AGREGATION DE DONNEES MULTIMODES : UN ETAT DES LIEUX ET UNE PROPOSITION PRAGMATIQUE POUR CONTROLER L'EFFET DE MESURE

Stéphane Legleye[1]

[1] *Insee, 88 Avenue Verdier, 92120 Montrouge, France ; CESP, Inserm, Hôpital Paul Brousse, 12 Avenue Paul Vaillant Couturier, 94800 Villejuif*

Résumé.

Les enquêtes multimodes ont ceci d'original qu'elles imposent de reconsidérer quelques aspects fondamentaux de méthodologie d'enquête qui sont souvent minimisés voire ignorés dans les enquêtes monomodes. Dans cette présentation, je rappellerai quelques-unes de ces évidences, relatives au mode de contact, au mode de collecte, à leur impact sur la représentativité des échantillons de répondants ainsi qu'à leur impact sur la qualité intrinsèque du recueil des données. Si l'effet mode de collecte est généralement perçu comme une nuisance, il peut être au contraire bénéfique et désiré et contribuer à dicter le choix du protocole multimode. Les situations doivent s'apprécier au cas par cas, suivant l'objectif des enquêtes, leur place dans les dispositifs d'observations, leurs utilisations publiques et pour la recherche, et enfin par l'existence de séries de mesures dans le temps.

Après avoir défini l'effet de sélection et l'effet de mesure, et rappelé les principales techniques permettant de les séparer, je présenterai quelques approches définies récemment pour tenter de le contenir voire de neutraliser l'effet de mesure. Enfin, je présenterai une approche pragmatique et parcimonieuse développée à l'Insee, fondée directement sur le concept d'effet de mesure, visant à le réduire au maximum lorsque cela est nécessaire.

Mots-clés. Multimodes, estimation, correction

Abstract.

The originality of mixed-mode surveys is that they require a reconsideration of some fundamental aspects of survey methodology that are often minimized or even ignored in single-mode surveys. In this presentation, I will recall some of these evidences, relating to the mode of contact, the mode of collection, their impact on the representativeness of the respondents' samples as well as their impact on the intrinsic quality of the data collection. While the collection mode effect is generally perceived as a nuisance, it can be beneficial and desired and help dictate the choice of a mixed-mode protocol. Situations must be assessed on a case-by-case basis, according to the purpose of the surveys, their place in observation systems, their public and research uses, and finally by the existence of series of measurements over time. After having defined the selection effect and the measurement effect, and recalled the main techniques to separate them, I will present some approaches recently defined to try to contain or even neutralize the measurement effect. Finally, I will present a pragmatic and parsimonious approach developed at INSEE, based directly on the concept of measurement effect, aiming to reduce it to the maximum when necessary.

Keywords. Mixed-modes, estimation, correction

1. Introduction

Les enquêtes multimodes sont à la mode. La raison n'est pas la recherche d'originalité mais principalement la volonté de s'adapter aux contraintes s'imposant aux concepteurs d'enquêtes depuis quelques années. La première est la baisse du taux de participation, très largement constatée dans de nombreux pays et pour presque tous les modes de collecte traditionnels (Czajka and Beyler 2016, Jonas 2016, Pew Research Center 2017). Certes, calculer un taux de participation de façon simple, transparente et transposable universellement est une tâche ardue sinon impossible, malgré les recommandations et autres grilles de calcul diffusées par certains organismes comme l'AAPOR (American association for public opinion research) (Skalland 2011). De plus, son lien avec la qualité de la mesure est généralement discutable, ce qui en fait un indicateur assez médiocre de la qualité global de la collecte (Groves 2006, Groves and Peytcheva 2008, Davern, McAlpine et al. 2010, Davern 2013). Dans certaines populations, par exemple les médecins (Gautier 2011, Legleye, Bohet et al. 2014, Legleye, Pennec et al. 2016), obtenir un taux de participation supérieur à 30% est une gageure : pourtant, les résultats ne sont que rarement mis en doute car l'on considère (implicitement) que la non-réponse est indépendante des comportements que l'on cherche à mesurer. Les données du Pew Research Center montrent qu'aux Etats-Unis, le taux de réponse moyen dans les enquêtes téléphoniques aléatoire avoisine les 9% en 2016, alors que nombre d'indicateurs semblent toujours correctement estimés (Pew Research Center 2017). Malgré tout, afficher un bon taux de réponse reste un objectif désirable et valorisé, et fournir un ou plusieurs modes de collecte adaptés peut y contribuer. Proposer un mode de collecte adapté aux souhaits de l'enquêté est donc souhaitable (Olson, Smyth et al. 2012, Smyth, Olson et al. 2014).

La deuxième raison du recours à un protocole multimode est d'améliorer la qualité du recueil des données : s'il est utile de recourir à un enquêteur pour persuader l'individu sélectionné de répondre à l'enquête, il peut être préférable d'utiliser un mode auto-administré et donc d'écarter l'enquêteur pour recueillir certaines informations, notamment des informations personnelles et sensibles.

La dernière motivation est bien sûr la contrainte budgétaire : les enquêteurs coûtent cher. Alors que l'immense majorité des sondages d'opinion se fait sur Internet, les instituts de statistique et les instituts de recherche sont soumis à de fortes pressions pour basculer tout ou partie de leurs enquêtes ou questionnaires sur Internet.

Les deux premières raisons de recourir au multimode trouvent place dans le cadre général de l'erreur totale d'enquête (Total survey error). Le TSE définit la liste de toutes les sources d'erreurs du processus d'enquête (Biemer 2010, Groves and Lyberg 2010) que les protocoles multimodes invitent à reconsidérer avec un regard attentif et renouvelé. Les principales sont le biais de non-réponse et l'erreur de mesure. L'erreur de mesure se présente souvent à l'esprit dès qu'il est question de multimode et de très nombreuses méthodes ont été développées pour tenter de l'estimer, et de l'éliminer. Toutefois, nous verrons que décider de le corriger est difficile.

Dans cette présentation, je rappellerai quelques évidences au sujet des enquêtes et de leur caractère intrinsèquement multimode en un sens large. Puis je passerai en revue les biais susceptibles d'être introduits par le recours à différents modes de collecte pour la phase de questionnaire, en distinguant notamment les questionnaires auto-administrés et les questionnaires intermédiés. Je présenterai brièvement les différentes méthodes employées pour estimer le biais de mesure et enfin celles pour le corriger, en exposant leurs limites. Parmi elles, je présenterai et discuterai une méthode originale mais simple développée à l'Insee. Je conclurai par quelques éléments généraux de bon sens.

2. Multimodes

2.1. Toutes les enquêtes sont multimodes

En un sens général, toutes les enquêtes ou presque sont multimodes, dès lors que leur protocole comprend une phase de contact (ou invitation) distincte de la phase de collecte (de Leeuw 2005). L'invitation à participer à une enquête se fait par le biais d'une lettre-annonce, plus rarement d'un avis de passage, d'un SMS ou d'un message téléphonique. Cette annonce indique au ménage ou à un individu qu'il est sélectionné pour répondre. Ce mode d'avertissement diffère souvent complètement, en termes de support, de l'administration du questionnaire. Par ailleurs, pour être efficace, l'annonce, quel que soit son medium, doit être perçue et comprise. Le support (format, maquette et présentation, matériau etc.) compte donc, pour chacun des envois. Et l'envoi ne garantit à coup sûr ni la réception, ni la lecture, ni la compréhension ni l'acceptation de l'enquête. Cette étape préliminaire d'invitation est généralement considérée comme favorisant beaucoup la réponse aux enquêtes téléphoniques (de Leeuw, callegaro et al. 2007, von der Lippe, Schmich et al. 2011) ou bien face à face (bien que la littérature soit quasiment muette sur ce dernier point).

La nature du support et le type de contenu sont déterminants pour le bon accueil de l'enquêteur, en face-à-face, au téléphone ou pour prendre la décision de se rendre sur l'adresse du site internet du questionnaire en ligne. En fait, toute information perçue sera utilisée par le destinataire pour motiver sa décision (rationnelle ou non…) de participer, résultat d'un arbitrage entre d'un côté son intérêt éventuel pour la thématique, son sens civique, sa perception du sérieux de la démarche et la notoriété du commanditaire d'un côté, et le fardeau de réponse de l'autre (Groves, Singer et al. 2000). Dans cette optique, le choix du support de communication peut sans doute jouer sur la réception de l'information, avec des effets variables suivant les publics (jeunes, vieux, diplômés etc).

2.2. La phase de contact comme mode de collecte

Si l'on exclut la phase de notification via la lettre-annonce ou ses avatars, avant la phase de collecte proprement dite se joue la phase de contact, qui est déterminante dans les enquêtes intermédiées. L'enquêteur, au téléphone ou en face-à-face tente de convaincre la personne sélectionnée de lui accorder un peu de temps, de l'écouter et finalement de participer à l'enquête. Il doit convaincre et donc séduire. Les outils et les moyens à sa disposition varient d'un mode à l'autre.

En cas de présence physique, l'enquêteur se caractérise par son sexe, son âge apparent, son habillement, sa posture, sa gestuelle, son sourire, sa voix (intonation, débit, accent) etc. Mais il dispose aussi d'attributs authentifiant sa fonction : carte professionnelle, mallette, questionnaires et ordinateurs à en-tête de la société, brochures de résultats etc. Il peut déployer une palette de stratégies pour convaincre et séduire (Joule and Beauvois 2014) et ce d'autant plus qu'il échappe au contrôle étroit de sa supervision : il est autonome et s'adapte à son public, quitte à enfreindre le protocole et la déontologie sur certains points de façon à être le plus efficace possible (Caveng 2012). Par ailleurs, bienséance et politesse imposent une relative bienveillance à l'égard de personnes se présentant à sa porte, ce qui impose généralement de leur accorder un minimum d'attention. De plus, le démarchage ne porte à porte est devenu rare, ce qui limite les confusions possibles avec une démarche commerciale et impose de consacrer un minimum de considération à l'enquêteur.

Au téléphone, l'enquêteur ne dispose que de sa voix, qui est son seul moyen de convaincre alors qu'il fait face à la concurrence nuisible du démarchage commercial téléphonique, très répandu et perçu comme indésirable. Presque tous les appels non sollicités provenant de numéros inconnus ou masqués sont des démarches commerciales, et les vendeurs imitent la démarche d'enquête véritable durant leur prise de contact. Au final, tout concourt donc à ne jamais répondre ou à refuser immédiatement toute sollicitation provenant d'un numéro inconnu ou masqué. C'est d'autant plus facile que l'appelant est distant et anonyme. La lettre annonce éventuelle est souvent loin du téléphone, même si le numéro d'appel y est mentionné, et elle a été oubliée depuis longtemps.

A l'autre bout du spectre, les contacts par courrier peuvent également être ignorés facilement, tout comme les invitations par courriel. Un filtrage systématique peut de plus être effectué. Et il est plus facile d'ignorer ou de feindre d'ignorer ces types de sollicitation, ou d'oublier de le faire que lorsque l'enquêteur est impliqué, physiquement ou au téléphone.

2.3. Le mode de collecte en général

Au sens large, un mode de collecte est la conjonction d'un support et véhicule d'information et d'une situation d'interlocution (ou non) qui s'étend du contact au remplissage du questionnaire (collecte des données) proprement dit. Relativement au remplissage du questionnaire, le mode de collecte peut se définir par l'implication d'un enquêteur, son degré de présence et les canaux de communication (Couper 2011). L'enquêteur peut seul avoir accès au questionnaire et lire les questions ou bien il peut uniquement distribuer ou présenter le questionnaire. Il peut être physiquement présent, (verbalement) présent au téléphone, présent en vidéo ou représenté par un avatar sur un ordinateur, etc. Le degré de proximité avec l'enquêté est donc également très variable : fort en présentiel, modéré au téléphone, distant en vidéo, symbolique en avatar, très faible en auto-administré, même si le questionnaire a été présenté par un enquêteur (en présentiel, au téléphone, en vidéo etc.) Le sentiment de proximité, la syntonie, l'empathie et la perception des enjeux varient. Ainsi que la confiance : les attributs de l'enquêteur légitime sont visibles en face-à-face, inexistants ou presque en avatar.

Le mode de collecte utilisé pour remplir le questionnaire utilise par ailleurs des canaux de communication différents : visuelle, orale. Avec un enquêteur en face-à-face, tous les canaux et tous les types de communication sont mobilisés, visuelle (présentation de l'enquêteur mais aussi cartes et supports s'il y en a), sonore et linguistique (voix), sémantique (contenu des questions), etc. Avec le téléphone, seul le canal sonore est utilisé. Avec le papier-crayon, seule de l'information visuelle du questionnaire est présente (mise en page, couleurs, police, etc.) en plus du contenu sémantique (questions). Sur l'ordinateur en auto-administré, que ce soit sur Internet ou via une application de type formulaire, on est dans un cas similaire, mais des informations additionnelles peuvent apparaître, conditionnellement ou non (demande d'information, consignes, modalités de réponse supplémentaires comme « ne sais pas » etc.) ; des contenus graphiques, vidéos et sonores peuvent aussi être inclus.

Une partie de l'information véhiculée, consciemment ou non, n'est donc pas de nature linguistique : la tenue de l'enquêteur, son ton, sa posture relativement aux questions et aux réponses, mais aussi la mise en page du questionnaire, les couleurs etc. sont des éléments qui influent sur l'interprétation des contenus, ou l'intérêt que l'enquêté portera aux questions.

Par ailleurs, l'enquête est une situation d'interlocution dans laquelle se distribuent les questions et réponses, l'initiative du rythme, des prises de paroles, du respect des modalités de réponse, du vocabulaire employé etc. C'est la question du locus de contrôle, de la maîtrise : elle est totale ou très large en auto-administrée (l'enquêté peut décider du moment qui lui convient pour répondre, il peut s'interrompre et reprendre etc.), plus réduite en situation intermédiée.

Le lieu et le moment, la présence d'un tiers, le sentiment de confidentialité peuvent également jouer sur la collecte. L'interview dans la rue ne peut sans doute être comparée à l'interview chez soi, dans un laboratoire, un café etc.

Lorsque l'on étudie l'effet de mode, on fait généralement référence au mode d'administration (remplissage) du questionnaire, mais en toute rigueur, toute la série des modes utilisés durant les phases successives d'annonce, de contact et de remplissage devrait être prise en compte, même si l'effet prépondérant est certainement celui occasionné par le mode d'administration du questionnaire.

2.4. Quelques exemples de protocoles

Les combinaisons entre les différents modes de contact et de collecte sont pléthoriques. Je citerai quelques exemples typiques.

- Enquête transversale avec notification par lettre annonce et questionnaire en face-à-face avec ordinateur (Computer assisted personal interview) : une large part des enquêtes de l'Insee.
- Enquête téléphonique avec génération aléatoire de numéros assortie d'un envoi d'une lettre annonce lorsque c'est possible (moins de 40% des numéros filaires) et envoi de SMS aux mobiles, ce qui fait que tous les enquêtés n'auront pas eu la même notification : le Baromètre santé de Santé publique France.
- Enquête transversale face-à-face avec notification par lettre, CAPI, avec sous-questionnaire passé sur un mode alternatif auto-administré sous casque (audio-CASI) : la première enquête de ce type au sein de la statistique publique fut EVS (événements de vie et santé) en 2005 (Cavalin 2009) ; c'est également le cas de CVS.
- Enquête aléatoire téléphonique avec bascule d'une partie du questionnaire sur un automate avec reconnaissance vocale (interactive voice recognition) : Enquête National STD and Behavior Measurement Experiment (NSBME) portant sur les comportements sexuels.
- Enquête transversale aléatoire téléphonique (avec envoi de lettre) avec bascule (volontaire ou contrainte) sur Internet pour le questionnaire après sélection au téléphone (Enquête Fecond mode de collecte de l'Ined) : protocole concurrentiel et séquentiel après sélection.
- Enquête aléatoire téléphonique (avec envoi de lettre-annonce et SMS lorsque c'est possible) avec bascule sur Internet pour une sous-partie de la population (Enquête Virage de l'Ined) : protocole concurrentiel.
- Enquête en face-à-face CAPI puis téléphone : enquête emploi en continu de l'Insee (protocole séquentiel).

Le multimode peut donc concerner l'articulation entre la phase de notification/contact et le questionnaire ; il peut, dans le questionnaire, concerner tous les individus de la même manière (comme dans l'enquête CVS) ou bien au contraire, tous les enquêtés ne sont pas soumis au même mode (comme dans Virage ou Fecond modes de collecte). Dans l'administration du questionnaire, la bascule d'un mode à l'autre peut être conditionnée par un choix ou une obligation ou bien une non-réponse : par exemple seuls les répondants seront relancés sur un autre mode.

Il est possible d'assigner des segments de la population aux modes qui leur sont les plus adaptés ou bien de proposer séquentiellement plusieurs modes différents, afin d'accroître le taux de participation : ce sont les pistes poursuivies en amont de la collecte par l'adaptive design (Wagner 2008, Wagner and Couper 2011, Calinescu, Bhulai et al. 2013) ou, durant la collecte, par le responsive design (Groves and Heeringa 2006). La complexité des protocoles se paie néanmoins : risques accrus d'erreurs, coût de gestion, développement d'applications de suivi, contrôles de qualité, etc.

3. Les effets de mode et leur estimation

On peut distinguer deux grandes familles d'effet du mode de collecte : la sélection des répondants et la qualité des informations livrées par les répondants.

3.1. L'effet de composition (sélection) : l'importance du contact

Au cours de la phase de contact, la personne contactée choisit ou non de coopérer à la sélection de la personne à enquêter. Le taux de coopération conditionne une grande partie du taux de réponse à l'enquête et de sa qualité (notamment la connaissance de l'éligibilité du ménage ou de l'individu sélectionné). Coopération si le principe de l'enquête est accepté et la sélection de l'enquêté faite,

réponse si l'enquêté répond effectivement. La réponse sera définie a posteriori en fonction du jugement du caractère complet ou non du remplissage ainsi que de sa qualité. En général, c'est bien durant la phase contact que se détermine en grande partie le taux de réponse et bien sûr la composition de l'échantillon de répondants (Groves, Singer et al. 2000). Mais dans des questionnaires auto-administrés longs, les abandons ne sont pas rares. L'effet premier du mode de collecte sera donc le taux de réponse et surtout la distribution de ce taux suivant les caractéristiques socio-démographiques des personnes contactées, autrement dit la représentativité de l'échantillon de répondants suivant des caractéristiques observables.

Le mode de collecte pour le questionnaire intervient aussi directement : si le ménage n'est pas équipé d'Internet par exemple, il ne pourra répondre par ce moyen.

3.2. Biais de composition et représentativité

Comme souligné dans l'introduction, un des intérêts d'un protocole multimode est de faciliter la participation à l'enquête et donc d'en améliorer la couverture ou la représentativité : en offrant des modalités de participation variées (même si l'enquêté ne se voit pas offrir de choisir explicitement ou de manifester sa préférence pour un mode), on espère théoriquement accroître la représentation de l'échantillon. On peut objectiver la proximité de l'échantillon de répondants et la population cible par des comparaisons bivariées ou multivariées. La première méthode repose sur la comparaison des marges socio-démographiques via notamment des différences standardisées (Austin and Stuart 2015), la seconde via un indicateur de variance du taux de participation comme le R-indicateur (Bethlehem, Cobben et al. 2009, Schouten, Cobben et al. 2009). Une variance faible de R indique une propension à répondre presque uniforme suivant les catégories socio-démographiques introduites dans l'analyse, ce qui suggère un faible biais de composition de l'échantillon final (A, B) relativement à la population cible. Cet examen de la représentativité de l'échantillon de répondants est essentiel pour comprendre le processus de collecte et apprécier la qualité finale de l'échantillon obtenu.

3.3. Les effets de mesure : satisficing et désirabilité sociale

Ces deux effets sont souvent ceux que l'on a en tête lorsqu'il est question de multimode : il s'agit de l'effet propre du recours à un mode de collecte pour le remplissage du questionnaire plutôt qu'à un autre sur la qualité des réponses. Par qualité on entend souvent sincérité (biais de réponse) et non-réponse partielle (donnes manquantes), catégorie qu'il faut enrichir de l'abandon. En effet, les abandons précoces dans le questionnaire vont diminuer le taux de réponse global à l'enquête et sans doute altérer la représentativité de l'échantillon de répondants.

Désirabilité sociale

Le biais de désirabilité sociale est la formulation de réponses insincères pour produire une image avantageuse et valorisante de soi, en accord avec les attendus normatifs prévalant dans la société. Les psychologues distinguent plus précisément l'autoduperie (et l'hétéroduperie (tendances à se mentir à soi-même et à autrui) (Paulhus 1998, Tournois, Mesnil et al. 2000). Ce biais de désirabilité est lié à des caractéristiques sociales et psychologiques propres des individus, sans lien avec une pathologie. Son ampleur dépendra du sujet de l'enquête, variera suivant les questions et sera lié aux conditions de passation : elle sera maximale lorsque le répondant a le sentiment que ses réponses sont jugées ou pourraient l'être : la présence d'un enquêteur, son rôle dans la passation, les assurances de confidentialité et d'anonymat etc. ont une influence déterminante.

Le face-à-face est réputé engendrer un fort biais de désirabilité sociale, devant le téléphone, puis l'auto-administré (de Leeuw 2005, De Leeuw 2008) sans distinction claire entre Internet et le papier. Lors d'une passation en face-à-face, certains modules de questions sensibles (violence, sexualité, opinions politiques ou autres) peuvent être néanmoins auto-administrées (l'enquêté est laissé seul devant l'ordinateur et peut être équipé d'un casque afin de faciliter la lecture des questions), sans que

l'enquêteur puisse voir le déroulement du questionnaire, lire les questions ou voir les réponses (Cavalin 2009, Insee 2016). Sous casque, le biais est évidemment réduit relativement au face-à-face, mais il n'est pas certain que tout effet soit effacé : on peut en effet toujours suspecter que les habitudes de réponse prises lors des questions posées par l'enquêteur, et donc soumises à un biais de désirabilité sociale, perdurent dans le module auto-administré, en vertu d'un principe d'engagement (Kiesler 1971, Joule and Beauvois 2014) et d'auto-duperie.

Il existe des échelles psychométriques pour évaluer la tendance à la désirabilité sociale, qui sont toutes des réductions de l'échelle initiale de Crowne et Marlowe (1960) de 33 items (Crowne and Marlowe 1960, Barger 2002, Haghighat 2007). Leur qualité métrologique est parfois incertaine.

Satisficing

Le second biais de mesure bien documenté est le *satisficing*. Le terme est un mot-valise formé de *satisfying* (satisfaisant) et *sufficing* (suffisant) qu'on pourrait traduire sur le même modèle par « suffisfaisance ». Le biais est dû au fait que l'enquêté se contente d'une réponse approximative et ne cherche pas à maximiser ou optimiser son choix de réponse. La direction du biais de réponse n'est pas due à un refus de livrer des informations sensibles ou supposément dévalorisantes, mais à un refus ou une incapacité de faire des efforts. Les causes sont le défaut de motivation (l'inintérêt pour les questions ou l'enquête) mais aussi la difficulté posée par les questions et les capacités cognitives de l'enquêté. Le satisficing est donc l'opposé du *maximizing*, tendance à maximiser ou optimiser ses choix. Il résulte d'un arbitrage entre l'effort déployé pour répondre et les intérêts à le faire, dans une perspective d'économie cognitive (Simon 1957). Les théoriciens ultérieurs, à la suite de (Krosnick 1991), ont développé des échelles pour tenter de mesurer cette propension en termes de maximisation des choix (13 items) et de l'expression de regrets relativement à des choix passés (5 items) (Schwartz, Ward et al. 2002). En 1996, Krosnick et ses collègues définirent ainsi la propension au satisficing : p=Difficulté/(motivation x capacité) (Krosnick, Narayan et al. 1996). La présence d'un enquêteur motive l'enquêté et est donc un élément important permettant de réduire le satisficing. Ce dernier se manifeste à travers des réponses stéréotypées, le choix trop fréquent de modalités moyennes, de « ne sais pas » ou par un taux élevé de non-réponses ; d'autant plus que les questionnaires sont longs et répétitifs, et surtout vers la fin de ceux-ci.

Il affecte fortement la qualité de la mesure (Barge and Gehlbach 2012) mais est limité par la présence d'un enquêteur.

Pour une version courte de l'échelle de maximisation/satisficing en 6 items, on peut se reporter à (Nenkov, Maurin et al. 2008).

3.4. L'effet des adaptations du questionnaire

A ces deux effets bien documentés il faut ajouter les effets dus à l'affichage de certaines modalités ou à des vérifications automatiques qui peuvent être opérées sur certains modes et pas d'autres. Par exemple, il est possible de conditionner l'affichage ou bien la proposition orale des modalités « Ne sais pas » et « Ne veux pas répondre » à l'absence de réponse dans les modes informatisés (Internet, face-à-face CAPI et téléphone CATI) alors que ce n'est pas possible sur papier. L'affichage d'aides contextuelles et d'info-bulles peut se faire sur Internet, mais pas au téléphone ni sur papier et de façon particulière en face-à-face. Des contrôles automatiques de saisie sont possibles dès qu'un ordinateur est mobilisé pour l'administration du questionnaire (Internet, mais aussi CAPI et CATI), alors que ce n'est pas le cas en auto-administré. Des contrôles de pertinence/vraisemblance sont également implémentables dans les modes informatisés.

La mesure du satisficing peut être affectée par ces différences d'affichage et de contrôle.

3.5. Des effets de modes spécifiques aux collectes par enquêteur

Les enquêtes intermédiées sont sujettes à un effet enquêteur intrinsèque (indépendamment de toute comparaison avec un autre mode de collecte). En effet, les questionnaires sont groupés par enquêteur, ce qui fait que les données sont multiniveaux, induisant une perte de puissance en vertu d'un grappage des données. Cet effet se manifeste via deux mécanismes. Par l'existence d'un biais de sélection des répondants (les jours et horaires de passage ou d'appel des enquêteurs peuvent en effet différer et accroître différentiellement les chances de contact, de coopération et de complétion des questionnaires de certains segments de la population ; mais aussi par l'influence que peuvent avoir les caractéristiques des enquêteurs (sexe, âge, expérience, voix, présentation, etc.) sur les réponses fournies par les répondants à certaines questions. Cet effet est mesurable même dans les enquêtes téléphoniques car le sexe de l'enquêteur est en effet immédiatement et presque infailliblement perçu (Davis, Couper et al. 2010, Chun, Tavarez et al. 2011). La littérature montre par exemple que les opinions féministes ou favorables aux minorités ethniques est plus facile lorsque la question est posée par une femme (Huddy, Billig et al. 1997), de même les déclarations de comportements et pratiques sexuelles (Wilson, Brown et al. 2002) ou de certains types de violences sexuelles subies (Dailey and Claus 2001). D'autres effets sur la participation, ont pu être démontrés en face-à-face (Durrant and Steele 2009, Durrant 2010), et il en existe certainement également d'autres au téléphone.

3.6. Définition des effets de mesure et de sélection

Plaçons-nous dans le cas simple de la comparaison de deux échantillons indépendants A et B définis par le mode de collecte M=A et M=B, où l'on dispose de variables auxiliaires et sociodémographiques X et SD, et une variable Y qui est la variable cible de l'enquête. A est le mode de référence.

L'effet de mode se définit par EM= BS + BM. Autrement dit, comme la somme du biais de sélection et du biais de mesure. Le premier se définit comme la contribution de la différence des compositions des échantillons A et B sur la mesure de Y ; le second comme la contribution de la différence des modes sur la mesure de Y qui n'est pas expliquée par la différence de composition ou, d'une façon plus restrictive, par la différence de mesure induite par le mode pour les individus ayant le même profil socio-démographique SD-X. La première définition fait référence aux méthodes de décomposition économétriques type Oaxaca-Blinder (Fortin, Lemieux et al. 2011) ; elle envisage le problème en termes de variables. La seconde fait plus spécifiquement référence à une approche en termes de profils des personnes (Nopo 2004).

C'est le biais de mesure que l'on a généralement en tête lorsqu'on parle d'effet de mode. Il se définit conceptuellement comme l'effet causal ou résiduel (c'est-à-dire net de l'effet de composition) du mode sur la mesure de Y pour chaque individu : autrement dit comme l'effet de répondre sur le mode B plutôt que A pour chaque répondant. Comme un unique répondant n'a répondu qu'à un seul questionnaire sur un seul mode, cet effet est généralement inestimable (il peut l'être dans les plans d'expérience en cross-over où chaque individu est son propre témoin, mais le biais de mémoire lié à la réinterrogation rendent les estimations fragiles). En revanche on peut l'estimer en moyenne en s'appuyant sur les échantillons de répondants similaires en SD-X sur les deux modes : BM=E[Y(B)-Y(A)|SD-X]. Cette estimation causale est celle du modèle contrefactuel de Rubin (Imbens and Rubin 2015).

Si les échantillons A et B étaient de grande taille et représentatifs de la population cible, si l'attribution du mode aux répondants était strictement aléatoire et si aucune non-réponse totale n'était observée dans les deux modes (le taux de participation à l'enquête sur le mode A étant de 100% comme celui observé sur le mode B), alors E(Y(A)-Y(B)) nous donnerait l'effet de mesure. En effet, il n'y aurait aucune sélection des répondants sur un mode plutôt que l'autre, les caractéristiques moyennes des répondants dans les deux échantillons seraient identiques et aucun effet de composition ne serait à l'œuvre.

Dans les cas réels, il y a plusieurs problèmes. D'abord, les échantillons sont de taille souvent réduite, ce qui peut amener à des erreurs d'échantillonnage. Ensuite, A et B souffrent de non-réponse totale, ce qui induit des biais de composition. Si la non-réponse totale ne dépend que de SD-X, et le mode est indépendant de Y, ce qui est le cas des mécanismes d'assignation réguliers (Imbens and Rubin 2015) alors on peut utiliser plusieurs méthodes équivalentes.

Notez qu'en vertu de ce qui a été rappelé à propos des types de contacts, ce que l'on observe et estime comme effet de composition est en réalité dû à la série des effets cumulés des différents types et modes de contact et de collecte. De même pour l'effet de mesure : on estime un effet dû à la conjonction des différents éléments caractérisant les types de questionnaires/supports mobilisés pour la collecte, y compris les déviations en termes de présentation et les adaptations des questions/modalités de réponse. Il est donc (dans l'absolu) présomptueux d'affirmer que c'est Internet (par exemple) qui cause cet effet, notamment en vertu des modifications/adaptations éventuelles des questionnaires et du contexte de passation (lieu, présence de tiers etc.). C'est un effet global dont le support est un élément important mais pas unique.

3.7. Estimation du biais de mesure

Parce que le biais de mesure est fondé sur une différence individuelle, on ne peut estimer un biais de mesure que pour les individus comparables dans les deux modes A et B, ce qui suppose l'existence d'un sous-ensemble non-vides d'individus ayant des profils SD-X similaires, ce que l'on appelle le support commun. Ce n'est pas toujours possible si les modes sont attribués à des ensembles disjoints de répondants (par sexe, âge par exemple). L'idée est de se rapprocher au maximum de l'expérimentation randomisée, soit d'une situation hypothétique où l'attribution du mode de collecte est indépendante de Y (idéalement faite avant toute mesure de Y) et des caractéristiques SD-X.

On utilise souvent un score de propension pour comparer globalement les échantillons A et B : on modélise la probabilité de répondre en B plutôt qu'en A conditionnellement à des variables SD et X choisies, c'est-à-dire en lien avec Y. La propriété équilibrante du score de propension fait que conditionnellement au score l'attribution de M est aléatoire (Rosenbaum and Rubin 1983, Imbens and Rubin 2015) : les deux échantillons A et B présentent alors des distributions SD-X équivalentes conditionnellement au score.

Pour mener l'estimation il faut se ramener au support commun. Cela peut se faire par matching direct ou par matching sur le score de propension ou bien encore par restriction à un sous-ensemble d'individus dans la plage commune du score de propension (Crump, Hotz et al. 2006, Imbens and Rubin 2015), ces méthodes s'avérant en pratique d'efficacités similaires avec les paramétrages ad hoc. L'estimation ne porte donc quasiment jamais sur la totalité des échantillons A et B et donc *a fortiori* pas sur la population cible. Plus on est exigeant dans le choix des variables SD-X, plus le support commun est constitué d'individus similaires, mais plus il est réduit : on gagne en validité interne ce que l'on perd en validité externe c'est-à-dire généralisation.

Les méthodes d'estimation (sur le support commun) sont très nombreuses : repondérations, régressions, régressions pondérées dites doublement robustes, régressions stratifiées sur le score de propension (méthode de subclassification), matching et régressions etc. (Imbens and Rubin 2015, Morgan and Winship 2015). Il existe aussi des méthodes par imputation (Suzer-Gurtekin 2013, Austin 2014, Park, Kim et al. 2016).

On trouve parfois des pondérations directes d'un échantillon sur l'autre, par calage ou post-stratification (Deville and Sarndal 1992), sans référence au support commun : l'idée est ici encore de neutraliser les biais de composition en SD-X dans les échantillons A et B. Le calcul de l'écart (pondéré) entre les deux échantillons fournit une estimation de l'effet de mesure moyen. Toutefois, cela est souvent insuffisant : il faut prendre en considération de nombreux croisements des variables SD-X ainsi que leurs différents moments (Imbens and Rubin 2015), et cela peut conduire à des poids très variables, ce qui peut nuire à la stabilité des estimations. En fait, tout est affaire de précision dans

le choix du support commun et de généralisation.

Le biais de mesure peut être estimé dans la population (ou du moins l'échantillon) complet, c'est alors l'Average treatment effect (le traitement étant ici le mode alternatif B par opposition au mode de référence A). Il peut aussi être estimé parmi les traités uniquement (ATT ou average treatment effect among the treated) ou bien parmi les contrôles uniquement (i.e. les observations du mode de référence A : ATC ou average treatment effect among the controls). On cherche en général à estimer l'ATE, mais l'ATT peut avoir son utilité, car il représente l'effet du mode B pour les personnes qui, typiquement, répondent sur ce mode plutôt que le mode de référence A.

Notez que le biais de sélection peut lui-même être affecté d'un biais de mesure : ce n'est pas parce que le biais de mesure n'est estimé que sur le support commun que la réponse à Y des individus hors support commun n'est pas différente suivant le mode. Le mettre en évidence, par exemple relativement à une variable donnée (disons l'âge) implique des analyses particulières (sur les individus hors support commun).

3.8. Hypothèses nécessaires à l'identifiabilité de l'effet de mesure

- Il doit exister un mécanisme connu ou modélisable de l'allocation du mode ;
- la sélection des répondants sur chacun des modes ne doit pas dépendre de la variable Y étudiée (non-réponse non-ignorable) : le mécanisme est régulier ;
- il doit y avoir indépendance de la variable de traitement (le mode) avec les variables contrefactuelles Y conditionnellement aux variables SD-X ou au score de propension prédisant la réponse sur A ou B ;
- la réponse au traitement (mode A ou B) de chaque individu *i* (Yi) ne doit pas dépendre de la réponse au traitement Yj d'autres individus *j* : les individus ne s'influencent pas les uns les autres (aucune interférence). Il n'existe qu'une seule forme de traitement ou du moins les éventuelles distinctions envisageables sont sans effet sur les Y observés et contrefactuels. Ces deux éléments composent l'hypothèse SUTVA (stable unit treatment value assumption) formulée par Rubin dans sa forme la plus récente (Imbens and Rubin 2015), p 10-13. Une des conséquences est qu'*a priori*, il ne faut pas confondre le choix d'Internet et l'imposition d'Internet dans une analyse de l'effet de mesure dans une enquête multimode Internet vs face-à-face.

  On notera que Morgan et Winship (2015, page 48) utilisent une formulation plus ancienne de SUTVA par Rubin, qui ajoute que la valeur Y d'un individu i ne dépend pas du mécanisme d'allocation du traitement à i ni du traitement reçu par les autres individus j (Rubin 1986), mais les deux sont proches.

Ces hypothèses sont généralement vérifiées dans les enquêtes ou du moins peu sujettes au doute, sauf sans doute la seconde, qui reste très problématique et dont ne peut pas s'assurer. Il faut donc argumenter pour en défendre la plausibilité.

3.9. Qu'est-ce que corriger un effet de mode ?

Corriger un effet de mode peut avoir deux significations. La première, impropre, veut généralement dire corriger l'effet de composition. Autrement dit, cela passera par la tentative de neutraliser les différences de composition des deux échantillons, rendre B comparable à A en termes SD-X. Cela n'est envisageable que pour les protocoles où A et B sont attribués de façon aléatoire. Cela peut se faire par repondération, par exemple par calage de B sur A (ce qui permet de mobiliser des variables

X et plus largement toutes les variables utiles mesurées dans l'enquête) ou bien de A et B sur la population cible (ce qui en général ne permet que d'utiliser les variables SD). En fait cette technique ne corrige pas l'effet de mesure s'il existe, mais permet de l'estimer pour B relativement à A (cf. 3.6).

Lorsqu'une enquête est véritablement multimode avec un protocole séquentiel ou concurrentiel, il est nettement plus difficile de procéder ainsi. De plus, dans le cas d'une enquête multimode, les échantillons collectés sur les différents modes sont complémentaires et non substituables : chaque échantillon est sujet à un biais de composition (même modélisable) et c'est leur union qui est censée représenter la population cible. Par conséquent cela interdit tout traitement simple par repondération visant à redresser un échantillon sur l'autre.

Dans tous les protocoles séquentiels ou concurrentiels dans lesquels l'attribution du mode n'est pas aléatoire, on ne va généralement pas corriger l'effet de composition : les observations de A et de B sont à considérer et traiter ensemble. En revanche on peut toujours, au moins partiellement, corriger l'effet de mesure.

La seconde signification, plus correcte, désigne la correction de l'effet de mesure. Il s'agit alors de modifier les réponses collectées sur le mode alternatif B jugées déviantes (influencées par le mode B relativement à A). Conceptuellement, l'effet de mesure est un problème de niveau de réponse à Y et non de poids : la solution viendra d'une imputation.

On peut toutefois tenter de « corriger » l'effet de mesure par repondération lorsqu'on est capable d'assigner la valeur de référence d'une variable Y au sous-échantillon interrogé sur le mode alternatif dans un calage. Si l'enquête est bimodale avec assignation aléatoire de deux modes A et B, chaque échantillon de répondants A et B vise la représentation de la même population cible et l'on peut alors assigner au mode alternatif B la proportion Y estimée dans le mode de référence A adéquatement redressé. Dans le cas où l'attribution du mode n'est pas aléatoire (protocole séquentiel ou concurrentiel), cela est plus délicat, mais il reste possible d'imposer une cale supplémentaire forçant la valeur de Y dans l'échantillon complet (A, B) redressé via calage. Cela nécessite de disposer d'une information exogène relativement au niveau de Y dans la population cible (via une autre enquête par exemple) ou bien d'être capable de calculer la « vraie » proportion Y dans la population à partir de l'estimation causale de l'effet de mesure dans B relativement à A. Cette méthode par repondération n'est que symptomatique : par exemple, les individus qui auraient dû déclarer tel type de comportement sur le mode B ne l'ont pas fait, mais plutôt que de modifier leur réponse par imputation, on va donner davantage de poids aux individus qui ont bien déclaré ce comportement pour obtenir une prévalence globale correcte. Une analyse détaillée comparant les deux modes A et B pour Y, même avec la pondération, mettra en évidence une différence de niveau en Y. Par contraste, la méthode d'imputation est curative.

3.10. Quand corriger un effet de mode ?

Les cas où un mode alternatif B est utilisé pour administré une partie du questionnaire administré sous le mode A (B est niché dans A), par exemple un sous-questionnaire en reconnaissance vocale ou en audio-CASI, ne posent évidemment aucun problème car tous les répondants sont assignés à ces deux modes de façon identique : il n'y a donc pas de problème d'agrégation des données.

Dans tous les autres protocoles, la décision dépendra des situations, du protocole, des objectifs du choix du multimode et des tailles d'échantillons A et B. La première chose à faire est de produire une estimation solide de l'effet de mesure (plus que de l'effet de sélection, qui est généralement perçu comme moins nuisible en raison du fait qu'il reflète la contribution des deux modes à la représentativité de l'échantillon de répondants). La seconde est l'interprétation de l'estimation et de ses conséquences.

Si le biais de mesure (estimé sur le support commun) est bénéfique on pourrait le conserver ; s'il est néfaste, on peut le corriger. Il faut savoir que la correction implique également une augmentation de

variance.

Si l'enquête fait partie d'une série réalisée jusqu'à présent réalisée en monomode (A), l'introduction de B va perturber la série. Cela peut gêner les interprètes et les utilisateurs des chiffres. On peut vouloir corriger ou bien rétropoler les anciennes mesures, mais cela peut être délicat. Prenons le cas de l'introduction d'Internet une année donnée. Rétropoler les anciennes mesures monomodes peut s'avérer difficile sinon impossible pour les années où Internet n'existait pas. Pour les années où Internet était suffisamment répandu, on peut imaginer redresser les chiffres pour les individus qui ont les caractéristiques du support commun A B de l'enquête courante, mais que faire pour les autres ? Ou alors il faut accepter une correction pour l'ensemble, mais moins fiable car effectuée sur l'ensemble des observations, donc sur un support commun défini très souplement (on fait l'hypothèse que les biais estimés grossièrement sont justes ou bien que les biais estimés avec précision sur le support commun sont extrapolables à toutes les observations).

Par ailleurs, si Internet devient le mode ultra majoritaire dans les années qui suivent, rétropoler deviendra impossible sauf à utiliser la première enquête multimode pour ce faire. Mais, une telle rétropolation fait implicitement l'hypothèse que les biais ne varient pas dans le temps, ce qui est discutable.

Estimer l'effet de mesure peut en revanche être utilisé à profit dans les enquêtes transversales pour réestimer le niveau de la variable Y. Par exemple, dans l'enquête Virage, l'estimation de l'effet mode de collecte permet de montrer qu'une orientation sexuelle LGB est presque deux fois plus souvent déclarée sur Internet qu'au téléphone, ce qui conduit à presque doubler l'estimation globale de la proportion de LGB dans la population : la conséquence est qu'il est possible de recaler l'échantillon initial avec cette cale supplémentaire afin d'obtenir la bonne proportion de LGB.

Dans les séries d'enquête où la part des modes évolue, on peut envisager de fixer ces parts à des valeurs moyennes pour caler la série d'enquêtes, afin de contenir l'effet de mode, sans le corriger. Cela ne revient qu'à corriger le biais de composition (Buelens and van den Brakel 2013) et à imposer une composition moyenne des modes de collecte, et cela ne peut être que temporaire si les évolutions concernant les parts des modes de collecte deviennent trop importantes ou si un mode disparaît.
Les méthodes par imputation sont théoriquement en position de pouvoir corriger tout l'effet de mesure sur l'ensemble des variables Y qui en sont victimes, mais leur mise en œuvre souffre des mêmes difficultés et limitations relativement à l'évolution des parts des modes de collecte au cours des exercices successifs de l'enquête. De plus, ces deux types de méthodes de correction posent quelques difficultés sur le plan déontologique.

3.11. Les difficultés spécifiques liées à l'estimation d'un effet de mesure

Si le cadre théorique de l'estimation causale est bien défini par l'approche contrefactuelle, une difficulté supplémentaire surgit pour les enquêtes multimodes dans lesquelles il faut estimer un effet de mesure. Dans l'approche de Rubin qui est historiquement liée à la mesure de l'effet causal d'un traitement dans le domaine de la santé, il faut se cantonner aux variables mesurées antérieurement à l'attribution du traitement. Ce sont ces variables prétraitement qu'il faut introduire dans le modèle de score de propension, et uniquement celles-ci. Il ne faut pas non plus examiner leurs liens avec l'outcome Y (Rubin 2007). L'idée est en effet de se ramener à l'essai randomisé canonique dans lequel l'outcome n'est connu qu'à l'issue de l'essai.

Toutefois, dans les enquêtes multimodes, le traitement étant le mode, et les informations recueillies l'étant généralement durant l'enquête, donc via un mode particulier, toutes les variables sont susceptibles d'être sujettes à un effet de mesure. Il n'y a pour ainsi dire aucune variable prétraitement, sauf à disposer de variables mesurées dans la base de sondage, ou vraiment antérieurement à l'attribution du mode, dans une vague antérieure de l'enquête ne mobilisant qu'un seul mode, ou par tout autre moyen sans lien avec le traitement. Sexe et âge sont peu suspects de tels biais, mais la profession, le diplôme, le statut d'emploi le sont plus si l'on compare un mode intermédié et un mode

auto-administré : les travaux expérimentaux menés à l'Insee dans le cadre de la refonte de l'enquête Emploi en témoignent.

Il est tentant d'utiliser des échelles psychométriques pour estimer le penchant à la désirabilité sociale ou au satisficing des enquêtés afin de le contrôler, mais même si ces échelles sont réputées fiables, elles sont sujettes aux biais qu'elles estiment en cas de passation sur des modes différents. Ce problème de stabilité de la mesure (Vannieuwenhuyze and Loosveldt 2013) est un problème conceptuel insoluble, sauf encore à disposer de telles mesures antérieurement à la réponse sur les modes étudiés (prétraitement).

Enfin, il faut rappeler que si le choix du mode est lié aux variables cibles (mécanisme d'assignation non régulier ou non-réponse non-ignorable) alors il n'est plus possible d'estimer l'effet de mesure sans biais : on est dans le cas d'une sélection non-ignorable (Imbens and Rubin 2015). On peut toujours suspecter que des variables non mesurées qui sont des facteurs de confusion causant le choix du mode et Y jouent. De plus, l'estimation ne peut se faire que sur le support commun.

4. Une méthode parcimonieuse

Il est possible de se fonder sur la définition du biais de mesure pour en proposer une correction. Pour cela on s'appuie sur la méthode du score de propension. On choisit un ensemble raisonnable de variables associées à Y et à M, sous-ensemble des variables socio-démographiques et auxiliaires SD-X. Une fois la modélisation faite, on peut apparier les individus du mode A à ceux du mode B, le mode A étant la référence.

On obtient alors la partition des observations suivantes. Un groupe A0 est issu de A sans contrepartie ; un groupe A1 est issu de A et est apparié à B1 ; un groupe B1 est issu de B et est apparié à A1 ; un groupe B0 est issu de B et est sans contrepartie. A1-B1 représente un sous-ensemble du support commun, tandis que A0 et B0 forment un ensemble un peu plus large que les représentants du biais de composition. A1-B1 est composé de paires de jumeaux en SD-X ; en son sein, on peut identifier les paires de jumeaux parfaits en SD-X, et Y, notées A1_1-B1_1 et les autres, notées A1_0-B1_0. Les paires imparfaites sont celles qui portent l'effet de mesure, par définition. Il suffit alors d'imputer la valeur de Y des observations de B de ces paires imparfaites. On prendra comme modèle de relation SD-X celui qui existe au sein de A1-B1_1. Cela peut être fait de manière stochastique, les imputations multiples pouvant permettre une estimation de la variance supplémentaire induite.

4.1. Exemple de CVS-VVS

Présentation des données

Nous traiterons ici d'un cas d'école construit à partir des données collectées en 2013 dans l'enquête annuelle de victimation Cadre de vie et sécurité (CVS), enquête en face-à-face, et dans sa réplication sur Internet, Vols, violence et sécurité (VVS) (Razafindranovona 2016). Les enquêtes mobilisées avaient été réalisées indépendamment en 2013 (taux de réponse de 63% et 32%). Nous nous plaçons dans l'hypothèse fictive où les deux échantillons devraient être agrégés, c'est-à-dire dans le cadre d'une enquête multimode à échantillons indépendants. CVS étant une enquête annuelle existant depuis 2007, nous considèrerons le face-à-face comme la référence en termes de mode.

Les deux enquêtes ont été réalisées en parallèle : il s'agit de deux échantillons indépendants. Les plans de sondages des deux enquêtes sont différents mais visent à représenter la même population cible, celle des résidents métropolitains francophones âgés de 14 ans et plus en ménages ordinaires.

Les variables cibles (Y) sont les six variables binaires portant sur la déclaration d'un des types

d'événement suivants au cours des 24 derniers mois : vol avec violence ; vol sans violence ; cambriolage ; vol de véhicule ; violence physique ; menaces.

Les variables sociodémographiques (SD) utilisées dans le redressement de CVS sont au nombre de huit : sexe, âge, diplôme, PCS, type de logement, statut d'occupation du logement, taille d'unité urbaine, type de ménage.

Les variables auxiliaires (X) d'opinion sont au nombre de quatorze (binaires ou ordonnées à quatre modalités ou cinq) et sont relatives aux opinions sur la police, la justice, et certains types de problèmes de sociétés (chômage, sentiment d'insécurité, etc.). Comme indiqué précédemment, elles sont très largement suspectes d'un effet de mesure. En revanche, elles sont très fortement associées aux Y.

Objectifs :

1. mesurer la représentativité des échantillons CVS, VVS et l'échantillon composite CVS-VVS ;
2. identifier et imputer les observations de VVS qui, dans CVS-VVS portent l'effet de mesure ;
3. comparer les estimations obtenues avec les initiales et les estimations de référence de CVS seul.

Représentativité de l'échantillon composite et des échantillons propres

Les marges de la population cible et des différents échantillons sont données Tableau 1. Elles montrent que CVS et VVS peinent à interroger des jeunes, que les hommes manquent dans CVS plus que dans VVS, que les tailles d'UU des répondants collent davantage aux marges du RRP dans VVS que CVS, que les types de ménages ou les PCS sont aussi mal représentées dans VVS que CVS, mais que la distribution des types de logements est de meilleure qualité dans VVS que CVS. La synthèse de ces résultats, exprimée en moyenne des écarts absolus de points de pourcentages est fournie tableau 2. Elle montre que globalement, VVS est un peu plus proche du RRP que CVS, et que l'agrégation occupe une position intermédiaire. Cette appréciation de la représentativité est toutefois assez grossière puisqu'elle n'est qu'une agrégation de mesures bivariées.

Pour avoir une mesure multivariée, on peut utiliser le R-indicateur (Bethlehem, Cobben et al. 2009). Il s'agit d'estimer la variation de la probabilité de répondre à une enquête suivant les critères socio-démographiques à l'issue d'une modélisation : plus la valeur est proche de 1, plus la variation est réduite et plus le taux de réponse est indépendant des caractéristiques utilisées dans la modélisation. La formule est la suivante : $R = 1 - 2 \times sd(\widehat{p})$, $\widehat{p}$ étant la probabilité estimée par régression logistique incluant les 8 variables socio-démographiques de calage et toutes leurs interactions bivariées.

R(CVS)=0.76 ; R(VVS)=0.63 ; R(agrégation)=0.69. CVS est donc de qualité supérieure à VVS, l'agrégation se situant entre les deux. Dans ce cas précis, la supériorité du taux de réponse dans CVS est bien corrélée à une représentativité supérieure.

Notons que le biais de composition de VVS est important, notamment pour les variables auxiliaires X. Il pourrait être corrigé par repondération avant agrégation, mais cela aboutit à des poids calés trop dispersés (coefficient de variation CV=504) comparativement à ceux d'un calage direct (CV=93). Par ailleurs, comme rappelé précédemment, ces variables X, pour utiles qu'elles soient dans chaque mode, sont elles-mêmes sujettes à un fort effet de mesure qui prévient leur usage dans ce but. Nous considérons donc l'échantillon CVS-VVS calé classiquement comme notre échantillon multimode fictif de travail.

Détermination du support commun et appariement

Au sein de l'échantillon CVS-VVS calé, nous modélisons par régression logistique l'appartenance à VVS. Sont intégrées les 8 variables SD plus leurs interactions bivariées. Le support commun classique est défini par la plage de variation du score de propension commune aux deux sous-

échantillons : seules 73 observations en sont exclues dans CVS et 126 dans VVS, ce qui est négligeable.

À partir du score, nous apparions les observations de VVS à celles de CVS, 1-1 sans remise, avec choix aléatoire dans une plage d'ex-aequo de 0,2 écart-type du logit du score, en imposant l'égalité des variables SD les plus liées aux variables cibles : sexe, âge, PCS, taille d'unité urbaine, type de ménage, type de logement et statut d'occupation du logement.

Au total, 7957 paires d'observations CVS-VVS sont constituées (représentant 55 % de CVS, 62 % de VVS). En l'absence de biais de mesure, il ne devrait pas y avoir de différences de victimations Y en leur sein. Les observations de VVS restantes représentent une bonne partie du biais de composition d'Internet du point de vue des variables SD. Nous reviendrons sur cette notion.

Stratégie d'imputation

Au sein des paires, 2512 (32%) présentent au moins une différence de Y, 1788 une seule et 169 au moins trois. Seule une fraction des paires sera imputée : elle est tirée aléatoirement de façon à ce que la proportion imputée de chaque victimation égale celle dans CVS. Le taux de tirage est défini variable par variable, en fonction du taux de victimation observé dans CVS. Les effectifs à imputer sont faibles : entre n=149 –vols avec violence- et n=381 –vols sans violence-, sur un total de 7957 observations à chaque fois (soit entre 1.8% et 4.7% des valeurs).

L'imputation est faite uniquement sur le score de propension, mais en prenant comme modèle la relation entre celui-ci et les variables Y estimée dans l'échantillon en face-à-face.

La syntaxe est la suivante :

```
proc mi data=out_rxmi(where=(apparie_x=1))   nimpute=5 out=imputeout_rxmi seed=1;
class   qaa                              vavn_rxmi vsvn_rxmi vphyn_rxmi vlogn_rxmi vvehn_rxmi men_rxmi
var     vavn_rxmi vsvn_rxmi vphyn_rxmi vlogn_rxmi vvehn_rxmi men_rxmi pscoresd
fcs logistic (vavn_rxmi vsvn_rxmi vphyn_rxmi vlogn_rxmi vvehn_rxmi men_rxmi=pscoresd /classeffects=include
likelihood=augment descending);
mnar model(vavn_rxmi vsvn_rxmi vphyn_rxmi vlogn_rxmi vvehn_rxmi men_rxmi/modelobs=(qaa='0') ) ; run ;
```

Résultats

Après imputations stochastiques multiples de toutes les valeurs manquantes sur l'ensemble des paires (n=5), l'effet de mesure (conditionnellement aux variables considérées) est éliminé au sein du support commun : toutes les différences sont à 0 (sans la pondération). Avec la pondération, presque toutes les différences de moyenne entre les observations CVS et les observations VVS sont à 0 (tableau 3).

Les victimations ainsi réestimées dans CVS-VVS représentent alors entre 74 % et 92 % des victimations initiales, soit, par rapport à CVS seul, une augmentation moyenne de 10 % (entre -3 % pour les violences physiques et +31 % pour les vols avec violence), contre un écart initial de +30 %. Les résultats sont fournis tableau 4.

La baisse est modeste, mais l'effet causal de l'interrogation sur Internet (estimé par la méthode de régression/subclassification sur le score de propension, contrôlant les variables SD-X (Imbens and Rubin 2015), est notablement réduit par rapport à la situation initiale : +0,48 % (p=0,008) vs +1,38 % (p=0,000) pour les vols avec violence, +0,68 % (p=0,016) vs +2,97 % (p=0,000) pour les vols sans violence, -0,29% (p=0,204) vs +0,13% (p=0,587) pour les violences physiques, -0,59 % (p=0,063) vs +1,28% (p=0.000) pour les cambriolages, -0,37% (p=0,264) vs 0,47% (p=0,167) pour les vols de véhicules et -0,52% (p=0,119) vs +1,14% (p=0,001) pour les menaces. La variance d'imputation est négligeable (<0,5%).

5. Discussion

Nous avons présenté une synthèse des propriétés des enquêtes en rappelant leur dimension multimode intrinsèque. Nous avons ensuite présenté les biais relatifs générés par l'emploi de plusieurs modes dans les enquêtes, en termes de représentativité et de mesure ainsi que les moyens pour les mesurer et les estimer. Enfin, nous avons présenté une méthode pour contrôler l'effet de mesure.

Les limites de notre méthode sont les suivantes. Elle impose d'abord de recourir à des variables indépendantes du mode de collecte pour la modélisation du score de propension et l'appariement, ce qui est classique pour ce type de données. Ensuite, elle ne corrige actuellement que l'effet de mesure sur une partie de l'échantillon, celle qui est issue d'un appariement 1-1 sans remise. Il est possible de modifier cette partie de la procédure pour assurer un appariement 1-1 avec remise dans le mode A, ce qui permettrait d'accroître le nombre de paires et donc de corriger davantage d'observations et donc assurerait un meilleur contrôle de l'effet de mesure.

La procédure ne fonctionne que sur des observations comparables : si le recouvrement des sous-échantillons A et B est faible, la correction ne s'effectuera que sur la partie commune, qui peut être réduite. La méthode n'utilise que les informations collectées dans l'enquête : elle n'incorpore pas d'information exogène, par exemple relative à l'effet du mode sur une autre sous-population que celle du support commun, qui aurait été mesuré dans une autre enquête.

Tableau 1 : marges des différents échantillons (pondérés par les poids de sondage)

	age				sexe				Diplôme				Tuu				Typmen				CSP				Typlog				Stoc			
	RRP	CVS	VVS	CVS VVS	RRP	CVS	VVS	CVS VVS	RRP	CVS	VVS	CVS VVS	RRP	CVS	VVS	CVS VVS	RRP	CVS	VVS	CVS VVS	RRP	CVS	VVS	CVS VVS	RRP	CVS	VVS	CVS VVS	RRP	CVS	VVS	CVS VVS
1	16	8	9	8	48	44	48	46	25	28	17	23	16	17	15	16	19	37	34	35	1	1	1	1	38	46	32	39	38	43	34	38
2	23	22	19	21	52	56	52	54	9	8	8	8	29	32	30	31	9	8	3	6	4	3	3	3	63	54	68	61	62	57	66	62
3	17	16	18	17					22	22	23	22	30	30	30	30	28	27	25	26	9	9	11	10								
4	16	17	19	18					18	16	19	17	25	22	25	23	45	28	38	33	14	13	16	14								
5	14	17	18	18					12	11	15	13									17	16	16	16								
6	15	21	17	19					14	15	18	16									13	11	10	11								
7																					26	34	31	33								
8																					17	12	12	12								
T		3	3	3		4	0	2		1	3	1		2	0	1		9	8	8		2	2	2		8	5	2		5	5	0

RRP=EAR 2013

Note : 0.6% des CSP ont été recodées dans CVS, contre 5.0% dans VVS.

1-8 : marges des variables socio-démographiques

T : Ecart moyen total à RRP sur l'ensemble des modalités de la variable socio-démographique

Tableau 2 : synthèse des écarts de distributions des variables de calage dans les trois échantillons

	CVS	VVS	Agrégation
age	19.8	20.9	19.6
sexe	7.1	0.4	4.0
Diplôme	8.7	19.6	8.8
Tuu	7.2	1.6	3.9
Typmen	35.6	31.0	33.4
CSP	17.8	18.2	16.9
Typlog	16.5	10.0	4.1
Stoc	9.1	9.4	0.4
Moyenne	3.6	3.3	2.7

Tableau 3 : T-tests au sein des 7957 paires, échantillon brut et pondéré, analyse poolée sur les 5 imputations

	Échantillon brut					Échantillon pondéré				
Variable	Moyenne (VVS-CVS)	StdErr	LCL	UCL	p	Moyenne (VVS-CVS)	StdErr	LCL	UCL	p
Vol avec violence	0.01%	0.14%	-0.27%	0.28%	0.972	-0.12%	0.15%	-0.42%	0.18%	0.428
Vol sans violence	0.02%	0.25%	-0.47%	0.51%	0.944	-0.10%	0.25%	-0.60%	0.39%	0.684
Violence physique	0.04%	0.20%	-0.36%	0.43%	0.863	0.00%	0.21%	-0.40%	0.41%	0.985
Cambriolage	-0.02%	0.29%	-0.60%	0.56%	0.952	-0.24%	0.31%	-0.84%	0.36%	0.433
Vol de véhicule	-0.01%	0.30%	-0.60%	0.59%	0.980	-0.84%	0.32%	-1.47%	-0.22%	0.008
Menaces	0.04%	0.32%	-0.58%	0.67%	0.893	0.60%	0.31%	-0.01%	1.22%	0.053

Tableau 4 : Synthèse des estimations de victimations

	CVS			CVS-VVS pré imputation				CVS-VVS post imputation				
variable	%	LCL	UCL	%	LCL	UCL	Ratio/ CVS	%	LCL	UCL	Ratio/ Pré-imputation	Ratio/ CVS
Vol avec violence	0.99%	0.83%	1.15%	1.75%	1.59%	1.90%	176%	1.30%	1.16%	1.43%	74%	131%
Vol sans violence	2.78%	2.51%	3.04%	4.43%	4.19%	4.67%	159%	3.38%	3.17%	3.60%	76%	122%
Violence physique	2.20%	1.96%	2.44%	2.34%	2.16%	2.52%	106%	2.13%	1.96%	2.30%	91%	97%
Cambriolage	3.74%	3.43%	4.05%	5.09%	4.83%	5.34%	136%	4.16%	3.92%	4.39%	82%	111%
Vol de véhicule	4.52%	4.18%	4.86%	5.14%	4.88%	5.40%	114%	4.72%	4.47%	4.98%	92%	105%
Menaces	4.53%	4.19%	4.86%	5.69%	5.42%	5.96%	126%	4.88%	4.63%	5.14%	86%	108%
Total	3.13%			4.07%			130%	3.43%			84%	110%

Références

Austin, P. C. (2014). "Double propensity-score adjustment: A solution to design bias or bias due to incomplete matching." Statistical methodes in medical research **0**(0): 1-33.

Austin, P. C. and E. A. Stuart (2015). "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies." Stat Med **34**(28): 3661-3679.

Barge, S. and H. Gehlbach (2012). "Using the theory of satisficing to evaluate the quality of survey data." Research in Higher education **53**(2): 182-200.

Barger, S. D. (2002). "The Marlowe-Crowne Affair: Short Forms, Psychometric Structure, and Social Desirability." Journal of Personality Assessment **79**(2): 286-305.

Bethlehem, J., F. Cobben and B. schouten (2009). "Des indicateurs de la représentativité aux enquêtes." Techniques d'Enquêtes(Recueil du symposium 2008 de Statistique Canada): 1-10.

Biemer, P. P. (2010). "Total survey error, Design ,implementation, and evaluation." Public Opinion Quarterly **74**(5): 817-848.

Buelens, B. and J. van den Brakel (2013). Measurement error calibration in mixed-mode surveys. Discussion paper. The Hague/Heerlen, CBS. **13**.

Calinescu, M., S. Bhulai and B. Schouten (2013). "Optimal resource allocation in survey designs." European Journal of Operational Research **226**(1): 115-121.

Cavalin, C. (2009). L'élaboration du questionnaire et du protocole de collecte: innovations et précautions méthodologiques. Violences et santé en France: Etats des lieux. F. Beck, C. Cavalin and F. Maillochon. Paris, La documentation française.

Caveng, R. (2012). "La production des enquêtes quantitatives." Revue d'anthropologie des connaissances **6**(1): 65-88.

Chun, H., M. I. Tavarez, G. E. Dann and M. P. Anastario (2011). "Interviewer gender and self-reported sexual behavior and mental health among male military personnel." Int J Public Health **56**(2): 225-229.

Couper, M. (2011). "The future of data collection modes." Public Opinion Quarterly **75**(5): 889-908.

Crowne, D. P. and D. Marlowe (1960). "A new scale of social desirability independent of psychopathology." J Consult Psychol **24**: 349-354.

Crump, R. K., J. V. Hotz, G. W. Imbens and O. A. Mitnik (2006). Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand. Discussion paper series. Bonn, Germany, Institute for the Study of Labor (IZA).

Czajka, J. L. and A. Beyler (2016). Declining response rate in federal surveys: trends and implications, U.S. dDpartment of Health & Human Services**:** 86.

Dailey, R. M. and R. E. Claus (2001). "The Relationship between Interviewer Characteristics and Physical and Sexual Abuse Disclosures among Substance Users: A Multilevel Analysis." Journal of drug issues **31**(4): 867-888.

Davern, M. D., T. McAlpine, J. Beebe, J. Y. Ziegenfuss, T. H. Rockwood and K. T. Call (2010). "Are lower response rates hazardous to your health survey? An analysis of three state health surveys." Health Serv Res **45**(5): 1324-1344.

Davern, M. E. (2013). "Nonresponse Rates are a Problematic Indicator of Nonresponse Bias in Survey Research." Health Serv Res **48**(3): 905-912.

Davis, R. E., M. P. Couper, N. K. Janz, C. H. Caldwell and K. Resnicow (2010). "Interviewer effects in public health surveys." Health Educ Res **25**(1): 14-26.

de Leeuw, E. D. (2005). "To mix or not to mix data collection modes in surveys." Journal of Official Statistics **21**(2): 233-255.

De Leeuw, E. D. (2008). Choosing the method of data collection. International handbook of survey methodology. E. D. De Leeuw, J. J. Hox and D. A. Dillman. New York, Lawrence Earlbaum Associates**:** 117-135.

de Leeuw, E. D., M. callegaro, J. Hox, E. Korendijk and G. Lensvelt-mulders (2007). "The influence of advance letters on reponse in telephone surveys, a meta-analysi." Public Opinion Quarterly **71**(3): 413-444.
Deville, J. and C.-E. Sarndal (1992). "Calibration estimators in survey sampling." Journal of the American Statistical Association **87**(418): 376-382.
Durrant, G. B. (2010). "Effects of interviewer attitudes and behaviors on refusal in household surveys." Public Opinion Quarterly **74**(1): 1-36.
Durrant, G. B. and F. Steele (2009). "Multilevel Modelling Of Refusal And Noncontact In Household Surveys: Evidence From Six Uk Government Surveys." Journal Of The Royal Statistical Society, Series A **172**(2): 361-381.
Fortin, N., T. Lemieux and S. Firpo (2011). Decomposition methods in econometrics. Handbook of Labor Economics. D. Card and O. Ashenfelter. Cambridge, MA, Elsevier North Holland**:** 1-102.
Gautier, A., Ed. (2011). Baromètre santé médecins généralistes [Health barometer among general practitioners]. Baromètre santé. Saint-Denis, INPES.
Groves, R. M. (2006). "Nonresponse Rates And Nonresponse Bias In Household Surveys." Public Opinion Quarterly **70**(5): 646-675.
Groves, R. M. and S. G. Heeringa (2006). "Responsive design for household surveys: Tools for actively controlling survey errors and costs." Journal of the Royal Statistical Society Series A **169**(3): 439-457.
Groves, R. M. and L. E. Lyberg (2010). "Total survey error: past, present and future." Public Opinion Quarterly **74**(5): 849-879.
Groves, R. M. and E. Peytcheva (2008). "The impact of nonresponse rates on nonresponse bias a meta-analysis." Public Opinion Quarterly **72**(2): 167-189.
Groves, R. M., E. Singer and A. Corning (2000). "Leverage-Saliency Theory of Survey Participation: Description and an Illustration." Public Opinion Quarterly **64**(3): 299-308.
Haghighat, R. (2007). "The development of the Brief Social Desirability Scale (BSDS)." Europe's Journal of Psychology **3**(4): 2007.
Huddy, L., J. Billig, J. Bracciodieta, L. Hoeffler, P. J. Moynihan and P. Pugliani (1997). "The Effect of Interviewer Gender on the Survey Response." Political Behavior **19**(3): 197-220.
Imbens, G. W. and D. B. Rubin (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge, MA, Cambridge University Press
Insee (2016). Enquête de victimation - Cadre de vie et sécurité /CVS. Paris, Insee.
Jonas, M. (2016). "7 things I learnt about survey response rates." Retrieved April, 2018, from natcen.ac.uk/blog/7-things-i-learnt-about-survey-response-rates.
Joule, R.-V. and J.-L. Beauvois (2014). Petit traité de manipulation à l'usage des honnêtes gens. Grenoble, France, Presses universitaires de Grenoble.
Kiesler, C. A. (1971). The psychology of commitment. Experiments linking behavior to belief. New York, Academic Press.
Krosnick, J. A. (1991). "Response strategies for coping with the demands or attitude measures in surveys." Applied Cognitive Psychology **5**(3): 213-236.
Krosnick, J. A., S. Narayan and W. R. Smith (1996). Satisficing in surveys: initial evidence. Advances in survey research. M. Braveman and J. Slater. San Francisco, Jossey-Bass**:** 29-44.
Legleye, S., A. Bohet, N. Razafindratsima, N. Bajos, T. Fecond Research and C. Moreau (2014). "A randomized trial of survey participation in a national random sample of general practitioners and gynecologists in France." Rev Epidemiol Sante Publique **62**(4): 249-255.
Legleye, S., S. Pennec, A. Monnier, A. Stéphan, N. Brouard, J. Bilsen and J. Cohen (2016). "Surveying end-of-life medical decisions in France: evaluation of an innovative mixed-mode data collection strategy." Interactive journal of medical research **5**(1): 8.
Morgan, S. L. and C. Winship (2015). Counterfactuals and Causal Inference: Methods and Principles for Social Research, Cambridge University Press.

Nenkov, G. Y., M. Maurin, A. Ward, B. Schwartz and J. Hulland (2008). "A short form of the Maximization Scale: Factor structure, reliability and validity studies." Judgment and Decision Making **3**(5): 371-388.
Nopo, H. (2004). Matching as a tool to decompose wage gaps. Discussion paper series. Bonn, IZA.
Olson, K., J. Smyth and H. M. Wood (2012). Does giving people their preferred survey mode actually increase survey participation rates? An Experimental Examination. Lincoln, University of Nebraska.
Park, S., J. K. Kim and S. Park (2016). "An imputation approach for handling mixed-mode surveys." The annals of applied statistics **10**(2): 1063-1085.
Paulhus, D. (1998). "Interpersonal and intrapsychic adaptativeness of trait self-enhancement: a mixed blessing?" Journal of Personality and Social Psychology **74**(5): 1197-1208.
Pew Research Center (2017). What low response rates mean for telephone surveys.
Razafindranovona, T. (2016). Exploitation de l'enquête expérimentale Vols, violence et sécurité. Document de travail. Paris, Insee. **M2016/03:** 48.
Rosenbaum, P. R. and D. B. Rubin (1983). "The central role of propensity scores in observational studies for causal effects." Biometrika **70**(1): 45-55.
Rubin, D. B. (1986). "Which Ifs Have Causal Answers (Comment on 'Statistics and Causal Inference' by Paul W. Holland) " Journal of the American Statistical Association **81**(961-962).
Rubin, D. B. (2007). "The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials." Statistics in Medicine **26**: 20-36.
Schouten, B., F. Cobben and J. Bethlehem (2009). "Indicators for the representativeness of survey response." Survey methodology **n°35**(Juin 2009): pp. 101-113.
Schwartz, B., A. Ward, J. Monterosso, S. Lyubomirsky, K. White and D. R. Lehman (2002). "Maximizing Versus Satisficing: Happiness Is a Matter of Choice." Journal of Personality and Social Psychology **83**(5): 1178-1197.
Simon, H. (1957). Models of Man: social and rational. Mathematical essays on rational behavior in a social setting. New York, WIley.
Skalland, B. (2011). "An alternative to the response rate for measuring a survey's realization rate of the target population." Public Opinion Quarterly **75**(1): 89-98.
Smyth, J., K. Olson and A. Kasabian (2014). "The Effect of Answering in a Preferred Versus a Non-Preferred Survey Mode on Measurement

" Survey Research Methods **8**(3): 137-152.
Suzer-Gurtekin, Z. T. (2013). Investigating the Bias Properties of Alternative Statistical Inference Methods in Mixed-Mode Surveys. Ph.D. thesis, University of Michigan.
Tournois, J., F. Mesnil and J. Kop (2000). "Autoduperie et hétéroduperie: un instrument de mesure de la désirabilité sociale [self-deception and other-deception: a social desirability questionnaire]." European review of applied psychologie **50**(1): 219-233.
Vannieuwenhuyze, J. and G. Loosveldt (2013). "Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to Disentangle Selection and Measurement Effects." Sociological Methods and Research **42**(1).
von der Lippe, E., P. Schmich and C. Lange (2011). "Advance letters as a way of reducing non-response in a National Health Telephone Survey: DIfference between listed and unlisted numbers." Survey Research Methods **5**(3): 103-116.
Wagner, J. (2008). Adaptive survey design to reduce nonresponse bias, Ann Arbor.
Wagner, J. and M. P. Couper (2011). Using Paradata and responsive design to manage survey nonresponse. World satistics congress of the International statistical institute.
Wilson, S. R., N. L. Brown, C. Mejia and P. W. Lavori (2002). "Effects of Interviewer Characteristics on Reported Sexual Behavior of California Latino Couples." Hispanic Journal of Behavioral Sciences **24**(1): 38-62