Corriger ou pas les effets de mesure ?

Journée d'études LifeObs 2025

24 novembre, 14h -17h30

Centre des colloques, Salle 100

Compte-rendu

Cette demi-journée visait à définir et échanger sur l'impact des effets de mesure lors de la réalisation d'une enquête multimode, et sur la nécessité ou non de les corriger. Elle a permis de confronter les besoins des utilisateurs (obtenir l'estimation la plus précise possible, comparer différentes enquêtes ou vagues d'une enquête, différentes sous-populations) et les méthodes qui peuvent être envisagées pour les corriger (regroupement de modalités, utilisation d'un seul mode privilégié, calage, imputation…).

69 personnes ont participé à l'évènement, dont 28 ont souhaité rejoindre un **groupe de travail,** en cours de constitution, sur le multimode.

L'après-midi était structuré en deux sessions, avec le programme suivant :

Introduction : Albane Gourdol (Ined) et Thomas Merly-Alpa (Insee).

1ère partie : Que sont les effets de mesure ? [Modérateur : Laurent Toulemon – Ined]

2nde partie : s'il y en a, faut-il et comment les corriger ? [Modérateur : Thomas Deroyon – SSMSI]

Conclusion de la journée et ouverture (Stéphane Legleye – ENSAI)

Ce programme a été élaboré par un **comité scientifique** composé de :

- Thomas Merly-Alpa, Responsable du développement de la collecte internet et multimode dans les enquêtes ménages, Insee, responsable du département Innovations de LifeObs
- Hélène Chaput, Directrice du CepiDc, Inserm
- Albane Gourdol, Cheffe du service des enquêtes et sondages, Ined, responsable du département collecte de LifeObs
- Guillaume Carette, Statistique d'enquête, sondage, redressement, analyse de la qualité, Ined

L'ensemble des supports de présentation sont à retrouver sur le site web LifeObs, et le présent compte-rendu synthétise les échanges ayant suivi chaque présentation.

Introduction [Albane Gourdol (Ined) et Thomas Merly-Alpa (Insee)]

Albane Gourdol et Thomas Merly-Alpa ont reprécisé le contexte de cette journée, qui s'est inscrite dans le cadre des activités de l'Observatoire des parcours de vie LifeObs[1]. Lauréat du programme « Équipements structurants pour la recherche » du PIA3 lancé en 2020, ce projet vise à développer un programme d'enquêtes longitudinales et innovantes sur les comportements familiaux, à accroître la diffusion des données, et à former les utilisateurs. Sept enquêtes sont prévues dans ce cadre, qui couvrent toutes les étapes du cycle de vie, de l'enfance à la vieillesse :

- Enfance : première cohorte d'enfants européens Guide/Eurocohorte
- Jeunes adultes : enquête ENVIE sur la vie affective des jeunes adultes (menée en 2023)
- Vie active avec quatre enquêtes :
    - Enquête Familles Employeurs (FamEmp) centrée sur l'évolution du niveau des conflits entre le travail et la vie personnelle (trois vagues de collecte : 2024, 2027, 2030).
    - L'étude des Relations Familiales et Intergénérationnelles (Erfi2) qui vise à décrire les situations familiales dans toute leur diversité (même calendrier que FamEmp).
    - Parfécomsa : enquête sur les parcours de fécondité et de santé reproductive en Outre-mer (2028).
    - Familles : menée par l'Insee et adossée au recensement (2025).
- Vieillissement : l'enquête SHARE sur les thématiques de la santé, des soins, de l'emploi et de la retraite, de la situation socio-économique et financière, des relations sociales et familiales, ou des conditions de vie et de logement envisagées sous le prisme de la dynamique de vieillissement. Les collectes sont menées tous les deux ans, avec trois vagues prévues dans le cadre de LifeObs.

Les activités du projet sont structurées en 4 départements : Collecte, Innovation, Diffusion et Formation. Elles reposent sur une coopération nationale entre des institutions clés dans le domaine des études sur les parcours de vie, la famille et la population : l'Ined (porteur du projet), les universités Paris Dauphine – PSL, de Bordeaux et de Strasbourg, l'Insee et la très grande infrastructure de recherche Progedo.

Cette journée d'étude s'inscrit dans les travaux des départements Innovation et Collecte. Afin d'approfondir le sujet, les personnes qui le souhaitent sont invitées à rejoindre le groupe de travail déjà mentionné sur le multimode.

Session 1 : que sont les effets de mesure ? [Modérateur : Laurent Toulemon – Ined]

Cette session a fait l'objet de trois présentations :

- Cadrage rapide sur ce que sont les effets de mesure (Guillaume Carette – Ined)
- Quelles bonnes pratiques pour les éviter ?
    - Bonnes pratiques de conception de questionnaire (Christine Fluxa – Insee)
    - Les effets enquêteurs dans l'enquête ERFI2 - Étude des relations familiales et intergénérationnelles (Linh Nguyen – Ined)

Ces présentations ont été suivies d'une session d'échange avec la salle sur les points suivants :

[1] LifeObs bénéficie d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du plan France 2030 portant la référence ANR-21-ESRE-0037.

Lorsque l'on veut identifier des effets de mesure par appariement, on utilise généralement des variables socio-démographiques mais on peut aussi utiliser d'autres variables du questionnaire. Comment choisir quelles variables utiliser ?

- ⇨ L'usage est de ne pas mettre de variables de questionnaire, qui sont à risque d'avoir des effets de mesure, et de se limiter à celles de la base de sondage. Cependant, les variables du questionnaire peuvent être cruciales (par ex : statut dans l'emploi), il faut alors essayer de les intégrer dans un second temps au modèle d'appariement pour améliorer le calcul du score.

L'effet de mode peut être dû à un effet de passation ou d'ergonomie. Au téléphone les options « refus » et « ne sait pas » (NSP) ne sont pas nécessairement présentées par exemple.

- ⇨ Effectivement, il n'est pas répété à chaque question dans ERFI 2 que les répondant.es peuvent répondre NSP ou ne pas répondre, même si cela a été dit au début.

Est-ce que les individus choisissent aléatoirement leur mode de collecte ?

- ⇨ Lors de la conception du protocole, il est possible de créer des échantillons dans lesquels le mode de collecte est imposé, afin de comparer les réponses dans une même enquête. Si ce n'est pas le cas, les travaux autour de l'analyse de sensibilité permettent de répliquer cette analyse en supposant que deux personnes qui se ressemblent n'ont pas nécessairement la même probabilité de répondre sur l'un ou l'autre support. L'analyse de sensibilité permet de faire varier cette hypothèse et voir si les résultats tiennent toujours.

Les variables soumises à des effets de mesure sont souvent des variables pour lesquelles les modalités de réponse sont une échelle (Likert, etc.). Ce n'est pas systématique mais cela peut indiquer un risque spécifique lié à ce type de questions.

Les bonnes pratiques de conception d'un questionnaire donnent un cadre théorique, mais il faut aussi garder en tête que l'adaptation de ces règles voire leur transgression est parfois nécessaire pour se conformer à la réalité. Par exemple, il peut être utile de commencer par une question ayant de très nombreuses modalités avant un filtre, pour ne pas exclure des individus qui ne se seraient pas reconnus dans l'activité (« faites-vous du sport » versus « faites-vous de la course, de la marche rapide, etc. »).

Il a été également souligné l'importance de découper les questions complexes en plusieurs questions. Il est souvent craint que cela rallonge le temps du questionnaire, mais cela permet en pratique de simplifier le questionnaire, et de gagner du temps et un effort de réflexion pour les enquêté.es.

Session 2 : faut-il et comment les corriger ? [Modérateur : Thomas Deroyon – SSMSI]

Cette session a fait l'objet de trois présentations :

- ° Introduction sur les méthodes existantes pour les corriger (Thomas Merly-Alpa – Insee)
- ° Présentation des travaux sur ERFI2 - Étude des relations familiales et intergénérationnelles (Guillaume Carette – Ined)
- ° Présentation des travaux sur le Baromètre santé (Noémie Soullier – Santé publique France)

Les échanges avec la salle ont porté sur les éléments suivants :

La notion d'imputation « déontologiquement contestable » a été rediscutée :

- ⇨ Pour les intervenants, imputer revient à changer la réponse d'une personne, ce qui invalide son choix, mais dans les faits il s'agit de faire comme de la non-réponse partielle alors qu'en fait cela n'en est pas puisque la personne a répondu. Cependant, comme l'analyse vise à obtenir des estimations sur une population plus large, l'imputation n'est pas moins éthique qu'une pondération.

Est-il pertinent en CAWI de récupérer des paradonnées sur les temps de réponse pour repérer des dérives ou mesurer la qualité du remplissage ?

- ⇨ Cela peut être assez riche. Ce n'est pas encore exploité sur ERFI2 mais il y a bien les données nécessaires pour que cela puisse se faire. En particulier pour les questions sur les connaissances du baromètre santé, SpF va mettre en place des mesures du temps de réponse pour les questions de connaissance. Si le temps de réponse est trop long on peut penser que les personnes sont allées chercher les réponses sur internet.

Est-il intéressant de creuser l'autoévaluation de la personne vis à vis de ses réponses ?

- ⇨ Ce n'est pas traité dans le cadre d'ERFI2 pour l'instant. Cela peut être une piste pour le baromètre mais il y a un risque que cette auto-évaluation soit elle-même entachée d'effets de mesure.

Existe-t-il des effets de mode entre portable, pc et tablettes ?

- ⇨ Une stagiaire de Santé publique France a travaillé sur ce sujet, il n'y a pas eu de différence sur la qualité des réponses entre PC et smartphone.

Avec les méthodes déterministes, quand s'arrête-t-on de les corriger ?

- ⇨ L'idée est de constituer une hypothèse selon laquelle on aurait une modalité sur ou sous-estimée dans un mode de collecte alternatif. A partir de là, l'approche déterministe consiste à corriger séquentiellement les observations les moins bien prédites selon un modèle dans le mode de référence jusqu'à équilibre des estimations entre les deux modes.

Corriger les effets de mode implique une augmentation du temps de traitement des données. Afin de le limiter, ne serait-il pas possible de faire une grande refonte, qui acte les changements et permette de repartir sur une nouvelle base ?

- ⇨ Les équilibres ne sont malheureusement pas stables (par exemple la part des répondants par téléphone continue de baisser). Il est donc trop tôt pour pouvoir acter de nouvelles bases pour les analyses.

Comment informer les utilisateurs des effets de mesure pour leur permettre d'utiliser correctement les données ?

- ⇨ A ce stade, on ne peut qu'alerter les chercheurs sur de potentiels effets de mesure. Il faut diffuser des consignes de bonne utilisation des données (ex : ne jamais utiliser une modalité seule). Cela ne répond cependant pas au problème de la diffusion de chiffres par le producteur pour des variables qui ont des effets de mesure.

Conclusion de la journée et ouverture (Stéphane Legleye – ENSAI)

La question des effets de mesure nous pousse aujourd'hui à revenir aux fondamentaux et ainsi redéfinir ce qu'est une enquête, ce qu'on mesure et quel biais on introduit. Cela implique de repenser le travail de collecte et de prévenir les biais dès la conception du questionnaire.

On n'a jamais la garantie de ce qu'on mesure, même avec le mode qui semble le plus sûr. On reporte actuellement le travail d'analyse de biais sur les statisticiens. Or, on peut déployer un arsenal de méthodes statistiques, mais on peut se demander s'il ne serait pas plus pertinent de corriger les biais de mesure à la source : une piste serait ainsi de poser des questions différentes entre les modes pour obtenir in fine la même mesure (ce qui est actuellement contraire aux bonnes pratiques, qui visent plutôt un questionnaire omnimode), via la modification des formulations des questions, échelles de réponse différentes, présentations visuelles différentes…

On peut également jouer sur l'introduction de la question : rappeler l'importance de la sincérité dans les réponses aux enquêtes, les niveaux d'usages, les conceptions répandues dans la population, les résultats d'enquête de référence, la perception des comportements à l'échelle de la société... Tout ceci peut permettre de « dédramatiser » les questions, renforcer la véracité des réponses ou permettre de mieux identifier les biais. Des changements dans la présentation visuelle ou l'ancrage de la question pourraient avoir un effet important.

Il est donc important de travailler sur la mise en condition des répondant.es. On pourrait en ce sens imaginer de travailler avec des laboratoires de psychologie sociale en menant des expérimentations.

Une fois les questions posées, il serait aussi possible de mettre en place une batterie de tests pour repérer les biais et les corriger. On peut se demander si cela est automatisable mais cela s'avère a priori compliqué, car d'une part il semblerait que les corrélations entre ces tests et le comportement de réponse soient faibles, et d'autre part, ces questions sont elles-mêmes à risque d'être entachées d'effets de mesure.

Comme il y a toujours des enjeux de mesure, il y a toujours un risque d'effets de mesure. Il faudrait ainsi se mettre d'accord sur une liste de thématiques ayant donné lieu à des effets de mesure importants et commencer à créer une liste de références. Cela permettrait de se concentrer sur ces questions et de ne pas perdre du temps à chercher des effets de mesure sur des sujets peu à risque, sauf à des fins de confirmation.

Il reste néanmoins toujours compliqué de déterminer si on a réussi à corriger un biais de mesure, car cela implique de savoir si on a des populations comparables. Un des enjeux est ainsi de savoir qui porte le biais de mesure et s'il est homogène entre les catégories de populations : quel est le bon agrégat ? Cela demanderait de multiplier les analyses et le travail, sans savoir comment s'arrêter. Faut-il chercher à corriger tous les biais jusqu'à qu'ils soient absents ou indiscernables ? Quel est réellement l'impact sur la décision publique liée à l'enquête s'il reste un petit effet de mesure ?

Cette conclusion a donné lieu à de nouveaux échanges avec la salle :

Est-ce qu'on peut faire fi de la différence entre biais de mesure et biais de sélection ?

- ⇨ Il n'est pas possible de savoir si on corrige entièrement un biais de mesure. Dès lors, est-il pertinent de tout de même chercher à le corriger ? On a tendance à penser qu'on corrige des

biais de mesure uniformes alors que la plupart du temps les biais sont variables par sous-groupe. Dans ce cas, on ne peut pas savoir si on observe un biais de composition ou un effet de mesure inhomogène entre groupes. Est-ce qu'il faut aller vers quelque chose de plus fin avec le "machine learning" ? Ou renoncer au contraire à corriger tous les biais ?

Comment corriger les effets de mesure sur les "batteries de questions" ?

- ⇨ La correction est d'autant plus difficile quand les questions ne sont pas binaires. On ne peut pas procéder comme s'il n'y avait pas de lien entre des questions cohérentes. Cela implique une modélisation et des traitements statistiques particuliers. Il faut imputer l'ensemble des réponses à cette batterie de façon conjointe.

Est-il vraiment important de corriger un biais ? ou peut-on se contenter d'annoncer une marge d'erreur ?

- ⇨ Il suffirait peut-être d'avoir des recommandations pour savoir quoi faire selon le type de situation, et à partir de quel moment la marge d'erreur liée aux effets de mesures est problématique. Il est donc nécessaire de mettre en commun les travaux qui sont faits sur ce sujet, en ne délaissant pas les questionnements de base.

Que faire en cas de biais de sélection massif ?

- ⇨ L'exemple d'Epicov nous a montré que cela peut arriver. Dans ce cas, il faut accepter que nous sommes dans cette situation, et conduire les travaux adaptés (méthodes économétriques telles que Heckman) pour en déduire la meilleure estimation possible. Cela doit rester des cas exceptionnels : il faut travailler à une présentation différente de l'enquête et trouver des données externes pour corriger du biais de non-réponse.

Est-ce qu'il y a des cas où l'effet de mode peut être désirable ?

- ⇨ Cela dépend des sous-groupes et des cas ! Par exemple, il peut y avoir des effets de scénarisation chez les adolescent.es pour qui les réponses exagérées en présence de leurs pairs peuvent donner du crédit social. Il faut avant tout capitaliser sur les expériences pour construire des éléments de référence.

Comment mesurer la préférence sur les modes, notamment pour les enquêtes répétées ?

- ⇨ Ce serait une piste de travail intéressante pour mieux comprendre ce qui motive les enquêtés. Mais la formulation des questions est à travailler en se mettant à leur place. En ce qui concerne les enquêtes répétées, on observe dans l'enquête Emploi que la majorité des individus conservent le même mode de réponse pour les cinq réinterrogations, cela semble prouver qu'il existerait des préférences.

En conclusion, il est trop tôt pour savoir s'il est pertinent ou non de corriger les effets de mesure. Le sujet va être approfondi au sein d'un groupe de travail dédié.