NOTE SUR LA QUALITÉ DES SOURCES DE RECRUTEMENT

1. CONTEXTE

Aujourd'hui, les enquêtes de référence de mesure d'audience sont réalisées par téléphone ou par internet. Néanmoins, nous avons souhaité dans cette note élargir notre analyse à l'ensemble des sources de recrutement qui permettent d'entrer en contact avec un individu/un foyer pour réaliser une enquête.

La diversification des sources de recrutement, c'est-à-dire le recours à d'autres sources que les bases de numéros de téléphone issues de l'annuaire ou générées aléatoirement, concerne aujourd'hui la plupart des études d'audience que le CESP contrôle. La baisse de la joignabilité pour les enquêtes par téléphone (avec une accélération depuis la mise en place des numéros polyvalents par l'ARCEP au 1er janvier 2023), la difficulté croissante à joindre certaines cibles (les jeunes et les CSP- par exemple) et le renchérissement des coûts que ces phénomènes entraînent, expliquent le recours à de nouvelles sources de recrutement par les instituts.

Ces nouvelles sources ont un point commun : elles sont le plus souvent qualifiées (par l'institut ou par un prestataire) et permettent de cibler des individus sur des critères socio-démographiques. Le recours à ces sources ne permet plus un recrutement en aléatoire ou sur des bases exhaustives.

Dans cette note, le CESP a souhaité :

- Lister les bonnes pratiques lorsque l'on introduit une nouvelle source de recrutement, en particulier quand il s'agit de recours à des bases qualifiées (sous-populations caractérisées par un ensemble de critères socio-démographiques ou autres).
- Donner une liste de critères permettant de valider la qualité de la source de recrutement, notamment sur le critère de la représentativité.
- Préciser le niveau de risque et donner son avis sur l'opportunité d'utiliser chacune de ces sources pour ce qui concerne les études auditées par le CESP.

Le recours à des sources qualifiées est un sujet qui concerne particulièrement le CESP car il est lié à la qualité de l'échantillon recruté et interrogé. Les études d'audience ont pour objectif principal de fournir au marché publicitaire des indicateurs sur lesquels s'appuient l'achat et la vente des espaces publicitaires. Pour garantir une monnaie fiable et des échanges justes, il est nécessaire que la mesure d'audience ne soit pas biaisée, c'est-à-dire qu'elle soit représentative des comportements et des habitudes de consommation des médias de la population étudiée.

En introduisant de nouvelles sources de recrutement, on peut améliorer la structure des échantillons mais on doit aussi s'assurer que cela n'introduit pas de nouveaux biais qui auraient un impact sur les audiences.

2. DÉFINITIONS

Il existe plusieurs sources de recrutement qui sont utilisées pour les enquêtes en général et qui peuvent, pour certaines, être utilisées dans les études d'audience de référence.

Dans cette partie, on tente de définir les différentes sources de recrutement, selon un certain nombre de critères. Les sources de recrutement se distinguent des modes de recrutement (face à face, CATI, CAWI…).

1. Bases exhaustives (ou quasi exhaustives) portant sur la population entière ou sur une sous-population

Ces bases permettent de constituer des échantillons probabilistes par tirage aléatoire.

- Base de numéros de téléphone (par déclinaison ou par génération aléatoire), base de coordonnées géographiques permettant de tirer des points d'enquêtes
- Bases produites par la Statistique publique : fichier Fidéli, fichier SIREN, base électorale… Le CESP précise que ces bases sont pour certaines très difficiles d'accès
- Mais aussi des bases « privées » comme des bases clients, des bases abonnés qui permettent d'interroger une population d'intérêt.

La question de la qualité de ces sources ne se pose généralement pas, en revanche les difficultés liées à la joignabilité peuvent altérer la représentativité des échantillons issus de ces sources. Dans ce cadre, le CESP recommande d'ailleurs de ne plus inclure dans les bases de téléphone les numéros en 09.

2. Panels ou access panels

Toutes ces sources ont un point commun : dès leur entrée dans le panel / l'access panel, les panélistes s'engagent à coopérer pour de futures collectes de données s'ils sont sélectionnés, généralement en échange d'une gratification. C'est une base de « volontaires » qui ont accepté d'être sollicités régulièrement à participer à des enquêtes. Il y a une relation qui s'inscrit dans le temps entre les membres du panel / de l'access panel - qui ont accepté l'engagement qui leur est demandé - et le prestataire en charge de cette base de sondages.

Plusieurs distinctions peuvent être faites :

- Entre panel et access panel : un panel est constitué de manière à représenter une population étudiée et pourra être réinterrogée dans le temps (analyses longitudinales), tandis qu'un access panel est un vivier plus large, dont la structure diffère a priori (plus ou moins) de celle de la population générale, mais permettant de tirer des échantillons qui seront sollicités pour des enquêtes
- Entre recrutement probabiliste *vs* non probabiliste : les futurs panélistes sont recrutés de manière aléatoire pour un panel probabiliste *vs* au travers de multiples sources dont il faut s'assurer de la qualité, pour les recrutements non probabilistes
- Entre le fait d'être géré par l'institut qui a la charge de l'enquête *vs* un prestataire extérieur : dans le premier cas, il est plus aisé d'obtenir des informations précises sur le panel/l'access panel
- Entre différents modes de sollicitation : une sollicitation pour une seule enquête ou - et c'est le cas pour de nombreux access panels - le choix laissé au panéliste entre plusieurs sujets, ce qui peut générer un biais d'auto-sélection.

Enfin, on doit ajouter, que des panélistes peuvent participer à plusieurs panels ou access panels.

3. Viviers d'adresses

Il en existe de plusieurs sortes. Il n'y a pas de gratification proposée pour être inclus dans un vivier.

- **Vivier d'adresses constitué par les instituts** : il s'agit d'une base d'individus ayant été interrogés sur un dispositif et ayant accepté de répondre à d'autres enquêtes mises en œuvre par l'institut. Ces bases appartiennent à l'institut et sont gérées en interne. Il n'y a pas de système de gratification proposé

  Cas particulier de la ré-interrogation sur une même étude : l'institut peut se donner la possibilité de réinterroger des individus ayant déjà participé à l'enquête

- **Fournisseurs de fichiers** : le prestataire est un agrégateur de fichiers ou de bases de contacts. Les individus listés dans ces fichiers n'ont pas de lien avec le fournisseur d'adresses (ils n'ont pas conscience de faire partie d'un fichier) et ils ne bénéficient pas de gratification pour leur participation à une enquête.

  Nous distinguons dans cette note deux types de fichiers :

  - Les fichiers « annuaires » qui couvrent une partie importante des numéros de téléphone actifs
  - Et les fichiers qualifiés qui permettent de cibler des profils spécifiques (jeunes, individus ou foyers hauts revenus, professions particulières…).

4. Autres sources de recrutement

- **« Intercepts » ou échantillonnage sur le terrain** : il existe une autre catégorie de sources de recrutement qu'ESOMAR désigne comme « Intercepts » (*ESOMAR Global Research – March 2021 – Questions to help buyers of online samples*). L'interception est une approche où les participants potentiels sont invités à répondre à une enquête (pour une gratification ou non) pendant qu'ils sont engagés dans une autre activité :
  - **À l'extérieur ou dans un espace ouvert au public (**centre commercial, gare, aéroport…). Dans ce cas, les échantillons peuvent être probabilistes si la population recherchée est la population qui fréquente ces espaces et si l'institut met en œuvre un mode de sélection assimilable à un tirage aléatoire (par exemple clients d'un magasin ou utilisateur d'un moyen de transport…)
  - **Ou en ligne** (jeu en ligne, lecture de contenus, actif sur un réseau social ou toute autre activité en ligne). Les participants « interceptés » peuvent être inconnus du fournisseur de l'échantillon ou avoir été pré-identifiés et qualifiés lors d'une d'enquête antérieure. Les modes de recrutements en ligne sont réalisés au travers de publicités, ciblages précis…
- **Parrainage** : un individu recruté pour participer à une enquête est sollicité (et gratifié) pour fournir des contacts de personnes de son entourage susceptibles, elles aussi, de participer à l'enquête (entourage : au sein du foyer ou en dehors).

5. Répondants synthétiques

On peut définir les répondants synthétiques comme des individus générés artificiellement qui ne correspondent pas directement à des individus ou à des événements réels. Ils imitent les propriétés statistiques du monde réel. Ce type de données doit être mentionné dans cette note, compte tenu des développements très rapides de l'IA et de la modélisation dans l'écosystème des sondages : on peut aujourd'hui avoir des échantillons constitués pour tout ou partie de répondants synthétiques.

Par exemple :

- En augmentant ou en complétant un échantillon/un panel par la création de répondants synthétiques, le plus souvent à partir des répondants réels de l'échantillon/du panel
- En générant un échantillon/un panel de répondants synthétiques, ou en utilisant une population synthétique pour les interroger.

On peut considérer ces développements comme de nouvelles sources pour constituer des échantillons et des panels. Dans ce cas, évaluer la qualité de ces sources sera primordial mais complexe.

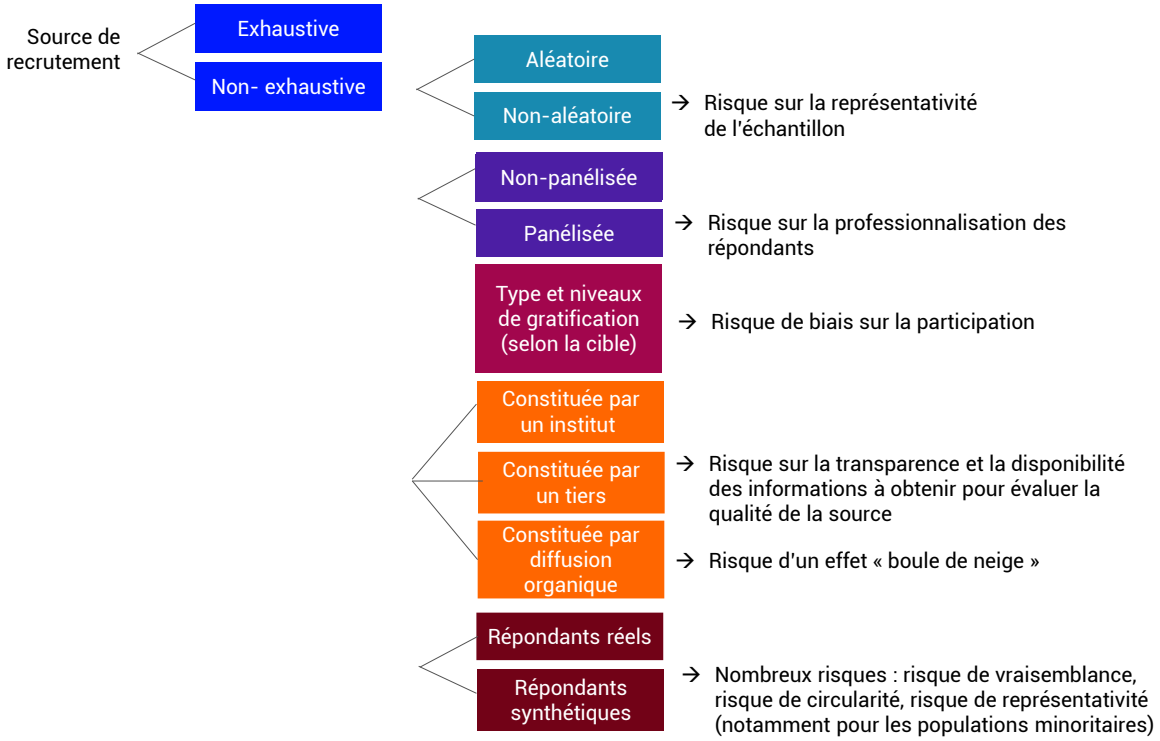
On ne mentionne pas dans cette note les autres utilisations qui peuvent être faites des données synthétiques (imputations, utilisation de l'IA pour remplacer les enquêteurs…), car cela sort du thème abordé.

Illustration 1. Panorama des différentes sources de recrutement

Bases exhaustives ou quasi exhaustives	Panels et access panels	Viviers d'adresses et fournisseurs de fichier	Autres sources	Répondants Synthétiques
• Permettent de constituer des échantillons probabilistes • Par exemple : base de numéros de téléphone, bases produites par la Statistique publique (Fidéli, SIREN, base électorale…), fichiers clients (abonnés…)	Un point commun : des individus "volontaires" pour être sollicités régulièrement à participer à des enquêtes (avec un système de gratification) Plusieurs distinctions peuvent être faites : • Panel *vs* access panel • Recrutement probabiliste *vs* non-probabiliste • Géré par l'institut en charge de l'enquête *vs* prestataire extérieur • Mode de sollicitation	On distingue : • Vivier d'adresses constitué par l'institut avec le cas particulier de la ré-interrogation • Fournisseurs de fichiers avec 2 types : fichiers « annuaires » et fichiers qualifiés pour cibler des profils spécifiques (jeunes…) • En général, pas de gratification proposée pour être inclus dans un vivier	Intercept (échantillonnage sur le lieu de présence de l'interviewé) • Intercept à l'extérieur ou dans un espace ouvert • Intercept en ligne (jeux en ligne, réseaux sociaux etc.) • Les modes de recrutement en ligne peuvent prendre plusieurs formes : pré-identification par un prestataire, publicité… (effet des algorithmes)	• Complément d'échantillons avec des répondants synthétiques • Recours à des populations synthétiques pour recruter un échantillon, un panel

Pour résumer, voici les questions à se poser pour identifier la nature de la source utilisée et évaluer le risque lié à la représentativité :

Illustration 2. Les questions à se poser et les risques liés



3. BONNES PRATIQUES ET LISTE DES CRITÈRES

3.1. Bonnes pratiques

Le CESP a listé les bonnes pratiques à suivre à chaque étape de l'intégration d'une nouvelle source de recrutement :

Au moment de la sélection de ces sources, en recueillant les informations nécessaires à l'évaluation de la qualité auprès du prestataire qui fournit les contacts :

- Définir de quels types de source il s'agit : vivier, panel, access panel, fournisseur de fichiers… pour comprendre le lien entre le prestataire et les individus susceptibles d'être contactés
- Obtenir des informations sur les sources qui ont permis la constitution du vivier, du panel, de l'access panel, du fichier d'adresses : l'objectif est de pouvoir déterminer s'il n'existe pas un défaut important de représentativité ou si une des sources de recrutement n'a pas un lien avec le sujet de l'enquête
- Demander les thématiques des études réalisées sur la source pour identifier s'il y a un lien entre le sujet de l'étude et les autres sujets traités le plus souvent par le prestataire
- Connaître la méthode d'échantillonnage utilisée par l'institut ou le prestataire pour élaborer l'échantillon qui sera fourni et les règles de sélection (nombre maximum de sollicitations autorisé, sur quelle période…)

- Obtenir des informations pour pouvoir évaluer la méthode de gestion des contacts et leur exploitation. Quelles sont les règles de gestion : sélection des adresses et intégration dans les bases, recrutements des panélistes, nombre de sélections, de sollicitations / de participations, règles de nettoyage, système de relances et programme de gratification… ? L'objectif est de s'assurer que les contacts ne sont pas sur-sollicités, « experts » des sondages, des questions marketing, de l'audience… ou « chasseurs de prime », ce qui pourrait remettre en cause la qualité de leurs réponses
- Enfin, décrire la méthode de qualification des profils et donner la récence et le rythme de mise à jour de ces informations
- Demander à disposer de variables auxiliaires pour évaluer la qualité de l'échantillon. Choisir des variables en lien avec la variable d'intérêt et pour lesquelles on dispose des statistiques officielles. Si l'information n'est pas disponible, intégrer ces variables dans le questionnaire d'enquête (mais pas en quotas)
- Concernant les répondants synthétiques : obtenir des informations sur les données utilisées pour entraîner le modèle et sur la méthodologie employée, et obtenir des indicateurs sur le biais, la précision, la pertinence (liens pauvres entre les variables), le phénomène d'hallucination (génération de données trompeuses).

En réalisant des analyses au moment du test ou de l'intégration de la source dans le dispositif d'enquête :

- Mener des tests en amont, avant d'introduire la nouvelle source dans les bases de recrutement ou d'enquêtes, notamment en utilisant des variables auxiliaires pour vérifier la représentativité de l'échantillon
- S'assurer de la déduplication de ces sources avec les bases d'enquêtes utilisées pour éviter les sollicitations multiples
- Garder, dans les fichiers, l'information « source de recrutement » pour pouvoir mener des analyses a posteriori (par exemple dans le recrutement d'un panel : y a-t-il un taux de churn différent selon la source ?)
- Mettre en place une validation des profils pour évaluer la qualité des sources sur ce critère (les jeunes sont-ils bien des jeunes ?)
- Mener des analyses pour estimer les impacts éventuels sur les niveaux de participation mais aussi sur les niveaux d'audience et partager ces analyses avec les commanditaires de l'enquête
- Intégrer une question sur la participation à des panels, la fréquence de participation à des enquêtes
- Pour les résultats issus des répondants synthétiques : les valider en les comparant à des résultats mesurés sur un échantillon réel.

Dans l'utilisation de ces sources dans la durée :

- Limiter le poids des sources qualifiées (pour privilégier les sources aléatoires)
- Maintenir stable le poids de chaque source dans le temps sauf s'il est démontré qu'il n'y a pas d'effet sur les résultats en changeant le poids des sources
- Garder une structure de l'échantillon proche d'une année sur l'autre pour limiter les impacts sur les niveaux d'audience liés à ces changements.

Cas particulier des ré-interrogations

Pour éviter le risque de réponses de mauvaise qualité dû à la connaissance du questionnaire :

- Limiter le recours à cette méthode car les individus qui acceptent la ré-interrogation sont probablement plus intéressés par le sujet (biais de réponse)

- Établir des règles strictes sur le nombre de participations autorisées et la durée minimum entre les ré interrogations et se donner les moyens de contrôler la bonne participation de ces individus (durée de réponse au questionnaire, mise en place d'une durée minimum acceptable…).

3.2. Liste des informations à obtenir/à produire pour évaluer la qualité

Dans cette partie, le CESP a élaboré une première liste de critères sur lesquels les bases peuvent être évaluées (certains critères ne s'appliquent pas à toutes les sources). La transparence est un élément important de l'évaluation. Il n'y a pas d'indicateurs chiffrés pour les critères listés, car les analyses par source doivent être faites de manière ad hoc en prenant en compte l'ensemble des informations données.

Critères	Oui
Risque de biais	
La source est-elle exhaustive ? Si non, les éléments fournis par le prestataire permettent-ils de vérifier si la source utilisée vise à la représentativité de l'univers que l'on cherche à représenter ?	
Le risque que la probabilité de présence dans la source soit corrélée au thème de l'enquête a-t-il été évalué (soit validé par le prestataire, soit par l'utilisation de variables auxiliaires) ?	
La taille de la base et la méthode de tirage de l'échantillon permettent-ils de considérer la sélection des individus comme étant aléatoire ?	
Pour la base totale et l'échantillon fourni : structure socio-démographique (sexe, âge, CSP de l'individu et région a minima) et toute autre information permettant de valider que l'univers recherché est couvert par la source	
Gestion des contacts (quand la source est gérée par un tiers)	
Quel est le nombre maximum autorisé de sélections / de sollicitations sur 12 mois, sur un mois ? Ce système permet-il d'éviter les sur sollicitations ?	
Pour l'échantillon fourni : quel est le nombre de participations à des enquêtes sur le même sujet au cours de l'année ? Cela permet-il d'éviter d'avoir des « experts » d'un sujet ou d'un type de questionnaire ?	
(*Pour les panels et access panels)* Le mode de sollicitation : a-t-on des informations sur le choix laissé à l'enquêté sur les sujets d'enquêtes (une seule enquête proposée ou plusieurs) ?	
Les critères pour exclure un contact, pour détecter la fraude sont-ils connus et permettent-ils d'écarter les mauvais participants ?	
(*Pour les panels et access panels)* Le protocole de gratification a-t-il été décrit ? Le montant de l'incentive pour la participation à l'enquête a-t-il été fourni ? Est-il adapté à l'enquête ?	
Combien de relances à la suite de la première sollicitation ? Ce système permet-il de maximiser la participation ?	
Le process et la fréquence de mise à jour des données ont-t-ils été fournis ? Permettent-ils d'avoir des informations de profil fiables ?	
Le taux de réponse moyen observé dans la base au cours des 12 derniers mois a-t-il été fourni ? (*ne s'applique pas aux fournisseurs de fichiers)*	
Qualité du fichier fourni (échantillon sollicité)	
Y a-t-il eu déduplication avec les autres bases d'enquêtes ?	
La cohérence entre profils vendus et profils observés a-t-elle été validée ?	
Les taux de réponse ont-ils été analysés ?	
Impact sur la qualité des réponses et sur les résultats (échantillon sollicité)	
Des analyses sur les durées par bloc de questions ont-elles été menées ?	
Des tests statistiques ou des analyses ont-ils été réalisés au cours des deux dernières années pour identifier s'il existe un lien entre la source et les niveaux d'audience ?	

Enfin, pour les recrutements de plusieurs personnes dans un même foyer, ou pour les recrutements par parrainage ou par diffusion organique, l'effet de grappe a-t-il été mesuré dans l'analyse des résultats ?	
Pour les données synthétiques	
Les informations sur les données d'apprentissage et sur la méthodologie ont-elles été fournies ?	
Une validation avec des données mesurées sur un échantillon réel a-t-elle été réalisée ?	
Les limites du modèle, y compris sous forme d'indicateurs chiffrés (biais, précision…) ont-elles été fournies ?	

4. AVIS DU CESP

Le CESP donne son avis, dans le tableau suivant, quant à l'utilisation dans les études d'audience de référence de chacune des sources de recrutement.

L'évaluation des risques qui portent sur la qualité des sources suit l'échelle suivante :

- Pas de risque : la qualité de la source est reconnue
- Risques faibles : la qualité de la source est reconnue et son utilisation acceptable, mais certains points doivent être validés (voir la liste des critères du CESP)
- Risques avérés : des risques sont reconnus ; l'institut peut envisager d'utiliser ce type de source en s'assurant que les critères de qualité sont respectés. Le CESP recommande que l'utilisation de ce type de source reste limitée
- Risques importants : le CESP attire l'attention sur les risques de perte de représentativité à utiliser cette source. Il recommande de l'utiliser avec précaution, sur des volumes très limités, et pour des cibles difficiles à recruter.

Type de source qualifiée	Indice de risque
Bases exhaustives	**Pas de risque**
Panel probabiliste	**Risques faibles** ; l'attention doit porter sur le risque de professionnalisation des répondants
Panel et access panel non probabiliste	**Risques avérés à risques importants** : les différents risques sont à déterminer en fonction des choix de gestion faits par le prestataire
Vivier d'adresses constitué par les instituts	**Risques faibles** si ces viviers sont constitués à partir d'enquête de bonne qualité (type enquêtes de cadrage) **Risques avérés** pour les autres types de viviers et les ré-interrogations
Fournisseurs de fichiers type annuaire	**Risques faibles à risques avérés** selon la couverture de l'annuaire
Fournisseurs de fichiers sur cibles	**Risques avérés**
Intercepts à l'extérieur ou dans l'espace public	**Risques faibles** si la sélection des interviewés se rapproche d'un mode aléatoire **à avérés** selon la population cible**.** Les risques identifiés sont le biais de sélection et le biais de couverture
Intercepts en ligne	**Risques importants :** le biais de sélection est renforcé par les algorithmes utilisés (effet « boule de neige algorithmique »)
Parrainage	**Risques importants** ; l'institut doit prouver par la réalisation d'analyses des résultats que le parrainage n'introduit pas de biais dans les résultats
Répondants synthétiques	**Risques importants** ; l'institut doit produire des données permettant d'évaluer la qualité des données synthétiques, notamment en les comparant à des données mesurées sur échantillon

Pour les études d'audience, qui permettent de produire des résultats en historique, le CESP rappelle l'importance de maintenir stable le poids de chaque source dans le temps sauf s'il est démontré qu'il n'y a pas d'effet sur les résultats en changeant le poids des sources.

Dans tous les cas, le CESP recommande d'évaluer la qualité des sources introduites dans le plan de sondage ainsi que leur impact sur les résultats et de communiquer aux utilisateurs cette information.