Composante du produit nº 11-522-X au catalogue de Statistique Canada
La série des symposiums internationaux de Statistique Canada : recueil

Article

Symposium 2008 :
Collecte des données : défis, réalisations et nouvelles orientations

Enquêtes internationales : motifs et méthodologies

par Mary E. Thompson

2009

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2008 de Statistique Canada
Collecte des données : défis, réalisations et nouvelles orientations

Enquêtes internationales : motifs et méthodologies[1]

Mary E. Thompson [2]

Résumé

Le contexte de la discussion est la fréquence croissante des enquêtes internationales, dont fait partie l'International Tobacco Control (ITC) Policy Evaluation Project, qui a débuté en 2002. Les enquêtes ITC nationales, qui sont longitudinales, ont pour but d'évaluer les effets des mesures stratégiques adoptées dans divers pays en vertu de la Convention-cadre pour la lutte antitabac de l'OMS. Nous examinons et illustrons les défis que posent l'organisation, la collecte des données et l'analyse des enquêtes internationales. L'analyse est une raison qui motive de plus en plus la réalisation d'enquêtes transculturelles à grande échelle. La difficulté fondamentale de l'analyse est de discerner la réponse réelle (ou le manque de réponse) aux changements de politiques et de la séparer des effets du mode de collecte des données, de la non-réponse différentielle, des événements extérieurs, de la durée de la présence dans l'échantillon, de la culture et de la langue. Deux problèmes ayant trait à l'analyse statistique sont examinés. Le premier est celui de savoir quand et comment analyser des données regroupées provenant de plusieurs pays, afin de renforcer des conclusions qui pourraient être généralement valides. Bien que cela paraisse simple, dans certains cas les avis sont partagés quant à la mesure dans laquelle ce regroupement est possible et raisonnable. Selon certains, les modèles à effets aléatoires sont conceptuellement utiles pour les comparaisons formelles. Le deuxième problème consiste à trouver des modèles de mesure applicables à diverses cultures et à divers modes de collecte de données qui permettent l'étalonnage des réponses continues, binaires et ordinales, ainsi que la production de comparaisons dont ont été éliminés les effets extérieurs. Nous constatons que les modèles hiérarchiques offrent un moyen naturel de relâcher les exigences d'invariance du modèle entre les groupes.

Mots clés : Enquêtes internationales, enquêtes longitudinales, analyse de données d'enquête, effets aléatoires, effets du mode de collecte des données, modèles hiérarchiques, modèles de mesure.

1. Introduction

J'ai choisi comme thème de mon exposé les enquêtes internationales, parce que les quelques dernières années, une grande part de mon activité a été consacrée à l'une de ces enquêtes, l'International Tobacco Control survey, et parce qu'il existe certains recoupements intéressants avec les domaines auxquels s'est intéressé Joseph Waksberg, en particulier les bases de sondage pour les enquêtes téléphoniques et les effets de la stratification avec des taux d'échantillonnage très variables.
L'article débute par une discussion des raisons qui motivent les enquêtes internationales et par certains exemples. Puis, il traite des défis que posent l'organisation, la collecte des données et l'analyse. Enfin, il aborde deux problèmes qu'il convient de résoudre dans le cadre de l'analyse, à savoir i) celui de la théorie de l'échantillonnage et du regroupement de données provenant de plusieurs pays et ii) celui de la mesure dans le contexte de différents modes de collecte des données et de différentes cultures.

La première grande enquête internationale a été l'Enquête mondiale sur la fécondité (EMF), menée durant les années 1970 par l'entremise de l'Institut international de statistique et financée par l'Agency for International Development des États-Unis et d'autres organismes parrains. Il s'agissait d'une enquête ponctuelle très ambitieuse. En fin de compte, l'EMF a été réalisée auprès de 330 000 femmes dans 61 pays, au coût d'environ 50 millions de dollars. Elle a fourni aux pays participants d'importantes données comparatives sur les tailles des familles et a abouti à l'adoption d'une politique de planification démographique dans plusieurs pays participants. Elle a également donné naissance à des centaines de projets analytiques, dont certaines études méthodologiques inédites, et a jeté les fondations de la méthodologie d'enquête internationale, particulièrement dans les pays en voie de développement (Verma, Scott et O'Muircheartaigh 1980 ; Cleland et Verma 1989).

---

[1] Cet article a initialement paru dans la livraison de décembre 2008 de Techniques d'enquêtes (Volume 34, No 2, pp 145-157). Il est republié ici dans ce recueil avec la permission des éditeurs

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

Un autre exemple bien connu est le Programme international pour le suivi des acquis des élèves, un projet de l'Organisation de coopération et de développement économiques, qui a été lancé en 2000. Le PISA est une enquête continue, exécutée tous les trois ans, auprès des jeunes de 15 ans dans les pays développés. Sa portée s'accroît, 67 pays devant, en principe, y participer en 2009. Les résultats permettent aux pays de surveiller le succès de leurs programmes d'enseignement en ce qui concerne l'acquisition de compétences en communication verbale et en compréhension de texte à contenu quantitatif.

La Global Youth Tobacco Survey (GYTS) est une enquête ponctuelle parrainée par l'Organisation mondiale de la santé et les Centers for Disease Control and Prevention qui a débuté en 2002. Elle a pour cible les jeunes de 13 à 15 ans des pays en voie de développement et, en 2004, des données avaient été recueillies dans 129 pays. L'objectif est de mesurer l'adoption du tabagisme chez les jeunes et la sensibilisation aux risques connexes pour la santé.

L'Enquête sociale européenne (ESE 2008) est une « enquête sociale appuyée par les universités » réalisée dans plus de 30 pays, financée par des organismes européens et nationaux, et conçue pour « expliquer l'interaction entre les institutions européennes en pleine évolution et les attitudes, les croyances et les comportements des diverses populations ».

Si le recours à des enquêtes locales et nationales est en hausse partout, il en est de même de la fréquence des enquêtes internationales, exécutées par des organismes internationaux, des organisations non gouvernementales et des entreprises du secteur privé. Cet essor semble s'inscrire dans une tendance vers une gouvernance mondiale et une préoccupation pour la santé et le bien-être des populations.

J'ai vu les objectifs des enquêtes internationales classés sous les rubriques de l'épidémiologie, de la surveillance et de l'évaluation des effets des politiques. De toute évidence, ces classifications se chevauchent. On peut soutenir que le PISA, la GYTS et l'ESE rentrent dans les catégories de la surveillance, parce que leurs données ne sont reliées qu'indirectement aux interventions. Par contre, l'EMF comporte un aspect d'évaluation directe dans les pays qui ont mis sur pied des programmes de planification familiale. L'International Tobacco Control (ITC) survey, dont il sera question plus loin dans cette section, est l'une des rares enquêtes dont l'objectif principal est l'évaluation.

Outre les préoccupations d'ordre scientifique, l'un des rôles importants d'une enquête internationale consiste à éveiller l'intérêt des gouvernements ; elle leur offre un moyen de participer à l'élaboration de politiques mondiales, même quand ils font face à des obstacles politiques et économiques.

Aux chercheurs, les enquêtes internationales permettent de comparer les populations de divers pays, d'interpréter les différences, voire même de mieux en comprendre les causes et les effets – habituellement avec l'objectif sous-jacent d'améliorer les conditions existantes et d'appuyer l'élaboration des politiques.

L'International Tobacco Control Policy Evaluation Project (projet ITC) a été lancé par Geoffrey T. Fong, Ph.D., du Département de psychologie à l'Université de Waterloo, et par des collaborateurs partout dans le monde (Fong, Cummings, Borland, Hastings, Hyland, Giovino, Hammond et Thompson 2006 ; Thompson, Fong, Hammond, Boudreau, Dreizen, Hyland, Borland, Cummings, Hastings, Siahpush, Mackintosh et Laux 2006). L'élan a été donné par la Convention-cadre sur la lutte antitabac (CCLA) de l'OMS, qui a été adoptée en mai 2003 et ratifiée par plus de 150 pays. En ratifiant le traité, les pays participants ont promis des mesures stratégiques de lutte antitabac, telles que l'apposition d'étiquettes portant des mises en garde musclées concernant le danger pour la santé, l'interdiction de la publicité au profit des cigarettes et l'interdiction de fumer dans les lieux publics. Parce que ces mesures requièrent une législation nationale, le moment où elles sont adoptées et la façon dont elles le sont varient. Par exemple, en décembre 2000, le Canada a commencé à utiliser des étiquettes de mise en garde explicites, créant un précédent international en ce qui concerne la taille de l'étiquette (plus de 50 % de la surface de l'emballage) et l'utilisation d'images aux couleurs vives. Depuis, quelques autres pays ont opté pour la même pratique, tandis que d'autres ont adopté des lois prévoyant l'utilisation d'avertissements textuels bien en vue. Pour des renseignements sur la situation actuelle de la réglementation concernant les mises en garde relatives à la santé dans le monde, consulter ITC (2008). Le rapport MPower (OMS 2008) décrit l'environnement stratégique mondial concernant la lutte antitabac et six stratégies recommandées pour la CCLA.

Le but du projet ITC est d'essayer de trouver des stratégies qui réduisent effectivement le taux d'adoption du tabagisme et qui aident les jeunes qui fument déjà à arrêter de le faire. S'ajoute à cela l'objectif ambitieux d'essayer d'expliquer comment fonctionnent vraiment les stratégies efficaces. L'équipe de recherche comprend des psychologues sociaux et des spécialistes du marketing social, ainsi que des épidémiologistes et des économistes.

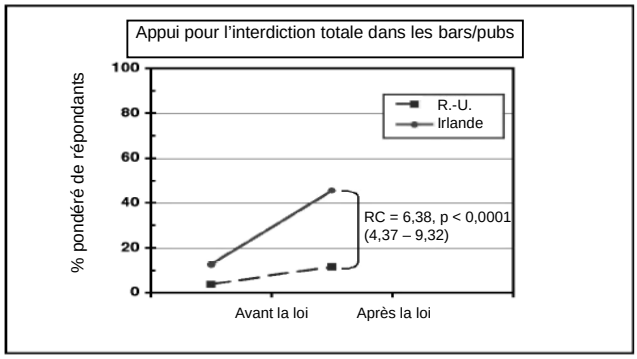
En septembre 2008, 17 pays réalisaient des enquêtes dans le cadre du projet ITC et d'autres s'ajouteront vraisemblablement à la liste. Les enquêtes ont débuté en 2002 au Canada, aux États-Unis, au Royaume-Uni et en Australie. Cette année-là, dans chacun de ces quatre pays, environ 2 000 fumeurs adultes ont été sélectionnés par téléphone en utilisant une base de sondage par composition aléatoire (CA) stratifiée géographiquement, dont le fondement scientifique a pour origine la fameuse méthode de Mitofsky-Waksberg (Waksberg 1978). Les fumeurs sélectionnés ont été interviewés une semaine ou deux plus tard, puis ont fait l'objet d'un suivi chaque année depuis, qu'ils aient continué ou non de fumer. La sixième vague de l'enquête ITC-Quatre pays s'est achevée en février 2008.

Comme une taille d'échantillon suffisante est nécessaire pour évaluer les effets des mesures introduites entre les vagues de l'enquête, les participants qui décrochent à chaque vague sont remplacés par une cohorte de nouvelles personnes. Dans le cadre de l'enquête ICT-Quatre pays, à chaque vague, les nouveaux participants ont été sélectionnées en utilisant le même plan de sondage qu'à la première vague, sans essayer de rechercher des personnes ayant les mêmes caractéristiques que celles ayant décroché. À chaque vague, la construction des poids est, en fait, exécutée séparément pour chaque cohorte, en procédant à un rajustement pour tenir compte des différences d'attrition selon la région et le groupe âge-sexe. Ce plan de sondage nous a permis de discerner des effets de « durée de la présence dans l'échantillon », si bien que cette caractéristique est introduite en tant que variable explicative dans les modèles analytiques (Thompson, Boudreau et Driezen 2005).

Après 2002, les premières mesures stratégiques nationales qui ont été prises étaient l'interdiction de la publicité et une amélioration des étiquettes de mise en garde entre les première et deuxième vagues au Royaume-Uni, ainsi que l'adoption d'étiquettes de mise en garde explicites entre les quatrième et cinquième vagues en Australie. L'enquête ITC-Quatre pays constitue ce que l'on a parfois appelé une expérience naturelle ou quasi-expérience (Cook et Campbell 1979), les pays où une politique particulière n'existe pas servant de contrôles externes ; en outre, l'aspect longitudinal du plan de sondage offre un contrôle interne. Le plan de sondage a été répété un certain nombre de fois, pour d'autres groupes de pays.

Par exemple, très vite, il est devenu évident que l'Irlande serait le premier pays à adopter une loi antitabac nationale, laquelle est entrée en vigueur en mars 2004. Les collaborateurs du projet ITC ont réussi à mettre sur pied des enquêtes parallèles en Irlande et au Royaume-Uni avant que la loi n'entre en vigueur et à rendre visite aux mêmes personnes une année plus tard. De nouveau, les échantillons ont été sélectionnés nationalement à l'aide d'une base de sondage par composition aléatoire (CA). En tout, 755 fumeurs ont été interviewés lors des deux vagues en Irlande, et 411 au Royaume-Uni. Une constatation intéressante concerne l'appui pour une interdiction de l'usage du tabac dans les pubs (Fong, Hyland, Borland, Hammond, Hastings, McNeill, Anderson, Cummings, Allwright, Mulcahy, Howell, Clancy, Thompson, Connolly et Driezen 2006). La figure 1-1 donne les proportions de personnes appuyant ou appuyant fortement l'interdiction dans les bars et les pubs, dans les deux pays, selon la vague.

Figure 1-1
Appui pour la loi antitabac durant deux vagues de l'enquête ITC en Irlande et au Royaume-Uni



Dans l'échantillon ITC de fumeurs, l'accroissement de l'appui pour l'interdiction entre les deux vagues a été faible au Royaume-Uni et important en Irlande. En outre, l'enquête n'a donné aucune preuve que la réduction de l'usage du tabac dans les lieux publics était associée à un accroissement dans les lieux privés. Révélant l'acceptation générale de la loi antitabac par les fumeurs, les résultats de l'enquête ITC et d'autres résultats comparables ont facilité l'adoption de lois

semblables en Écosse, en France, en Allemagne, dans le reste du Royaume-Uni et aux Pays-Bas. Une enquête ITC a été réalisée avant et après la mise en application de l'interdiction, en avril 2006, en Écosse, en utilisant le reste du Royaume-Uni comme contrôle, et les résultats ont été reproduits, sauf qu'à ce moment-là, l'appui dans le reste du Royaume-Uni avait augmenté considérablement (Hyland, Hassan, Higbee, Fong, Borland, Cummings, Thompson, Boudreau et Hastings 2008).

Le modèle utilisé pour le test était simple, à savoir un modèle EEG, où Y est une mesure binaire de l'appui pour l'interdiction, w désigne le pays, t représente le temps, le terme wt représente une interaction, et x est un vecteur de covariables fixes au niveau individuel :

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

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

Le coefficient δ représente la différence d'accroissement de l'appui dans les deux pays, et nous avons testé l'hypothèse H_0: $\delta = 0$. D'autres paramétrisations sont possibles, mais celle-ci a l'avantage de concorder avec le graphique de la figure 1-1, qui représente des proportions marginales ; la méthodologie est généralement reconnue et prise en charge par les logiciels pour données d'enquêtes complexes.

2. Défis

La réalisation d'une enquête internationale pose de nombreux défis. Les articles sur l'EMF publiés par Verma et coll. (1980) et par Cleland et Verma (1989) contiennent des discussions réfléchies à ce sujet qui ne sont guère dépassées aujourd'hui. À la présente section, en guise d'illustration, nous décrivons certains problèmes survenus durant l'organisation et la collecte des données de l'enquête ITC.

Contrairement à l'EMF, l'enquête ITC a été financée, en premier lieu, par des programmes nationaux de subventions, principalement les National Institutes of Health des États-Unis et les Instituts de recherche en santé du Canada. L'infrastructure centrale, ayant à sa tête M. Fong à l'Université de Waterloo et M. K. Michael Cummings au Roswell Park Cancer Institute à Buffalo, collabore directement avec les organismes des divers pays. Nous avons dû apprendre à travailler avec des groupes provenant de sociétés, de régimes politiques et de cultures très différents. Rien que les coûts et les budgets de l'enquête diffèrent étonnamment d'un pays à l'autre. Quand les gouvernements participent au financement, ils ont leurs propres exigences et des ententes concernant la propriété des données doivent être négociées. Puisque les niveaux d'infrastructure et d'expertise peuvent différer assez bien d'un pays à l'autre, la coordination étroite qui existe dans le cas de l'enquête ITC-Quatre pays est difficile à reproduire à une plus grande échelle.

Par exemple, durant la première moitié de 2008, le travail sur le terrain a été effectué pour la troisième vague d'une enquête parallèle (l'enquête ITC-Asie du Sud-Est) en Thaïlande et en Malaisie, pays qui sont géographiquement proches et semblables à certains égards, mais qui diffèrent en ce qui a trait à de nombreux aspects. Du point de vue ethnique, la Thaïlande est assez homogène, alors que la Malaisie compte trois grands groupes ethniques et de nombreux groupes mineurs. En Thaïlande, plus de la moitié de la population vit dans les régions rurales, mais en Malaisie, la plupart de la population est urbaine et la mobilité résidentielle est forte. La Thaïlande a accumulé une grande expérience des enquêtes, y compris les études de cohorte, mais quand la première vague de l'enquête a démarré en 2005, la Malaisie entreprenait ce genre d'étude de cohorte pour la première fois. Nous avons essayé de recommander des plans d'échantillonnage parallèles dans les deux pays, mais nous avons dû faire des compromis. Par exemple, il a été découvert, au moment de la première vague, que les bases de sondage officielles étaient constituées, au niveau le plus bas, de blocs élémentaires de tailles différentes correspondant à des grappes de ménages. Cette différence a rendu l'échantillon de ménages plus dispersé en Malaisie, où les blocs étaient plus petits. La dispersion plus forte signifiait plus de travail et des coûts plus élevés. (Les effets de plan sont encore plus importants pour la Malaisie que pour la Thaïlande, à cause de la plus forte hétérogénéité au niveau des unités de premier degré.)

Un aspect important du projet consiste à essayer d'accroître la capacité d'exécution d'enquêtes sur la santé longitudinales dans des pays où ce genre de travail est relativement nouveau. Nous fournissons des protocoles détaillés, des manuels de formation et des modèles de saisie des données. Nous avons appris à insister davantage sur la détermination des compétences locales, particulièrement en statistique. La communication quotidienne se fait par courriels et par

téléconférences. L'épuration des données finales et la construction des poids d'enquête se font normalement à l'Université de Waterloo, mais les équipes de certains pays désirent vivement participer à ces étapes des opérations.

Nous procédons à des enquêtes téléphoniques, avec sélection des participants par CA modifiée dans les quatre pays originaux, ainsi qu'en Irlande, en Corée du Sud, en France et en Allemagne et avec sélection à partir de l'échantillon de la National Health Survey en Nouvelle-Zélande, ainsi qu'à des enquêtes par interview sur place en Thaïlande, en Malaisie pour la première vague, en Chine, au Bangladesh, au Mexique et en Uruguay.

En Malaisie, pour la deuxième vague, nous avions l'intention de procéder à des interviews sur place, mais comme la reprise de contact ainsi que la sélection de nouveaux participants se sont avérées difficiles à cause d'une combinaison de facteurs, nous sommes passés à l'interview téléphonique dans la mesure du possible dans certaines régions. La possibilité de comparer les modes de collecte était limitée, mais dans le grand État principalement urbain de Selangor, 137 fumeurs interrogés à la première vague (qui n'avaient pas arrêté) ont été réinterviewés sur place, et 63 ont été réinterviewés par téléphone, ce qui rend possible certaines inférences provisoires. Pour la troisième vague, nous avons essayé de procéder à des interviews par téléphone et sur place dans certains des mêmes districts de recensement, en vue de permettre une meilleure évaluation des effets du mode de collecte des données, et cette étude est en cours. Parallèlement, la proportion de l'échantillon de fumeurs de l'enquête ITC-Malaisie interviewés sur place a diminué régulièrement, pour passer de 100 % à la première vague à 63,5 % à la deuxième, et à 44,4 % à la troisième. Pour la quatrième vague, nous prévoyons utiliser seulement la collecte par téléphone dans les États continentaux.

Pour la première vague de l'enquête, les Pays-Bas ont utilisé parallèlement un panel Internet et l'interview téléphonique avec CA, avec des tailles d'échantillon d'environ 400 et 1 800, respectivement. Cet exercice offrira la meilleure occasion jusqu'à présent de pouvoir tenir compte des effets de mode dans la modélisation. Ces effets ont fait l'objet de nombreux travaux de recherche récemment. Par exemple, certaines études ont montré que les personnes interviewées par téléphone choisissent les options extrêmes d'une échelle de Likert plus fréquemment que celles qui répondent par Internet (Wichers et Zenderink 2006 ; Bronner et Kuijlen 2007).

Aux Pays-Bas, l'échantillon de l'enquête en ligne est constitué de fumeurs sélectionnés aléatoirement à partir d'un grand panel polyvalent prérecruté d'environ 200 000 personnes créé par la firme TNS NIPO. L'échantillon de l'enquête téléphonique, sélectionné parmi les fumeurs abonnés à un service téléphonique par ligne terrestre, pourrait fort bien représenter une population de fumeurs différente. Le faible taux de réponse téléphonique montre clairement qu'aux Pays-Bas, les membres du public ne sont pas aussi réceptifs aux enquêtes téléphoniques que dans la plupart des autres pays participant au projet ITC. Nous avons souhaité que l'on demande à chaque groupe s'il pouvait être rejoint par l'autre mode, afin de pouvoir utiliser des méthodes à base de sondage double (Lohr et Rao 2000) pour calculer les poids d'enquête appropriés. Nous modéliserons également la propension (Rosenbaum et Rubin 1984) à répondre par téléphone (disons), sachant les variables démographiques et les variables d'accessibilité, et nous tiendrons compte du score de propension dans les comparaisons des profils de réponse selon le mode.

Les taux de réponse varient beaucoup, même dans le cas de l'enquête ITC-Quatre pays, les taux de réponse et les taux de rétention étant les plus élevés en Australie et les plus faibles aux États-Unis. Cela compromet certainement la capacité de faire des comparaisons entre pays, en ce sens que nous ne pouvons comparer que les populations représentées par les répondants, c'est-à-dire ceux qui, dans chaque pays, répondraient s'ils étaient sollicités aux termes de notre protocole. La situation paraît un peu meilleure si nous décomposons les taux de réponse en leurs composantes. Par exemple, nous avons constaté, d'après les résultats des tentatives d'appel et ce que nous savons de l'utilisation croissante d'appareils de filtrage des appels, qu'il est beaucoup plus difficile de prendre et de reprendre contact avec les adultes américains qu'avec ceux résidant dans les trois autres pays. Cependant, une fois que le contact est établi, le taux d'accord ou de non-refus aux États-Unis (plus de 80 %) est très semblable à ceux observés dans les trois autres pays.

Nous nous heurtons à des problèmes de mesure, même pour des faits ordinaires, tels que les habitudes d'achat et de consommation de tabac. Dans certains pays, comme l'Inde, le Bangladesh et le Soudan, dont l'inclusion dans le projet est en cours de discussion, de nombreuses formes de tabac sont utilisées couramment. Dans les pays développés, rien que tenir à jour la liste des marques de cigarettes est un travail à temps plein. Vient aggraver la difficulté le fait que, quand nous demandons aux participants à l'enquête quelles sont leurs habitudes d'achat ou s'ils ont remarqué les publicités, nous leur demandons de se souvenir de ce qu'ils ont fait au cours des deux semaines précédentes ou au cours d'une période plus longue. Dans la majorité des cas, nous nous fions à des données autodéclarées, mais pour un certain nombre de raisons, l'autodéclaration pourrait ne pas être exacte.

En ce qui concerne les attitudes et les croyances, nous savons depuis toujours que les questions doivent être adaptées à la langue et au niveau de littératie des participants, mais nous avons néanmoins été surpris et brusquement rappelés à la réalité de constater une forte prévalence de non-réponses partielles dans les régions éloignées d'un des pays, ce qui donne à penser que les questions sur les attitudes et les croyances présentent de grandes difficultés. Dans le cas de l'enquête pilote réalisée en Inde, l'interview a duré, en moyenne, 1,5 heure par participant, bien que le questionnaire ait été raccourci et simplifié.

Les mesures psychosociales doivent être validées dans chaque culture et chaque langue. Ainsi, nous avons commencé à inclure une très courte échelle de dépression. Voici la version utilisée pour l'enquête ITC-Quatre pays.

- Au cours du dernier mois, avez-vous souvent éprouvé peu d'intérêt ou de plaisir à faire les choses ?
- Au cours du dernier mois, avez-vous souvent éprouvé le sentiment d'être abattu(e), déprimé(e) ou désespéré(e) ?
- Au cours de la dernière année, un médecin ou un autre fournisseur de soins de santé vous a-t-il dit que vous faisiez une dépression ?

Et voici la version que nous avons finalement adoptée pour la deuxième vague de l'enquête ITC-Chine, en suivant les conseils d'autres chercheurs qui ont signalé qu'ils avaient pu en valider une version.

> Voici une liste d'états que vous pourriez avoir ressentis ou de comportements que vous pourriez avoir eus. Dites-moi à quelle fréquence vous vous êtes senti(e) de cette façon durant la dernière semaine.
>
> 1. Je n'avais pas envie de manger ; mon appétit n'était pas bon.
> 2. Je me sentais sans espoir au sujet de l'avenir.
> 3. Je me sentais triste.
> 4. J'avais l'impression que les gens ne m'aiment pas.

Ryder, Yang, Zhu, Yao, Yi, Heine et Bagby (2008) ont compilé les résultats d'une étude comparative fort intéressante de l'expression de la dépression.

La liste des problèmes de mesure continue. Même si la question est censée être la même dans les deux langues, il peut être difficile de trouver des équivalences. Nous nous efforçons d'obtenir une traduction de bonne qualité en utilisant un comité de la traduction ou en comparant des traductions indépendantes, mais nous devons souvent accepter un résultat imparfait. Par exemple, les traductions littérales de l'anglais au français ou à l'allemand sont généralement plus longues dans la langue d'arrivée et l'obtention d'une traduction facile à utiliser par téléphone requiert beaucoup de compétence. Thrasher, Quah, Borland, Awang, Sirirassamee, Boado, Miller, Watts et Dorantes (2008) décrivent une étude d'évaluation cognitive de certaines des questions les plus importantes qui a été réalisée dans plusieurs pays.

Il existe aussi des différences culturelles plus subtiles, particulièrement la mesure dans laquelle les répondants donnent une réponse socialement désirable. Nous avons relevé ce qui pourrait être une tendance plus prononcée à le faire chez les Mexicains et chez les Canadiens anglophones. Johnson et Van de Vijver (2003), entre autres, ont examiné la possibilité que les différences entre pays, pour ce qui est des réponses socialement désirables, soient reliées aux « systèmes de valeurs culturelles, comme dans la dimension de l'individualisme/collectivisme » de Hofstede (1980).

Dans une enquête longitudinale, nous devons aussi nous préoccuper de la validité et de la fiabilité des mesures répétées. Comme nous l'avons déjà indiqué, il est fréquent d'observer ce que l'on appelle des « effets de durée de la présence dans l'échantillon », en vertu desquels la proportion de réponses a tendance à évoluer à la hausse ou à la baisse à mesure que la cohorte progresse, simplement à cause du fait d'être mesurée.

Tous ces problèmes accroissent les défis analytiques auxquels font face les chercheurs. Fondamentalement, le but de l'analyse doit être de discerner la réponse réelle (ou le manque de réponse) au changement de politiques et de la séparer des effets du mode de collecte des données, de la non-réponse différentielle, des événements extérieurs, de la durée de la présence dans l'échantillon, de la culture et de la langue. Il s'agit là d'une tâche de taille.

3. Regroupement de données provenant de divers pays

Selon le paradigme classique de l'analyse des données d'enquête (Binder 1983 ; Godambe et Thompson 1986 ; Skinner 1989), nous avons un modèle pour les réponses y avec le paramètre θ, et nous imaginons comment nous estimerions θ si nous possédions les réponses pour l'ensemble de la population grâce à un recensement. Nous utiliserions une équation d'estimation sans biais efficace telle que :

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

pour définir une *estimation par recensement*. Pour obtenir l'estimation par sondage, nous utilisons une somme pondérée des termes de la fonction d'estimation par sondage :

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

pour obtenir un estimateur approximativement sans biais de la fonction d'estimation par recensement. Nous construisons les poids de sondage de manière à tenir compte du plan d'échantillonnage et de la sous-représentation de certains groupes à cause de la non-réponse et de la non-couverture. L'interprétation habituelle de w_i est le nombre de membres de la population représentée par i. L'utilisation de cette fonction d'estimation par sondage est séduisante, à cause de la réduction probable du biais dû à l'échantillonnage informatif et à la non-réponse ; mais, si les poids sont très variables et que le modèle pour les termes est correct, la deuxième équation donne un moyen inefficace d'estimer θ.

Or, quand nous combinons les données provenant de deux pays pour lesquels les fractions d'échantillonnage sont très différentes, comme dans le cas de l'enquête Irlande/R.-U., les poids pour un pays (R.-U.) seront beaucoup plus grands que ceux pour l'autre pays (Irlande). Si nous appliquons littéralement le paradigme, les données provenant du Royaume-Uni domineront l'analyse. Si le modèle est correct, l'estimation par recensement la plus efficace est la moyenne de y sur l'ensemble des deux pays. Mais alors, l'estimation par sondage correspondante représente une utilisation inefficace de l'échantillon. Ce problème est semblable à celui qui se pose dans les études cas-témoins, qui a été discuté par Scott (2006).

Un moyen de produire de meilleures estimations tout en retenant le paradigme classique consiste à considérer que la valeur du paramètre pour le Royaume-Uni est $\theta - \Delta$, que celle pour l'Irlande est $\theta + \Delta$, et que nous essayons d'estimer θ, la moyenne arithmétique des deux. Un système de fonctions d'estimation par recensement efficace pour θ et Δ équivaut à un système qui se divise en une partie pour chaque pays. Puisque le rééchelonnement des poids dans un pays n'a alors aucun effet sur les estimateurs ponctuels et sur leurs propriétés, la version par sondage pondérée de ce système produit une estimation efficace.

De surcroît, l'analyse qui s'ensuit est approximativement la même que celle que nous obtiendrions suivant le paradigme original si nous avions des tailles d'échantillon égales dans les deux pays et que nous rééchelonnions les poids de manière que leur somme soit égale à la taille de l'échantillon dans chaque pays. Comme l'a souligné Scott (2006), cette façon de rééchelonner les poids est très courante chez les épidémiologistes. Il s'agit, dans un certain sens, d'une application partielle de la méthode du poids q de Pfeffermann et Sverchkov (1999), où le poids égal à l'inverse de la probabilité d'inclusion est divisé par une sorte d'espérance du poids, conditionnellement à une variable explicative (pays).

Pour estimer un paramètre moyen θ, une proposition un peu plus attrayante est de considérer un modèle à effets aléatoires, où $Y = \theta + u + e$, et u est un effet aléatoire de pays, puis d'élaborer un système de fonctions d'estimation par recensement qui est efficace pour l'estimation du paramètre θ. Par exemple, si

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

et si les composantes de la variance correspondant à u et à e sont connues, alors la meilleure combinaison des deux moyennes de pays pour estimer θ est

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

où

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

Notons que, si $\sigma_u^2 = 0$, l'estimateur par recensement devient la moyenne de y sur les deux pays combinés. Cependant, si σ_u^2 est dominante, le meilleur estimateur est la moyenne arithmétique des moyennes de pays. Si l'on part d'un échantillon groupé, le paradigme habituel donne la même combinaison convexe d'estimateurs de la moyenne intra-pays fondée sur l'échantillon.

D'une manière plus générale, nous pouvons remplacer θ dans chaque fonction d'estimation par recensement de pays par $\theta + u$, où u désigne de nouveau un effet aléatoire de pays. Alors, la meilleure combinaison des deux fonctions d'estimation par recensement de pays pour θ est

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

où $c_1 = [\mathrm{Var}(\sum_{i=1}^{N_2} E(\phi_{2i}|u_2)) + E(\sum_{i=1}^{N_2} \mathrm{Var}(\phi_{2i}|u_2))][\sum_{i=1}^{N_1} E(\partial\phi_{1i}/\partial\theta)]$, et c_2 est défini symétriquement. Dans c_1, si le premier terme entre crochets domine, la fonction d'estimation par sondage correspondante pondère les termes de manière comparable dans les deux échantillons.

Même dans le cas simple d'une moyenne, les paramètres du modèle à effets aléatoires seront inconnus et difficiles à estimer si l'on n'a affaire qu'à deux pays, mais conceptuellement, le modèle semble être utile. Quand il existe plusieurs pays ou régions raisonnablement semblables (par exemple, les sept villes de l'enquête ITC-Chine), des modèles linéaires avec effets aléatoires peuvent être estimés suivant le paradigme habituel, comme il est décrit, par exemple, dans des conditions plus générales par Pfeffermann, Skinner, Holmes, Goldstein et Rasbash (1998).

À titre d'à-côté, l'analyse par EEG des données recueillies en Irlande décrite plus haut était une analyse groupée et tous ses « effets » ont été considérés comme étant fixes. Le modèle est presque « saturé », deux points dans le temps et deux pays représentant les quatre paramètres principaux. On peut voir qu'avec les poids de sondage ordinaires, l'estimation de β et de ρ serait dominée par les données du Royaume-Uni. Cependant, si β et ρ sont connus, comme dans le cas des paramètres θ et Δ dans l'exemple de la moyenne, les équations pour les paramètres principaux se séparent en deux paires, l'une pour α_0 et γ, et l'autre pour $\alpha_0, \alpha_1, \gamma$ et δ, chacune comprenant des poids ne provenant que d'un des deux pays. Donc, l'estimation des paramètres principaux est moins affectée par le rééchelonnement des poids. Si l'estimation de β nous importait aussi, nous pourrions considérer qu'il s'agit de la moyenne d'une variable aléatoire au niveau du pays, ce qui mène naturellement à l'obtention de l'influence appropriée pour chacun des deux échantillons. (En fait, dans notre analyse, nous n'avons pas choisi cette option ; nous avons rééchelonné les poids de manière que leur somme soit égale à la taille de l'échantillon dans chaque pays.)

La discussion qui précède de l'analyse de données groupées repose sur l'hypothèse qu'il existe un paramètre θ dont l'interprétation et la pertinence sont les mêmes dans tous les pays. La plupart des analyses multinationales partent de cette hypothèse. En effet, de Leeuw et Hox (2003) énoncent comme exigence pour une méta-analyse « que toutes les études doivent estimer le même paramètre fixe et qu'il est supposé que toute la variance est égale à la variance d'échantillonnage ». Mais en fait, une question essentielle est celle de savoir s'il est approprié ou non de construire un modèle qui doit s'appliquer simultanément aux données provenant de plusieurs pays. Parfois, il pourrait être plus indiqué de considérer simplement les modèles de pays comme étant distincts, mais parallèles. Par exemple, pour des pays à différents stades de développement, on peut s'attendre à ce que l'introduction de la même augmentation relative du prix réel des cigarettes entraînera une diminution de la consommation de cigarettes ; mais, puisque le modèle linéaire est, au mieux, une approximation locale utile de la relation complexe entre le prix et la consommation, il n'y a aucune raison de supposer que les diminutions seront du même ordre de grandeur ou que les deux estimations par régression mesureront la même quantité.

Voici un autre exemple. L'un des modèles d'intérêt dans le projet ITC est le modèle médiateur de la figure 4-1, dans lequel il est postulé de quelle façon « remarquer » les étiquettes de mise en garde concernant la santé pourrait influer sur l'intention d'arrêter de fumer.

À la période de référence, la distribution de l'intention d'arrêter varie assez bien selon le pays. Il en est de même des autres variables du modèle. Serait-il raisonnable d'espérer que les relations entre ces variables puissent devenir moins différentes entre les pays ? En fait, il semble que, pour les quatre pays originaux, il en soit ainsi. Même si les fumeurs du Royaume-Uni étaient beaucoup plus susceptibles de déclarer qu'ils n'avaient aucune intention d'arrêter, il n'en reste pas moins que la « préoccupation pour la santé » (prise de conscience des effets nocifs pour la santé déclenchée par les étiquettes) prédisait l'intention d'arrêter et que la prévalence de cette préoccupation augmentait avec le fait de remarquer les étiquettes ((Hammond, Fong, Borland, Cummings, McNeill et Driezen 2007). Donc, il n'est pas déraisonnable d'explorer un modèle tel que celui de la figure 4-1 pour les données en provenance des quatre pays, regroupées. Indépendamment des problèmes de pondération, dans la régression du médiateur « préoccupation concernant la santé » sur le « fait de remarquer » les étiquettes de mise en garde concernant la santé, de même que dans la régression de l'intention d'arrêter de fumer sur ces deux variables, nous avons trouvé qu'il était commode de considérer les moyennes de pays comme étant des effets fixes. Par ailleurs, puisque les coefficients de régression estimés pour les pays modélisés séparément varient modérément, il est naturel, dans l'analyse groupée, de conceptualiser ces coefficients comme possédant des composantes de pays aléatoires.

Cette discussion peut se résumer et s'expliquer comme il suit.

- Toute analyse qui consiste à regrouper des données provenant de plusieurs pays devrait être effectuée avec prudence. Pour qu'une telle analyse soit appropriée, la structure du modèle (l'équation de régression et ses variables) doit être correcte pour tous les pays et l'hypothèse que les paramètres sont communs doit être appuyée par la théorie et l'observation. Une estimation de variance robuste qui respecte les plans de sondage des pays sera nécessaire si ces plans sont complexes.
- Si l'ensemble de paramètres d'un modèle groupé peut (par transformation) être séparé en sous-ensembles disjoints correspondant aux pays, l'estimation de ces paramètres n'est pas affectée par de grands écarts entre les fractions d'échantillonnage appliquées dans les divers pays, ni par le rééchelonnement des poids dans les pays.
- Si une moyenne ou un paramètre de régression fixe est considéré comme étant commun aux pays, l'estimation en utilisant comme pondération l'inverse des probabilités d'inclusion sera inefficace si les fractions d'échantillonnage sont très variables.
- L'alternative au simple rééchelonnement des poids consiste à faire de la moyenne ou du paramètre de régression un effet fixe qui varie selon le pays (ce qui aboutit à la séparation en sous-ensembles disjoints, mais augmente le nombre de paramètres et élimine le « caractère commun ») ou à faire de la moyenne ou du paramètre de régression un effet aléatoire, variable selon le pays (ce qui aboutit à une séparation approximative et retient le « caractère commun »).
- Du point de vue conceptuel, il est séduisant de rendre l'ordonnée à l'origine fixe et la pente, aléatoire, puisque le niveau de référence a tendance à varier beaucoup plus selon le pays que la pente. À l'étape de la mise en œuvre, cette approche requiert un nombre suffisant de pays pour que l'estimation des composantes de la variance soit faisable et un petit nombre d'effets aléatoires à intégrer.
- Quand une analyse groupée pose des problèmes, une comparaison moins formelle des résultats des analyses exécutées parallèlement dans les divers pays permet parfois d'accomplir la plupart de ce qu'il est souhaité.

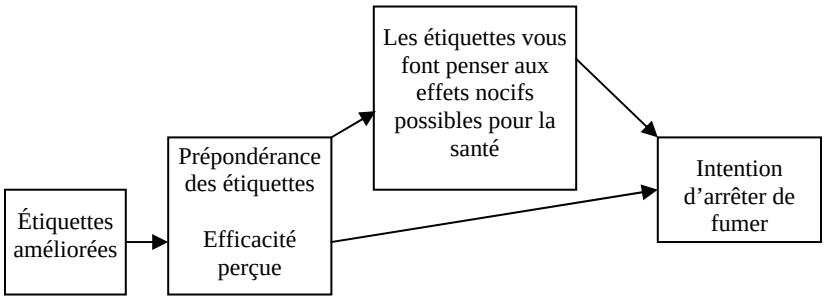
4. Étalonnage des mesures et comparaisons transculturelles

L'autre problème statistique que je souhaite mettre en relief est l'utilisation de modèles de mesure pour essayer d'étalonner les mesures obtenues par divers modes de collecte et de comparer les mesures provenant de diverses cultures. Une approche courante consiste à considérer qu'à l'aide de chaque item du questionnaire, on mesure un concept, comme la « dénormalisation sociale » (perception d'une désapprobation sociétale) et à traiter le concept comme une variable continue η. La distribution de η, sachant les variables explicatives, détermine une distribution des réponses aux items du questionnaire.

Si nous disposons de plusieurs items du même type pour mesurer un concept, un modèle conceptuel pour les mesures continues y pourrait être $Y_{ik} = b_i\eta_k + a_i + e_{ik}$ pour l'item i et le participant k. Ici b_i représente un facteur d'échelle

positif pour l'item i, a_i, un changement d'emplacement et e_{ik}, une erreur de mesure normale de moyenne nulle et de variance σ_{ei}^2 indépendante de k. Supposons que toutes les e_{ik} sont indépendantes les unes des autres et indépendantes des η_k. (Cela revient en fait à émettre l'hypothèse que η est le seul déterminant latent de Y.) Si nous posons que la distribution de η est $N(0, 1)$, ce que nous pouvons faire si η est normal sans variables explicatives, alors la distribution de Y_{ik} est $N(a_i, b_i^2 + \sigma_{ei}^2)$, et, pour un item i unique, les paramètres a_i et $b_i^2 + \sigma_{ei}^2$ peuvent être estimés d'après les données de marge pour un grand nombre de participants. S'il existe au moins deux items ayant la même variance, puisque les covariances des réponses aux items d'un participant sont de la forme $b_{i_1} b_{i_2}$, tous les paramètres peuvent être estimés d'après les données de marge sur de nombreux participants. Sachant les valeurs des paramètres d'item, la valeur de η pour un participant peut être « prédite » d'après la distribution a posteriori de η, sachant les réponses aux items du participant.

Figure 4-1
Modèle médiateur des effets des politiques : étiquettes de mise en garde



Si la mesure y est binaire, il est fréquent de choisir un modèle fondé sur la théorie de la réponse à l'item (TRI) $\text{Prob}(Y_{ik} = 1|\eta_k) = H(b_i\eta_k - \gamma_i)$, où H est la loi normale standard ou la fonction de répartition (c.d.f.) logistique. Le paramètre b_i est le « paramètre de discrimination » pour l'item et γ_i est un seuil tel que la probabilité de réponse 1 soit supérieure à 1/2 quand le concept rééchelonné par b_i excède γ_i. La probabilité inconditionnelle que $Y_{ik} = 1$ s'obtient par intégration par rapport à la distribution de η_k, étant donné des variables explicatives fixes pour le participant k. Dans le cas le plus simple, il semble qu'au moins trois items sont nécessaires (avec de nombreux participants) pour que tous les paramètres puissent être estimés, puisqu'ils produiraient sept probabilités conjointes pour l'estimation de six paramètres. De nouveau, sachant les valeurs des paramètres d'item, la valeur de η pour un participant peut être prédite, sachant son ensemble de réponses aux items. Voir, par exemple, Lu, Thomas et Zumbo (2005). Un logiciel standard d'estimation des variables latentes peut être utilisé pour produire ces inférences, ainsi que leurs analogues dans le cas de mesures ordinales.
Commençons par examiner le problème d'étalonnage. Supposons qu'il existe deux modes de collecte des données et que, pour l'item i dans le mode j avec le participant k, nous avons la mesure continue.

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

Ce modèle, dans lequel α_j et β_j ne dépendent pas de l'item i, pourrait convenir pour un ensemble d'items qui sont tous du même type général. Les exemples plausibles ne sont pas nombreux, mais l'un d'eux pourrait être une série de questions de la forme : « Pendant quel pourcentage du temps diriez-vous que vous vous sentez… », où il est demandé au répondant de donner un pourcentage par téléphone, ou de marquer une position sur une ligne dans le cas d'un questionnaire papier.

Si nous posons que les a_i et b_i sont les paramètres des items lorsqu'on utilise le premier mode de collecte des données, nous pouvons fixer $\alpha_1 = 0$ et $\beta_1 = 1$. Si β_2 est plus grand que 1, les réponses ont tendance à être plus variables ou plus extrêmes sous le deuxième mode de collecte. Si α_2 est plus grand que 0, les répondants ont tendance à donner une réponse plus élevée sous le deuxième mode de collecte que sous le premier. Notons que les échantillons pour les deux

modes de collecte sont constitués de participants différents. Si nous pouvons émettre l'hypothèse que la distribution de η est la même pour les deux échantillons (une hypothèque qui, en fait, requiert une randomisation par rapport au mode), nous obtenons la distribution de Y_{i1k} comme auparavant, $N(a_i,\ b_i^2+\sigma_{ei}^2+\sigma_\varepsilon^2),$ tandis que la distribution de Y_{i2k} est $N(\beta_2 a_i + \alpha_2,\ \beta_2^2 b_i^2 + \beta_2^2 \sigma_{ei}^2 + \sigma_\varepsilon^2).$

Si $\sigma_\varepsilon^2 = 0,$ sachant les données sur un item i dans les deux modes, nous pouvons estimer α_2 et $\beta_2,$ en supposant que β_2 est positif. Si $\sigma_\varepsilon^2 > 0,$ les paramètres α_2, β_2 et σ_ε^2 peuvent être estimés, à condition que l'on dispose d'au moins deux items – du même type, mais dont les valeurs de a et b diffèrent.

Ces considérations peuvent être étendues au cas plus habituel des items avec réponses ordinales en imaginant une probabilité de réponse ordinale devant être déterminée par une réponse continue sous-jacente. Pour des données binaires, nous poserions le plus simplement

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

avec $\alpha_1 = 0$ et $\beta_1 = 1.$ Si la distribution de η_k est la même pour les deux modes, alors, avec des données provenant de nombreux participants et trois items, nous pouvons déterminer tous les paramètres. L'ajout d'une variable explicative réduirait le nombre d'items requis.

L'hypothèse que la distribution de η est la même pour les échantillons correspondants aux deux modes est cruciale pour ce type d'étalonnage, et il est difficile de garantir qu'elle sera vérifiée. Elle l'est si nous avons des échantillons probabilistes superposés pour les deux modes dans une seule enquête ; alors, en principe, nous pouvons imaginer une mise en correspondance des réponses d'un mode à l'autre, par la voie des valeurs estimées de α_2 et β_2. Nous ne devons pas estimer les concepts proprement dits pour cela. Plus rigoureusement, nous pouvons inclure α_2 et β_2 comme paramètres dans un modèle pour toutes les réponses à un ensemble d'items similaires.

Dans certains pays développés, les bases de sondage pour les ménages et les particuliers semblent évoluer vers des registres d'adresses et des listes de personnes. Cependant, même s'il existe une base de sondage commune pour (disons) les enquêtes par téléphone et en ligne, il est difficile de randomiser les répondants par rapport aux modes de collecte des données. La façon dont la non-réponse dépend des variables démographiques pourrait fort bien différer selon le mode. En outre, la nécessité de maximiser les taux de réponse oblige souvent à permettre aux répondants de choisir. En principe, nous pourrions imaginer que la distribution de η peut être déplacée ou inclinée en fonction de la « propension » à choisir un mode ou l'autre. Après avoir modélisé cette propension en fonction des variables explicatives et introduit un ou deux paramètres pour la dépendance de la distribution de η à l'égard de la propension, nous pourrions estimer les paramètres d'item a_i et b_i d'après les réponses obtenues par le premier mode de collecte des données. L'estimation des paramètres de mise en correspondance α et β s'ensuivrait de la même manière qu'auparavant.

Dans d'autres circonstances, nous pourrions utiliser les deux modes de collecte des données dans différents groupes de population. Dans ce cas, l'effet de mode fait partie de l'effet de groupe ; il ne peut pas être distingué d'une différence sous-jacente dans la distribution du concept.

Le problème de la comparaison des mesures entre différentes cultures ou entre d'autres groupes n'est pas le même que celui de l'étalonnage, puisque la randomisation des participants par rapport aux groupes, afin que la distribution du concept demeure constante, est hors de question. On pense généralement que, pour comparer la moyenne d'un concept d'un groupe à l'autre, les items mesurés doivent avoir la même relation avec le concept dans les deux groupes. Quand il existe plusieurs concepts, la comparaison de la relation entre les concepts d'un groupe à l'autre requiert une sorte d'« invariance de mesure » ou d'équivalence pour tous les items concernés. La littérature sur les comparaisons et les mesures transculturelles est abondante. Par exemple, Johnson (1998) énumère, pour l'équivalence transculturelle, 52 termes qui ont été introduits par les auteurs dans diverses disciplines.

Le modèle d'analyse factorielle confirmatoire multigroupes est utile pour les items dont la mesure est continue et prend la forme :

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

où Y_k^g est le vecteur des réponses observées aux items pour le répondant k dans le groupe g, Λ^g est une matrice des pentes ou des « saturations factorielles », le vecteur d'ordonnées à l'origine τ^g indique la valeur attendue de Y_k^g quand $\eta_k^g = 0$ et e_k^g est une erreur de mesure de moyenne 0. Alors, $E(Y_k^g) = \tau^g + \Lambda^g \kappa^g$, où κ^g est la moyenne du concept η dans le groupe g. La matrice de variance-covariance entre les valeurs observées y_k^g peut être exprimée sous la forme $V(Y_k^g) = \Lambda^g \Phi^g \Lambda^{g'} + \Theta^g$, où Φ^g est la matrice de covariance des concepts latents et Θ^g est la matrice diagonale des variances de l'erreur de mesure. À cet égard, voir de Jong, Steenkamp et Fox (2007), Davidov (2008) et les références qu'ils fournissent.

La version TRI du modèle peut être définie simplement. En utilisant la même notation pour les paramètres, dans le cas des items binaires, nous avons

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

et il existe une extension naturelle au cas ordinal.

Les paramètres du modèle ne peuvent pas être identifiés, à moins d'imposer certaines contraintes. Dans le modèle d'analyse factorielle confirmatoire multigroupes, de nombreux auteurs postulent un item « marqueur » pour chaque concept, avec une saturation factorielle de 1 et une ordonnée à l'origine de 0 pour tous les groupes, de sorte que la moyenne du concept est déterminée dans chaque groupe. Cette hypothèse est très forte. Alternativement, nous pourrions choisir les unités pour les concepts de sorte qu'elles soient marginalement N(0, 1) dans le groupe 1. Les paramètres des items (avec un nombre suffisamment grand d'items) sont donc identifiés pour le groupe 1. Si nous supposons que les variances et les relations du diagramme des trajectoires restent vraies dans le groupe 2, alors nous pouvons vérifier si les paramètres d'item restent également les mêmes et, dans la négative, essayer de reconcevoir l'ensemble d'items afin d'en produire un s'approchant davantage de l'invariance de mesure. Par ailleurs, si les paramètres d'item sont contraints de demeurer les mêmes, nous pouvons vérifier si la distribution conjointe sous-jacente des concepts est également la même. Cependant, le rejet formel de l'hypothèse nulle est difficile à interpréter. S'inspirant de Rensvold et Cheung (1998), Barrera Ceballos (2007) a procédé à ce genre d'analyse multigroupes pour les données de l'enquête ITC-Mexique et de l'enquête ITC-Uruguay, en remplaçant « préoccupation concernant la santé » dans le modèle de la figure 4-1 par « dénormalisation sociale », c'est-à-dire la mesure dans laquelle le répondant a l'impression que la société désapprouve le tabagisme. (Les deux autres concepts sont la prépondérance des étiquettes de mise en garde et l'intention d'arrêter.) Les relations semblent étonnamment différentes dans les deux pays sous les contraintes d'invariance des items de mesure, résultat qui pourrait être dû à des différences sociétale réelles ou à une correspondance imparfaite entre les items proprement dits (c'est-à-dire échec des contraintes). Il faut reconnaître que, ne contenant qu'un très petit nombre de concepts possédant des items multiples, le questionnaire de l'enquête ITC n'était pas conçu pour ce genre d'analyse.

En fin de compte, les relations entre les concepts sont de la plus haute importance, de même que la question de savoir si ces relations peuvent être considérées comme étant semblables, quoique pas forcément identiques, d'un groupe à l'autre. Il en est ainsi, que les distributions marginales des concepts soient les mêmes ou que les items mesurés possèdent les mêmes paramètres d'un lieu à l'autre ou d'un mode à l'autre. Intuitivement, les deux catégories de contraintes mentionnées au paragraphe précédent semblent trop fortes. Une approche hiérarchique proposée par De Jong et coll. (2007) offre une solution.

Si l'item i comprend C options de réponse ordonnées, nous pouvons écrire

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

$c = 1, \ldots, C$. Ici, les saturations factorielles sont remplacées par les paramètres de discrimination b, et les ordonnées à l'origine sont remplacées par les seuils γ. Au lieu d'insister sur le fait que ces paramètres sont indépendants de l'étiquette de groupe avant le traitement, l'approche consiste à les modéliser avec des effets aléatoires propres au groupe :

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

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

L'hétérogénéité de la variable latente est modélisée par une structure hiérarchique :

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

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

Si le nombre d'items est suffisant, un tel modèle peut être estimé et ajusté en utilisant des méthodes Monte Carlo à chaîne de Markov. Les tests d'invariance de l'analyse multigroupes peuvent encore être exécutés dans ce cadre de travail.

5. Discussion et conclusion

De nouveau, dans le contexte du projet ITC, le but de l'analyse doit être de discerner la réponse réelle (ou le manque de réponse) au changement de politiques et de la séparer des effets du mode de collecte des données, de la non-réponse différentielle, des événements extérieurs, de la durée de la présence dans l'échantillon, de la culture et de la langue. Il n'est pas toujours nécessaire de faire la distinction entre tous les facteurs confusionnels, mais il est important de leur permettre de contribuer au modèle. Dans le présent article, nous n'avons pas abordé la question des événements externes, qui peuvent être modélisés de manière évidente s'ils sont reconnus. Nous n'avons pas discuté en détail de la modélisation des effets d'attrition et de durée de la présence dans l'échantillon, mais en principe, chacun de ces effets peut être considéré comme faisant partie de l'ensemble. Ceux qui sont retenus d'une vague à l'autre d'une enquête pourraient être considérés comme une sorte de groupe culturel. Par ailleurs, les effets de durée de la présence dans l'échantillon constituent une forme particulière d'échec de l'invariance de la mesure, au cours du temps, plutôt que d'un groupe à l'autre. Une analyse complète tiendrait compte de ces effets, ainsi que d'autres effets de la culture, de la langue et du mode de collecte de données.

Il n'est absolument pas certain que les effets des politiques puissent systématiquement être identifiés dans un modèle complet. Mais les chances augmentent si le plan de sondage tient compte des comparaisons de données longitudinales entre pays et de la répétition qui provient de l'observation de cohortes qui ont des points de départ différents.

Un thème commun aux deux sections précédentes est l'introduction d'effets aléatoires en tant qu'outil. Le procédé consistant à introduire des effets aléatoires pour les pays et les groupes dans les paramètres clés est naturel et (pour les échantillons de grands groupes) conceptuellement compatible avec l'analyse classique des données d'enquête fondées sur les fonctions d'estimation pondérées. Il existe certains obstacles à l'application pratique, dus aux limites d'identifiabilité et d'estimabilité, ainsi qu'au calcul des fonctions de vraisemblance si l'on envisage plus qu'un ou deux effets aléatoires. Parallèlement, étant donné la disponibilité croissante de méthodes numériques pour traiter ce genre de modèle, la poursuite des travaux de recherche en vue de les adapter à des enquêtes internationales complexes devrait être très fructueuse.

Remerciement

Ces travaux sont financés partiellement par une subvention du Conseil de recherche en sciences naturelles et en génie du Canada. Le projet ITC est financé par des subventions du National Cancer Institute des États-Unis (P50 CA11236), du Roswell Park Transdisciplinary Tobacco Use Research Center et des Instituts de recherche en santé du Canada (57897). L'auteure remercie un examinateur anonyme de ses commentaires très constructifs.

Bibliographie

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. *Revue Internationale de Statistique*, 51, 279-292.

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

Cook, T., et 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. et 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., et Hox, J. (2003). The use of meta-analysis in cross-national studies. Dans *Cross-Cultural Survey Methods* (Éds., J.A. Harkness, F.J.R. Van de Vijver et 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. et 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. et 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., et Thompson, M.E. (1986). Parameters of superpopulation and survey populations: Their relationships and estimation. *Revue Internationale de Statistique*, 54, 127-138.

Hammond, D., Fong, G.T., Borland, R., Cummings, K.M., McNeill, A. et 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. et Hastings, G. (2008). The impact of smokefree legislation in Scotland: Results from the Scottish International Tobacco Control Policy Evaluation Project. Travaux en cours.

ITC (2008). Tobacco Labelling Resource Centre. http://www.igloo. org/tobacco_labelling. Accessible le 24 avril 2008.

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

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

Lohr, S.L., et 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. et 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. et Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Séries B, 60, 23-40.

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

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

Rosenbaum, P.R., et 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. et 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.

Scott, A. (2006). Études cas-témoins basées sur la population. *Techniques d'enquête*, 32, 137-147.

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

Thompson, M.E., Boudreau, C. et Driezen, P. (2005). Incorporating time-in-sample in longitudinal survey models. *Recueil* : *Symposium 2005*, *Défis méthodologiques reliés aux besoins futures d'information.* Session 12 : Défis pour l'emploi des données provenant d'enquêtes longitudinales. Statistique 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. et 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. et Dorantes, A. (2008). Ensuring valid cross-cultural comparisons in survey research on tobacco: Development, implementation, and results from a transnational cognitive interviewing study. Travaux en cours.

Verma, V., Scott, C. et O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, Séries 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 Accessible le 14 avril 2008.

Wichers, B., et 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.