parameter θ, and we imagine how we would estimate θ if we had responses from the whole population, in a census. We would use an efficient unbiased estimating equation like this:

$$\sum_{i=1}^{N} \phi_i(y_i,\ \theta) = 0,$$

to define a *census estimate*. To obtain the sample estimate, we use a weighted sum of the sample estimating function terms:

$$\sum_{i \in s} w_i\, \phi_i(y_i,\ \theta) = 0$$

to give an approximately unbiased estimator of the census estimating function. The survey weights are constructed to take into account the sampling design, and under-representation of some groups due to non-response and non-coverage. The usual interpretation of w_i is the number of population members represented by $i.$ The use of this sample estimating function is appealing because of the likely reduction of bias due to informative sampling and non-response; but if the weights are highly variable and the model for the terms is correct, the second equation gives an inefficient way of estimating θ.

Now when we are combining data from two countries with very different sampling fractions, as in the Ireland/UK survey, the weights for one country (the UK) will be much greater than the weights for the other country (Ireland). A literal application of the paradigm would have the data from the UK dominating the analysis. If the model is correct, the most efficient census estimate is the mean of y over the two countries combined. But then the corresponding sample estimate is an inefficient use of the sample. This problem is similar to that arising in case-control studies, as discussed by Scott (2006).

One way of producing better estimates while remaining in the traditional paradigm is to consider that the parameter value for the UK is $\theta - \Delta,$ that the parameter value for Ireland is $\theta + \Delta,$ and that we are trying to estimate $\theta,$ the arithmetic mean of the two. An efficient census estimating function system for θ and Δ is equivalent to one which separates into a part for each country. Since rescaling of weights within a country has then no effect on the point estimators and their properties, the survey-weighted sample version of that system yields efficient estimation.

Moreover, the ensuing analysis is the approximately the same as we would obtain from the original paradigm if we had equal sample sizes in the two countries, and rescaled the weights to sum to sample size within each country. As noted by Scott (2006), rescaling the weights in this manner is a very common practice among epidemiologists. It is in a sense a partial application of the q-weight method of

Pfeffermann and Sverchkov (1999), where the inverse inclusion probability weight is divided by a kind of expectation of the weight, conditional on an explanatory variable (country).

In estimating a mean parameter θ, a somewhat more appealing suggestion is to consider a random effects model, where $Y = \theta + u + e$, and u is a country random effect, then to develop a census estimating function system which is efficient for estimating the parameter θ. For example, if

$$Y_{1i} = \theta + u_1 + e_{1i} \quad \text{and} \quad Y_{2i} = \theta + u_2 + e_{2i}$$

and if the variance components corresponding to u and e are known, then the best combination of the two country means for estimating θ is

$$a\bar{Y}_1 + (1 - a)\bar{Y}_2,$$

where

$$a = \frac{1}{2}\left\{\frac{\sigma_u^2 + \sigma_e^2/N_2}{\sigma_u^2 + \frac{\sigma_e^2}{2N_1} + \frac{\sigma_e^2}{2N_2}}\right\}.$$

Notice that if $\sigma_u^2 = 0$, the census estimator becomes the mean of y over the two countries combined. However, if σ_u^2 is dominant, the best estimator is the arithmetic mean of the country means. From a pooled sample, the usual paradigm gives the same convex combination of within-country sample-based mean estimators.

More generally, we can replace θ in each country census estimating function by $\theta + u$, where again u is a country random effect. Then the best combination of two country census estimating functions for θ is

$$c_1 \sum_{i=1}^{N_1} \phi_{1i}(Y_{1i}, \theta, u_1) + c_2 \sum_{i=1}^{N_2} \phi_{2i}(Y_{2i}, \theta, u_2)$$

where $c_1 = [\mathrm{Var}(\sum_{i=1}^{N_2} E(\phi_{2i} | u_2)) + E(\sum_{i=1}^{N_2} \mathrm{Var}(\phi_{2i} | u_2))][\sum_{i=1}^{N_1} E(\partial\phi_{1i}/\partial\theta)]$, and c_2 is defined symmetrically. If the first term in square brackets in c_1 dominates, the corresponding sample-based estimating function weights the terms comparably in the two samples.

Even in the simple case of a mean, the parameters of the random effects model will not be known, and will be difficult to estimate when there are only two countries, but conceptually the model seems to be a useful one. When there are several reasonably similar countries or regions (for example the seven cities of the ITC China survey), linear models with random effects are estimable in the usual paradigm, as described for example in a more general setting by Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998).

As an aside, the GEE analysis of the Ireland data described earlier was a pooled analysis, and all its "effects" were regarded as fixed. The model is nearly "saturated", with two time points and two countries accounting for the four main parameters. It is possible to see that with ordinary survey weights, the estimation of β and ρ would be dominated by the UK data. However, if β and ρ are known, then as in the case of the parameters θ and Δ in the example of the mean, the equations for the main parameters separate into two pairs, one pair for α_0 and γ, and one pair for $\alpha_0, \alpha_1, \gamma$ and δ, each involving weights from only one of the two countries. Thus the estimation of the main parameters is less affected by the scaling of the weights. If the estimation of β were also important to us, we might consider it to be the mean of a country level random variable, leading naturally to each of the two samples having appropriate influence. (In fact, in our analysis, we did not do this; the weights were rescaled to sum to sample size within country.)

The foregoing discussion of pooled analysis assumes that there exists a parameter θ that has the same interpretation and relevance across countries. Most multi-country analyses start from this assumption. Indeed, de Leeuw and Hox (2003) state as a requirement for a meta-analysis that "all studies must estimate the same fixed parameter, and all variance is assumed to be sampling variance". But in fact a central issue for debate is the question of when it is appropriate to make a model that is to apply to the data from several countries simultaneously. Sometimes it may be most appropriate simply to consider the country models to be separate but parallel. For example, in countries at different stages of development, introducing the same relative increase in real price of cigarettes can be expected to lead to decreases in cigarette consumption; but since the linear model is at best a useful local approximation to the complex relationship between price and consumption, there is no reason to suppose that the decreases will be of the same magnitude, or that the two regression estimates are measuring the same quantity.

For another example, one of the models of interest to the ITC project is the mediational model of the figure 2, postulating how "noticing" health warning labels might affect intention to quit.

The distribution at baseline of the intention to quit in the various countries is quite variable. The same is true for the other variables in the model. Is it reasonable to hope that the relationships among these variables might be less variable across countries? In fact it appears that for the original four countries, they are. Even though UK smokers were much the most likely to say they had no plan to quit, it was still the case for them that "health concern" (label-triggered consciousness of health effects) predicts quit intention, and that "health concern" was elevated with increased noticing of labels (Hammond, Fong, Borland, Cummings, McNeill and Driezen 2007). Thus it is not unreasonable to explore a model like the one in Figure 2 for the data from the four countries, pooled. Regardless of weighting issues, in the regression of the mediator "health concern" on "noticing" health warning labels, and in the regression of intention to quit on both of these, we have found it convenient to take the country means to be fixed effects. On the other hand, since the estimated regression coefficients for the countries modeled separately vary moderately, it is natural in the pooled analysis to conceptualize the regression coefficients as having random country components.

This discussion can be summarized and elaborated in the following points:

- An analysis which pools data across countries should be adopted with caution. For such an analysis to be appropriate, the model structure (the regression equation and its variables) should be correct for all countries, and the assumption of common parameters should be supported by theory and observation. Robust variance estimation which respects the country sampling designs will be necessary when the sampling designs are complex.
- If the set of parameters of a pooled model can (through transformation) be separated into disjoint subsets corresponding to the countries, the estimation of those parameters is not affected by large differences in sampling fractions among countries, and is not affected by rescalings of the weights within countries.
- If a fixed mean or regression parameter is deemed to be common to the countries, estimation using inverse inclusion probability weights will be inefficient if the sampling fractions are widely variable.
- Alternatives to simple weight rescaling are to make the mean or regression parameter a fixed effect varying by country (which leads to separation into disjoint subsets, but increases the number of parameters and removes the "common-ness"); or to make the mean or regression parameter a random effect, varying by country (which leads to approximate separation and retains the "common-ness").
- It is conceptually appealing to make the intercept fixed and the slope random, since baseline levels tend to vary much more by country than slopes do. In implementation, this approach requires enough countries to make estimation of variance components feasible, and a small number of random effects to be integrated.
- When a pooled analysis is problematic, less formal comparison of the results of parallel country analyses may accomplish most of what is desired.

4. Calibration of measurements and cross-cultural comparisons

The other statistical problem which I would like to highlight is the use of measurement models to try to calibrate measurements across modes and compare measurements across cultures. A common approach is to consider that with each questionnaire item we are measuring a construct, like "social denormalization" (perception of societal disapproval), and to think of the construct as a continuous variable η. The distribution of η, conditional on explanatory variables, determines a distribution of responses to the questionnaire item.

If we have several items of the same kind measuring a construct, a conceptual model for continuous measurements y might be $Y_{ik} = b_i \eta_k + a_i + e_{ik}$ for item i and participant k. Here b_i represents a positive scaling for item i, a_i a location shift, and e_{ik} a normal mean zero measurement error with variance σ^2_{ei} not dependent on k. Assume all e_{ik} to be independent, and independent of the η_k. (This is effectively an assumption that η is the only latent determinant of Y.) If we take the distribution of η to be $N(0, 1)$, as we may if η is normal with no explanatory variables, then the distribution of Y_{ik} is $N(a_i, b_i^2 + \sigma^2_{ei})$, and for a single item i, the parameters a_i and $b_i^2 + \sigma^2_{ei}$ are estimable from the marginal data on many participants. If there are at least two items with the same variances, then since the item responses for a participant have covariances of form $b_{i_1} b_{i_2}$, all parameters are estimable from the marginal data on many participants. Given values for the item parameters, the value of η for a participant can be "predicted" from the posterior distribution of η, given the participant's item responses.

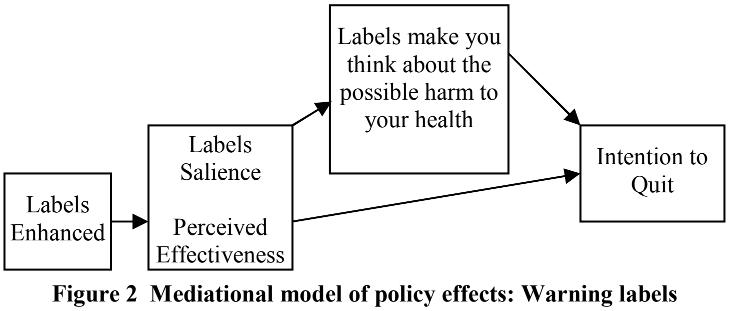

Figure 2 Mediational model of policy effects: Warning labels

If the measurement y is binary, it is common to take an Item Response Theory (IRT) model $\text{Prob}(Y_{ik} = 1|\eta_k) = H(b_i\eta_k - \gamma_i)$ where H is the standard normal or the logistic cumulative distribution function (c.d.f.). The parameter b_i is the “discrimination parameter” for the item, and γ_i is a threshold, such that the probability of response 1 exceeds 1/2 when the construct scaled by b_i exceeds γ_i. The unconditional probability that $Y_{ik} = 1$ is obtained by integrating with respect to the distribution of η_k, given fixed explanatory variables for participant k. In this simplest case it appears that at least 3 items are needed (with many participants) for all parameters to be estimable, since they would yield 7 joint probabilities for the estimation of 6 parameters. Again, given values for the item parameters, the value of η for a participant can be predicted, given his or her set of item responses. See for example Lu, Thomas and Zumbo (2005). Standard latent variable estimation software can be used to produce these inferences, and their analogues in the case of ordinal measurements.

First let us consider the calibration problem. Suppose there are two data collection modes, and for item i in mode j with participant k we have the continuous measurement

$$Y_{ijk} = \beta_j(b_i\eta_k + a_i + e_{ijk}) + \alpha_j + \varepsilon_{ijk}.$$

This model, in which α_j and β_j do not depend on the item i, might be appropriate for a set of items all of the same general type. Plausible examples are not abundant, but one such might be a series of questions of form: “What percentage of the time would you say you feel …”, where the respondent is asked to give a percentage over the telephone, or asked to mark a position on a line on paper.

If we take the a_i and b_i to be the parameters of the items using the first data collection mode, we may set $\alpha_1 = 0$ and $\beta_1 = 1$. If β_2 is greater than 1, there is a tendency for a wider variation, or more extreme responses, under the second collection mode. If α_2 is greater than 0, respondents tend to give a higher response under the second collection mode than under the first. Note that the samples for the two modes involve different participants. If we can assume that the distribution of η is the same for the two mode samples (an assumption which effectively requires randomization to mode), we have the distribution of Y_{i1k} as before, $N(a_i, b_i^2 + \sigma_{ei}^2 + \sigma_\varepsilon^2)$, while the distribution of Y_{i2k} is

$$N(\beta_2 a_i + \alpha_2, \beta_2^2 b_i^2 + \beta_2^2 \sigma_{ei}^2 + \sigma_\varepsilon^2).$$

If $\sigma_\varepsilon^2 = 0$, then given data on one item i in the two modes, we can estimate α_2 and β_2, assuming β_2 is positive. If $\sigma_\varepsilon^2 > 0$, the parameters α_2, β_2 and σ_ε^2 are estimable provided that there are at least two items available $-$ of the same type, but with differing values of a and b.

These considerations can be extended to the more usual case of items with ordinal responses, by imagining an ordinal response probability to be determined by an underlying continuous response. For binary data, we would most simply set

$$P(Y_{ijk} = 1|\eta_k) = H(\beta_j(b_i\eta_k + a_i) + \alpha_j),$$

with $\alpha_1 = 0$ and $\beta_1 = 1$. If the distribution of η_k is the same for both modes, then from data on many participants and three items we can identify all parameters. Adding an explanatory variable would decrease the number of items required.

The assumption that the distribution of η is the same for the two mode samples is crucial for this kind of calibration, and is difficult to guarantee. It is satisfied if we have interpenetrating probability samples for the two modes in a single survey; then in principle we can imagine a mapping of responses from one mode to the other, through estimated values of α_2 and β_2. We do not have to estimate the constructs themselves to do this. More rigorously, we can include α_2 and β_2 as parameters in a model for all responses to a set of similar items.

In some developed countries, sampling frames for households and individuals appear to be moving in the direction of address registries and lists of persons. However, even when there is a common frame for (say) telephone and internet surveys, it is difficult to randomize respondents to data collection modes. The dependence of non-response on demographic variables may well be different for the modes. Moreover, the need to maximize response rates often dictates allowing respondents to choose. In principle, we might imagine that the distribution of η might be shifted or tilted according to the “propensity” to choose one mode or the other. Having modeled this propensity in terms of explanatory variables, and having introduced one or two parameters for the dependence of the distribution of η on the propensity, we could estimate the item parameters a_i and b_i from the respondents for the first data collection mode. The estimation of the mapping parameters α and β would follow in the same manner as before.

In another kind of circumstance, we might use the two data collection modes in different groups of the population. In that case, the mode effect becomes part of the group effect; it cannot be disentangled from an underlying difference in the distribution of the construct.

Trying to compare measurements across cultures or other groups is different from the calibration problem, since randomizing participants to groups, to keep the distribution of the construct constant, is out of the question. The common wisdom is that to compare the mean of a construct from one group to another, the measuring items must have the same relationship with the construct in the two groups.

When there are several constructs, to compare the relationship among constructs from one group to another requires a kind of "measurement invariance" or equivalence for all items involved. There is a vast literature on cross-cultural comparisons and measurement. For example, Johnson (1998) lists fifty-two terms for cross-cultural equivalence that have been introduced by authors in various disciplines.

The multi-group confirmatory factor analysis model is useful for continuous measurement items, and takes the form:

$$Y_k^g = \tau^g + \Lambda^g \eta_k^g + e_k^g$$

where Y_k^g is the vector of observed responses to the items for respondent k in group g; Λ^g is a matrix of slopes or "factor loadings"; the intercept vector τ^g indicates the expected value of Y_k^g when $\eta_k^g = 0$; and e_k^g is a measurement error with 0 mean. Then $E(Y_k^g) = \tau^g + \Lambda^g \kappa^g$, where κ^g is the mean of the construct η in group g. The variance-covariance matrix among the observed values y_k^g can be expressed as $V(Y_k^g) = \Lambda^g \Phi^g \Lambda^{g'} + \Theta^g$, where Φ^g is the covariance matrix of the latent constructs and Θ^g is the diagonal matrix of measurement error variances. See de Jong, Steenkamp and Fox (2007), Davidov (2008) and references therein.

The IRT version of the model can be defined in straightforward manner. Using the same parameter notation, in the case of binary items, we have

$$P(Y_k^g = 1 \mid \eta_k^g) = H(\tau^g + \Lambda^g \eta_k^g),$$

and there is a natural extension to the ordinal case.

The model parameters are not identifiable unless some restrictions are made. In the multi-group confirmatory factor analysis model, many authors postulate a "marker" item for each construct, with a factor loading of 1 and an intercept of 0 for all groups, so that the mean of the construct is identified in each group. This is a very strong assumption. Alternatively, we might imagine choosing the units for the constructs so that they are marginally N(0,1) within group 1. The parameters of the items (with sufficiently many items) are thus identified for group 1. If the variances and relationships of the path diagram are assumed to remain true in group 2, then we can test whether the item parameters also remain the same, and if not, try to redesign the set of items to produce something closer to measurement invariance. On the other hand, if the item parameters are constrained to remain the same, we can test whether the underlying joint distribution of the constructs is also the same. However, formal rejection of the null hypothesis is difficult to interpret. Following Rensvold and Cheung (1998), Barrera Ceballos (2007) has carried out this kind of multi-group analysis for the data of the ITC Mexico survey and the ITC Uruguay survey, replacing "health concern" in the model of Figure 2 by "social denormalization", or the extent to which the respondent perceives society to disapprove of tobacco use. (The other two constructs are warning label salience and intention to quit.) The relationships appear unexpectedly different in the two countries under the constraints of measurement item invariance, a finding which could be due either to real societal differences or to an imperfect correspondence between the items themselves (*i.e.*, failure of the constraints). Admittedly, with very few constructs having multiple items, the ITC survey instrument was not designed for this kind of analysis,

Ultimately the relationships among the constructs are of paramount importance, along with the question of whether the relationships of the constructs can be said to be alike, though not necessarily identical, from group to group. This is so regardless of whether the marginal distributions of the constructs are the same, or whether measurement items have the same parameters from place to place, or mode to mode. Intuitively, the two kinds of restrictions of the previous paragraph seem too strong. A hierarchical approach of De Jong *et al.* (2007) offers a way forward.

If item i has C ordered response options, we can write

$$P(Y_{ik}^g = c \mid \eta_k^g, b_i^g, \gamma_{i,c}^g, \gamma_{i,c-1}^g) = H(b_i^g \eta_k^g - \gamma_{i,c-1}^g) - H(b_i^g \eta_k^g - \gamma_{i,c}^g),$$

$c = 1, \ldots, C$. Here the factor loadings are replaced by the discrimination parameters b, and the intercepts are replaced by the thresholds γ. Instead of insisting that these parameters are independent of group label before proceeding, the approach is to model them with group-specific random effects:

$$\gamma_{i,c}^g = \gamma_{i,c} + e_{i,c}^g,\ e_{i,c}^g \sim N(0, \sigma_{\gamma_i}^2),$$

$$b_i^g = b_i + r_i^g,\ r_i^g \sim N(0, \sigma_{b_i}^2).$$

The heterogeneity in the latent variable is modeled by a hierarchical structure:

$$\eta_k^g = \kappa^g + v_k^g,\ v_k^g \sim N(0, \sigma_g^2),$$

$$\kappa^g \sim N(\kappa, \xi^2).$$

With sufficiently many items, such a model is estimable, and can be fitted using Markov Chain Monte Carlo methods. The invariance tests of multi-group analysis can still be performed within this framework.

5. Discussion and conclusions

Again, the aim of analysis in the ITC context must be to discern the real response (or lack of response) to policy change, separating it from the effects of data collection mode, differential non-response, external events, time-in-sample, culture, and language. It may not be necessary to distinguish among all of the confounders, but it is important to allow them to contribute to the model. In this paper we have not addressed external events, which can be modelled in an obvious way if recognized. We have not discussed modelling attrition or time-in-sample effects in detail, but in principle, each one of them can be regarded as part of the mix. Those who are retained from wave to wave of a survey might be regarded as a kind of cultural group. On the other hand, time-in-sample effects are a particular kind of failure of measurement invariance, over time rather than from one group to another. A comprehensive analysis would take account of these, and of other effects of culture, language and data collection mode.

It is by no means the case that the effects of policy would always be identifiable in a full model. But the chances increase if the design involves between country comparisons of longitudinal data, and the replication which comes from observing cohorts with different starting points.

A unifying thread of the two previous sections is the introduction of random effects as a device. The device of introducing random effects for countries and groups in key parameters is natural, and (for large group samples) conceptually compatible with traditional survey analysis, based on weighted estimating functions. There are some obstacles to practical implementation, arising from identifiability and estimability limitations, and the calculation of likelihood functions if more than a few random effects are entertained. At the same time, with increasing availability of numerical methods to handle such models, further research to adapt them to complex international surveys should be very fruitful.

Acknowledgments

This work is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The ITC project is supported by grants from the National Cancer Institute of the United States (P50 CA11236) Roswell Park Transdisciplinary Tobacco Use Research Center and the Canadian Institutes of Health Research (57897). Thanks are extended to an anonymous reviewer for very helpful comments.

References

Barrera Ceballos, J.A. (2007). Cross-national comparison of the impact of cigarette warning labels and social denormalization on intention to quit from the International Tobacco Control Survey. Research paper. Statistics and Actuarial Science, University of Waterloo.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Bronner, K., and Kuijlen, T. (2007). The live or digital interviewer: A comparison between CASI, CAPI and CATI with respect to differences in response behaviour. *International Journal of Market Research*, 49, 167-190.

Cleland, J., and Verma, V. (1989). The World Fertility Survey: An appraisal of methodology. *Journal of the American Statistical Association*, 84, 756-767.

Cook, T., and Campbell, D. (1979). *Quasi-Experimentation*. Chicago: Rand McNally.

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European Social Survey. *Survey Research Methods*, 2, 33-46.

De Jong, M.G., Steenkamp, J.-B.E.M. and Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical irt model. *Journal of Consumer Research*, 34, 260-278.

de Leeuw, E.D., and Hox, J. (2003). The use of meta-analysis in cross-national studies. In *Cross-Cultural Survey Methods* (Eds., J.A. Harkness, F.J.R. Van de Vijver and P. Ph. Mohler). Hoboken, NJ: Wiley, 329-346.

ESS (2008). European Social Survey. http://www.europeansocialsurvey.org/.

Fong, G.T., Cummings, K.M., Borland, R., Hastings, G., Hyland, A., Giovino, G.A., Hammond, D. and Thompson, M.E. (2006). The conceptual framework of the International Tobacco Control Policy Evaluation Project. *Tobacco Control*, 15(Supp 3): iii3-iii11.

Fong, G.T., Hyland, A., Borland, R., Hammond, D., Hastings, G., McNeill, A., Anderson, S., Cummings, K.M., Allwright, S., Mulcahy, M., Howell, F., Clancy, L., Thompson, M.E., Connolly, G. and Driezen, P. (2006). Reductions in tobacco smoke pollution and increases in support for smoke-free public places following the implementation of comprehensive smoke-free workplace legislation in the Republic of Ireland: Findings from the ITC Ireland/UK Survey. *Tobacco Control*, 15(Supp. 3): iii51-iii58.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey populations: Their relationships and estimation. *International Statistical Review*, 54, 127-138.

Hammond, D., Fong, G.T., Borland, R., Cummings, K.M., McNeill, A. and Driezen, P. (2007). Text and graphic warnings on cigarette packages: Findings from the International Tobacco Control Four Country Study. *American Journal of Preventive Medicine*, 32, 210-217.

Hofstede, G. (1980). *Culture's Consequences. International Differences in Work-Related Values*. Beverly Hills, CA: Sage.

Hyland, A., Hassan, L., Higbee, C., Fong, G.T., Borland, R., Cummings, K.M., Thompson, M., Boudreau, C. and Hastings, G. (2008). The impact of smokefree legislation in Scotland: Results from the Scottish International Tobacco Control Policy Evaluation Project. In progress.

ITC (2008). Tobacco Labelling Resource Centre. http://www.igloo.org/tobacco_labelling. Accessed April 24, 2008.

Johnson, T.P., and Van de Vijver, F.J.R. (2003). Social desirability in cross-cultural research. In *Cross-Cultural Survey Methods* (Eds., J.A. Harkness, F.J.R. Van de Vijver and P. Ph. Mohler). Hoboken, NJ: Wiley, 195-204.

Johnson, T.P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. In *Cross-Cultural Survey Equivalence* (Ed., J.A. Harkness). *ZUMA-Nachrichten Spezial 3*. Mannheim: ZUMA, 1-40.

Lohr, S.L., and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lu, I.R.R., Thomas, D.R. and Zumbo, B.D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling*, 12, 263-277.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Series B, 60, 23-40.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā*, series B, 61, 166-186.

Rensvold, R.B., and Cheung, G.W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, 58, 1017-1034.

Scott, A. (2006). Population-based case control studies. *Survey Methodology*, 32, 123-132.

Rosenbaum, P.R., and Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-542.

Ryder, A.G., Yang, J., Zhu, X., Yao, S., Yi, J., Heine, S.J. and Bagby, R.M. (2008). The cultural shaping of depression: Somatic symptoms in China, psychological symptoms in North America? *Journal of Abnormal Psychology*, 117, 300-313.

Skinner, C. (1989). Introduction to Part A. In *Analysis of Complex Surveys* (Eds., C. Skinner, D. Holt and T.M.F. Smith), Chichester: Wiley. 2.

Thompson, M.E., Boudreau, C. and Driezen, P. (2005). Incorporating time-in-sample in longitudinal survey models. *Proceedings*: *Symposium 2005*, *Methodological Challenges for Future Information Needs.* Session 12: Challenges in Using Data from Longitudinal Surveys. Statistics Canada.

Thompson, M.E., Fong, G.T., Hammond, D., Boudreau, C., Dreizen, P., Hyland, A., Borland, R., Cummings, K.M., Hastings, G.B., Siahpush, M., Mackintosh, A.M. and Laux, F.L. (2006). Methods of the International Tobacco Control (ITC) Four Country Survey. *Tobacco Control*, 15(Supp 3): iii12-iii18.

Thrasher, J., Quah, A., Borland, R., Awang, R., Sirirassamee, B., Boado, M., Miller, K., Watts, A. and Dorantes, A. (2008). Ensuring valid cross-cultural comparisons in survey research on tobacco: Development, implementation, and results from a transnational cognitive interviewing study. In progress.

Verma, V., Scott, C. and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, Series A, 143, 431-473.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

WHO (2008). *WHO Report on the Global Tobacco Epidemic*, *2008*. http://www.who.int/tobacco/mpower/mpower_report_full_2008.pdf Accessed April 14, 2008.

Wichers, B., and Zengerink, E. (2006). It's the culture, stupid! A cross-cultural comparison of data collection methods. *Panel Research 2006, Part 4/The respondent - Cross cultural insights*, ESOMAR.