Les nouveaux défis posés par les sondages internet

Présenté par Claire Durand, professeur, département de sociologie, Université de Montréal,

Dans le cadre *Colloque International sur les Sondages & Mesures d'Audience* en Algérie, Alger, 11-13 mars 2011.

Il peut apparaître surprenant que je propose d'aborder les sondages Internet dans un pays où la pénétration d'Internet n'est pas encore très importante (14% en 2010) Toutefois, si certains modes de communication ne se répandent pas aussi rapidement en Afrique du Nord que dans les pays plus développés, il demeure que, dans un avenir rapproché, ils se développeront ici comme ailleurs et ceci d'autant plus que les accès sans fil et sur téléphone cellulaire se développent peut-être plus rapidement et facilement ici que dans un pays très étendu et peu peuplé comme le Canada par exemple.

Début mars 2011, les journaux français font leur première page avec une supposée avance de Marine Le Pen sur Nicolas Sarkozy pour les élections françaises de 2012, ceci sur la foi d'un sondage Internet réalisé par Harris Interactive, le premier sondage électoral de cette firme en France. Une telle situation se produit régulièrement dans plusieurs pays du monde. Au Québec, la seule firme ayant fait des sondages lors de l'élection municipale de Montréal en novembre 2008 a produit des erreurs d'estimation de sept pourcent et n'a prédit l'ordre d'arrivée d'aucun des trois principaux candidats. Cette erreur importante n'a pas que les estimations des sondages de type Panel Internet fassent la Une des journaux par la suite. Or, il faut être d'autant plus prudent que ces estimations peuvent influencer, voire biaiser le vote subséquent. Évidemment, si les estimations des sondages Internet étaient toujours mauvaises, ces sondages auraient disparu. Ce n'est pas le cas. Souvent, l' estimation des intentions de vote se distinguent peu de celles des sondages utilisant d'autres méthodologies.

Pourquoi utiliser un sondage Internet dans des pays dits « avancés » où il y a une couverture presque complète des ménages par téléphone filaire ou cellulaire alors que la couverture par Internet est nettement moins élevée, atteignant difficilement les 70%? La première motivation est sans aucun doute le coût. Une fois absorbés les coûts d'infrastructure, les sondages Internet ne demandent pas d'interviewers comme pour les modes de passation par entrevue ni d'impression et de frais de poste comme pour les sondages postaux. Ils sont donc très économiques. Au-delà toutefois de cette motivation non négligeable, on peut penser que les sondages Internet ont aussi l'avantage de tous les sondages auto-administrés, c'est-à-dire la réduction des réactions de prestige et de conformité sociale qui peuvent survenir en présence d'un interviewer. On a donc tendance à penser que les réponses données dans un sondage Internet sont plus « vraies », plus « fiables » que celles données aux sondages par entrevue. Enfin, la possibilité de présenter des stimuli visuels est aussi un aspect valorisé particulièrement dans les sondages Marketing.

Qu'est-ce qu'un sondage Internet?

Mais qu'est-ce qu'un *sondage Internet*. Le terme est généralement utilisé pour tous les sondages auquel le répondant est invité à répondre sur un site Web. Toutefois, les modes de recrutement sont très variés et correspondent à des objectifs divers. Couper (2000) liste huit modes de recrutement :

Les modes non probabilistes

1. Les sondages faits à des fins de divertissement ou d'animation sur les sites des media où l'on pose des questions quotidiennes.
2. Les sondages recrutant des volontaires au moyen d'annonces où l'on offre habituellement une rémunération faible ou la possibilité de gagner des prix.
3. Les panels de volontaires, très utilisés pour les sondages marketing et électoraux, dont le recrutement se fait surtout par des invitations à joindre le panel placées sur divers sites, y compris ceux des firmes de sondage.

Les modes probabilistes

4. Les sondages de « sortie de site » où l'on interroge au hasard des personnes ayant utilisé un service ou fréquenté un site web.
5. Les sondages faits à partir de listes complètes de population (employés d'une organisation, étudiants d'un collège ou d'une université, etc.).
6. Les sondages multi-mode où on offre aux répondants contactés par un autre mode – téléphone, poste -- la possibilité de répondre sur un site Web.
7. Les panels d'utilisateurs d'Internet recrutés via un autre mode – téléphone ou face à face -- avec sélection probabiliste.
8. Enfin, les échantillons probabilistes de l'ensemble de la population où le recrutement se fait par une autre mode mais où l'on fournit l'équipement nécessaire pour faire le sondage sur le Web (WebTv ou ordinateur) lorsque le répondant n'a pas ce type d'équipement.

Des huit modes de recrutement utilisés, seul le huitième permet théoriquement de mesurer de façon non biaisée l'opinion de l'ensemble de la population et le septième, l'opinion non biaisée des internautes. Les modes 4 et 5 permettent de recueillir l'opinion de populations ciblées restreintes. La plupart des sondages Internet dont nous entendons parler sont faits à partir d'échantillons de volontaires, non probabilistes (méthode 3). Les sondages internet visant la couverture de l'ensemble de la population et utilisant un échantillon probabiliste sont très chers. Devant la possibilité de recruter des répondants « professionnels » et d'avoir par la suite

un accès facile et peu coûteux à ces répondants, la plupart des firmes de sondage, et même de plus en plus de chercheurs, ont recours à des panels à échantillon non probabiliste (méthode 3).

Quels sont les problèmes connus des sondages Internet utilisant des panels non probabilistes?

Outre le problème de base du non recours à un échantillon aléatoire, la première question qui se pose lorsque l'on étudie les panels Internet est celui de la couverture. Jusqu'à quel point l'ensemble de la population a-t-elle accès à Internet? Jusqu'à quel point la population y ayant accès est-elle bien représentée dans les panels? La population ayant accès à Internet est-elle similaire à celle qui n'y a pas accès, toutes choses égales par ailleurs?

Bigot, Croutte et Recours (2010) examinent précisément cette dernière question et ceci, dans le cas de la France. Dans le cadre d'une enquête probabiliste sur les conditions de vie et les aspirations des Français menée début 2009, ils proposent une analyse des différences entre les répondants selon leur accès à Internet. Ils montrent que les personnes ayant accès à Internet se distinguent significativement sur 37% des 191 variables mesurées avant redressement. C'est particulièrement le cas pour les variables ayant trait à la composition des ménages, à leur équipement et à leur logement ainsi que pour les opinions ayant trait aux mœurs (homosexualité, etc.). Après avoir redressé pour amener l'échantillon d'internautes à une composition sociodémographique similaire à celle de l'ensemble des Français, les différences demeurent sur 12% des variables, en particulier celles ayant trait au logement et à l'équipement des ménages, aux opinions sur les mœurs (plus libérales chez les Internautes) et aux pratiques culturelles. Notons qu'il n'y a pas ici de biais d'autosélection.

Loosveldt et Sonck (2008) examinent quant à eux les différences selon l'accès Internet en Belgique néerlandophone. Ils constatent des différences similaires à celles présentées par Bigot et coll. (2010). Les personnes ayant accès à Internet sont plus jeunes, plus scolarisées, plus actives. Seulement 18% des personnes ayant 10 ans de scolarité ou moins ou de celles ayant 60 ans ou plus ont accès à Internet comparativement à 80% de ceux qui ont 30 ans ou moins ou qui ont une scolarité universitaire. A partir de ces informations, ils redressent les données d'un panel Internet par un score de propension basé sur la probabilité d'inclusion dans l'échantillon. Cette pondération permet d'ajuster pour les différences dans la proportion d'urbains et de personnes en emploi mais a très peu d'effets sur les variables d'attitudes : les différences demeurent significatives pour ce qui est de la satisfaction face à l'emploi -- moins élevée chez les membres du panel--, des attitudes face à la politique --moindre intérêt mais plus forte confiance dans la capacité de comprendre les enjeux et de se faire une opinion chez les membres du panel-- et des attitudes face à l'immigration -- significativement plus négatives chez les répondants du panel Web.

Blasius et Brandt (2010) tentent une autre stratégie de comparaison, soit de produire, à partir d'un panel Internet, un échantillon aléatoire qui respecte la composition de la population et de poser à cet échantillon des questions pour lesquelles il existe de très bonnes données validées.

Premier problème rencontré, le panel Internet recruté manquait à ce point de personnes de 50 ans et plus qu'il était impossible de constituer un échantillon qui aurait pu être représentatif de la population sur le plan de l'âge et de la scolarité. Ils décident donc de tenter de représenter la population des 18 à 49 ans et comparent les données recueillies au micro-recensement et au General Social Survey allemands, restreints aux mêmes groupes d'âge. Ils montrent que, dans l'échantillon issu du panel, même après pondération, il y a nettement plus de personnes célibataires ou divorcées et sans enfant et moins de personnes croyantes ou fréquentant l'église. De plus, ces répondants se distinguent dans leurs attitudes face à la société. Ils ont plus tendance à valoriser la réalisation de soi, le laisser faire, la richesse et la contribution politique personnelle.

Pour ce qui est des États-Unis, Malhotra et Krosnick (2007) comparent des sondages faits par panel de volontaires en 2000 et 2004 aux sondages face à face faits dans le cadre de l'American National Election Study (ANES). Ils concluent que non seulement les panels de volontaires se différencient sur le plan des caractéristiques des personnes, et ceci même après pondération, mais les relations entre les variables diffèrent également. Les sondages utilisant un panel de volontaires étaient composés de moins de Noirs et de personnes peu scolarisées mais de plus personnes plus intéressées à la politique, plus susceptibles d'aller voter, plus favorables à la guerre en Irak et plus favorables à Bush. On aurait pu penser que les différences entre panelistes et répondants aux échantillons probabilistes s'estomperaient avec le temps mais Pasek et Krosnick en 2010 confirment les conclusions de la recherche précédente, cette fois en comparant un panel Internet à un échantillon téléphonique de type GANT (génération aléatoire de numéros de téléphone) avec un taux de réponse relativement bas de 10,2%, ceci dans le cadre d'une étude sur les déterminants de l'intention de participer au recensement et de la participation effective. Ils concluent que le panel internet a une moins bonne distribution démographique, qu'il diffère du sondage téléphonique de 13 points en moyenne dans la proportion de réponses modales à divers indicateurs d'opinion et jusqu'à 30 points dans un cas et enfin qu'il présente des différences significatives et non insignifiantes dans la prédiction de la participation au recensement de même que dans l'évolution dans le temps des opinions et dans les relations entre les variables.

Enfin, une dernière étude, menée par Stephenson et Crête(2011) au Québec, compare un sondage de type panel Internet et un sondage téléphonique de type GANT faits tous les deux par la même firme. Ils concluent que 40 des 52 variables montrent des distributions différentes dans les deux échantillons avant pondération et 36 après pondération. Les différences varient entre 2 points et 28 points de pourcentage. Les membres du Panel ont des opinions significativement différentes sur l'importance des divers thèmes de campagne, sur les institutions politiques, sur le rôle du gouvernement et sur un certain nombre d'attitudes relatives à la politique. En général, les membres du panel sont plus cyniques face à la politique. Ils sont également significativement mais faiblement plus intéressés à la politique mais moins intéressés aux élections. Pour ce qui est du vote, l'échantillon Web comprenait plus de personnes déclarant avoir voté, moins de personnes ayant voté pour le Parti Québécois et plus

de partisans du Parti Vert. L'échantillon téléphonique donnait une meilleure estimation du vote et la pondération n'améliorait pas la situation, au contraire. Les déterminants du vote pour le candidat déjà en poste diffèrent également selon le mode en analyse multivariée, ceci pour trois des 23 déterminants analysés.

Les études présentées constituent un échantillon des études comparant les panels Internet aux sondages utilisant des échantillons probabilistes. Toutefois, tous les articles font également une recension des écrits. Tous concluent à des différences significatives et non insignifiantes entre les panels Internet et les sondages à échantillon probabiliste. De plus, Bigot et coll. (2010) concluent de la recension des écrits que les éléments sur lesquels les panels Internet diffèrent varient de façon non-systématique et difficilement prévisible selon les études. Ceci est corroboré par notre échantillon d'études.

Ces résultats montrent que la pondération ou le redressement des données des panels internet ne permet pas nécessairement de corriger les biais reliés à la couverture ni ceux dus à l'auto-sélection des répondants. En résumé, même s'il y avait un bon nombre de femmes de 65 ans et plus peu scolarisées dans l'échantillon, celles-ci représenteraient-elles adéquatement les femmes ayant les mêmes caractéristiques qu'elles? Les recherches suggèrent que non. Le recours à la pondération est régulièrement invoqué pour légitimer les panels internet. L'argument est qu'à partir du moment où l'on peut corriger l'échantillon en rétablissant les proportions réelles des divers groupes dans la population, il ne devrait pas y avoir de biais. Ce raisonnement est le même que celui qui justifie l'utilisation d'échantillons par quotas. Deux points doivent être soulignés ici. D'une part, les échantillons par quotas sont maintenant constitués à partir d'échantillons probabilistes, ce qui n'est pas encore le cas pour les panels Internet. D'autre part, les échantillons par quotas sont fréquemment pointés du doigt comme responsables des erreurs des sondages électoraux, par exemple en Grande Bretagne en 1992 et en France en 2002. Enfin, un problème demeure, soit qu'il faudrait parfois attribuer des poids beaucoup plus élevés à certains répondants qu'à d'autres pour rétablir un semblant de distribution acceptable des groupes d'âge et de scolarité dans la population et que l'absence de certains groupes dans les panels entraîne que les membres de ces groupes sont fortement sollicités.

Quels sont les problèmes liés à l'utilisation même d'Internet pour faire des sondages?

Certains problèmes -- ou parlons plutôt de difficultés -- sont communs à tous les sondages Internet. Les faibles taux de réponse sont souvent mentionnés. Pour ce qui est des panels de volontaires, on pourrait penser que, comme les répondants se sont inscrits volontairement et qu'ils ont droit à une rémunération, même faible, ils seraient de « bons » répondants. Or lorsque les informations sur les taux de réponse de ce type de sondages sont rendu publiques, on parle parfois de taux de réponse généralement plus bas que pour les autres modes et aussi bas que moins de 1 pourcent. On peut penser que l'échantillon de répondants devient alors fortement biaisé en faveur des personnes « qui ont du temps ». S'il faut un échantillon de départ de 100 000 personnes pour être en mesure d'obtenir 1000 répondants, cela a des

conséquences. Avec de tels taux de réponse, certains membres des panels sont sollicités pour tous les sondages faits par l'institut de sondage et cette abondance de requêtes entraîne à la baisse le taux de réponse. Par ailleurs, les taux de réponse et la difficulté du recrutement amènent les instituts à ne pas renouveler leur échantillon de sorte que l'on se retrouve avec des échantillons de répondants professionnels. Doug Rivers de la firme YouGov dans une présentation à la conférence de AAPOR (American Association of Public Opinion Research) en 2008 indiquait que des estimations voulaient que 90% des sondages Internet soient complétés par trois pour cent des Internautes. Ceci dit, en général, tous les sondages utilisant internet, y compris ceux utilisant des échantillons probabilistes, ont tendance à avoir de moindres taux de réponse que les sondages effectués par d'autres moyens auprès de populations similaires (Lozar Manfreda et coll., 2008), ceci pour diverses raisons dont la plus grande difficulté à avoir recours à des « outils de motivation » et le fait que les requêtes « disparaissent » plus rapidement de la vue qu'un sondage papier qui peut demeurer visible sur la table, le comptoir ou le bureau.

Un autre problème invoqué à l'occasion est lié à la mesure et donc au questionnaire. Couper (2000) souligne que, en l'absence d'interviewer, personne ne peut aider le répondant en expliquant certains termes, par exemple. Toutefois ce problème est présent dans tous les sondages auto-administrés. Quelle est donc la spécificité des sondages Internet sur ce plan? Le questionnaire administré en format papier donne au départ un certain nombre d'informations au répondant sur ce qui l'attend, ne serait-ce que par le format et la mise en page – plus ou moins professionnelle – du questionnaire et par le nombre de pages. Le répondant peut se repérer. Ce format ne peut se transposer tel quel aux sondages Internet puisque, aussitôt répondue, la question « disparaît » et est oubliée, ce qui rend plus difficile la tâche du répondant. L'élaboration du questionnaire demande donc de porter une attention presque maniaque à faciliter la vie du répondant si l'on veut obtenir des réponses fiables. Par ailleurs, la tentation peut être grande pour le répondant de répondre rapidement à ce type de sondage, l'opinion devenant alors moins fiable. De façon à pallier ce type de problème, les Instituts de sondage utilisant des panels ont tendance à éliminer les questionnaires remplis trop rapidement.

Les recherches sur les questionnaires Internet se sont multipliées au cours des dernières années. On recommande entre autres de donner des indications sur la progression du questionnaire, de regrouper les listes de variables comportant une même échelle de réponse dans un format de type tableau sur une seule page, de permettre la non réponse sauf exception, de permettre le retour en arrière dans le questionnaire, d'insérer des phrases de transition lorsque nécessaire. Ces recommandations ne sont toutefois pas toujours suivies par les instituts.

Un dernier aspect mérite notre attention. Le répondant doit accorder une confiance presque aveugle à la personne qui le sollicite puisque, qu'il s'agisse d'un chercheur ou d'un institut de sondage, cette personne peut l'identifier. Bien sûr, les chercheurs et instituts de sondage prennent des moyens pour s'assurer de protéger l'anonymat des répondants et respectent un code d'éthique strict. Toutefois, certains instituts recueillent des informations très détaillées sur les membres des panels – adresse civique incluant le code postal, numéros de téléphone, date

de naissance, etc. On peut penser que, dans des situations tendues socialement ou politiquement, la confiance des répondants dans la stricte confidentialité des informations qu'ils fournissent peut devenir centrale, nuire au recrutement et entraîner un fort taux de non réponse aux questions politiques sensibles comme l'intention de vote chez les répondants.

Pourquoi donc fait-on des sondages Internet?

Bien qu'habituellement personne n'a tendance à accorder une crédibilité aux sondages « ouverts à tous » sur un site web, l'industrie du sondage a développé un discours légitimant son recours aux panels de volontaires. On postule qu'à partir du moment où Internet est relativement accessible dans une société donnée, les personnes qui y ont accès sont représentatives des personnes ayant les mêmes caractéristiques sociodémographiques mais n'ayant pas accès à Internet. Cette assertion n'est pas validée, comme on l'a vu précédemment, ceci parce que le profil des personnes ayant accès à Internet est, en ce moment, très différent de celui des personnes n'y ayant pas accès et que ce profil est lié à ce que l'on veut habituellement mesurer. Toutefois, il est évident qu'avec la généralisation de l'accès à Internet, ces différences s'estomperont. Il est aussi évident que, la crédibilité des instituts étant en jeu, ceux-ci tentent par tous les moyens d'évaluer et de corriger les biais possibles de leurs échantillons. Enfin, étant donné la prolifération de ce type de sondages et la non-pertinence de la marge d'erreur pour estimer leur fiabilité, il faut trouver de nouveaux moyens d'estimer cette fiabilité, entre autres en comparant les réponses à des variables sociopolitiques dont la répartition est connue et validée par d'autres méthodes.

Il restera alors à se demander comment procéder pour constituer des échantillons probabilistes de répondants, ce qui demeure le cœur du problème. En effet, au-delà de l'accès à Internet, c'est d'abord et avant tout le type d'échantillon qui est problématique. Les instituts ont développé un discours voulant que, comme les taux de réponse aux sondages téléphoniques diminuent, ces sondages sont en quelque sorte, eux aussi, composés de volontaires. C'est faire fi du fait que les répondants aux sondages téléphoniques ne se sont pas auto-sélectionnés au départ. Toutefois, il se développe en ce moment un mode de recrutement pour les panels appelé « river sampling » : Les logiciels sollicitent au hasard des personnes ayant visité certains sites. On ne sait pas jusqu'à quel point cette méthode donne des résultats mais elle montre au moins qu'il y a des efforts faits pour constituer des échantillons où la probabilité jouerait un certain rôle.

Les sondages Internet, présent et avenir

Au vu de ce qui a été mentionné précédemment, dans quelles circonstances peut-il apparaître justifié d'utiliser Internet pour faire un sondage? Le recours à Internet est tout à fait justifié, efficace et peu coûteux lorsque l'on désire recueillir l'opinion d'employés d'une organisation,

d'étudiants d'une Université, de membres d'une association alors que l'on a généralement accès à une liste d'adresses courriel couvrant tous les membres de la population. Dans ce cas, le problème de couverture étant réglé, il faut concentrer les efforts sur les autres problèmes, soit la qualité du questionnaire, la préservation de la confidentialité des réponses et le taux de réponse. Ce dernier est fortement lié aux deux premiers aspects (confidentialité et questionnaire) mais il est également tributaire des moyens utilisés pour inciter à la réponse. Dillman (2000) présente plusieurs moyens pour améliorer la collaboration entre autres en insistant sur des aspects différents à chaque fois que l'on fait un rappel aux répondants. Pour ce qui est de la confidentialité, le recours à des firmes externes qui ont seules accès à l'identificateur des répondants peut aider à rassurer les inquiets. Enfin, les questionnaires ne doivent pas recueillir des informations personnelles permettant d'identifier les répondants si celles-ci ne sont pas essentielles à la recherche. Ces éléments étant pris en compte, les sondages Internet constituent un moyen très efficace et peu coûteux de recueillir une information fiable.

Pour ce qui est des sondages utilisant des panels de volontaires, leur prolifération est en partie due au manque de formation et d'information de la population et en premier lieu, des journalistes censé informer cette population. C'est à mon avis d'abord sur ce plan que les efforts doivent porter. Sinon, on risque de voir perdurer des situations où la tentation sera grande de faire des sondages éclair sur les événements du jour, sondages qui feront la Une des médias du lendemain et qui pourront influencer l'opinion de la population sur des bases qui peuvent être trompeuses ou dont on ne peut pas mesurer la fiabilité. Seule la capacité de la population et des medias à distinguer les sondages fiables des moins fiables, à estimer la fiabilité des estimations et à comprendre et expliquer les notions de base que sont les taux de réponse et les marges d'erreur peut permettre de prévenir une telle dérive.

Références :

Bigot, R, P. Croutte et F. Recours (2010). *Enquêtes en ligne : peut-on extrapoler les comportements et les opinions des internautes à la population générale?* Centre de recherche pour l'étude et l'observation des conditions de vie (CREDOC), texte manuscrit, 63 p.

Blasius, J. et M. Brandt (2010). Representativeness in Online Surveys through stratified samples, *BMS, 107,* p. 5-21.

Couper, M. P. (2000) Web Surveys : A review of Issues and Approaches, *Public Opinion quarterly, 64(4),* 464-494.

Dillman, D. (2000). *Mail and Internet Surveys, the Tailored Design Method*, New York: Wiley and Sons, 463 p.

Loosveldt, G. et N. Sonck (2008). An evaluation of the weighting procedures for an online access panel survey, *Survey Research Methods, 2 (2),* 93-105.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J. Haas, I. et V. Vehovar (2008). Web surveys versus other survey modes. A meta-analysis comparing response rates. International Journal of Market Research, *50 (1),* 79-104.

Malhotra, N et J. A. Krosnick (2007). The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability, Samples *Political Analysis, 15(3)*, 286-323.

Pasek, J. et J.A. Krosnick (2010). *Measuring Intention to Participate and Participation in the 2010 Census and their Correlates and Trends: Comparison of RDD Telephone and Non-Probability Internet Survey Data*, Statistical Research Division, US Census Bureau, Washington, Study Series Survey Methodology #2010-15, 71 pages.|

Stephenson, L. B. et J. Crête (2011). Studying Political Behavior: A Comparison of Internet and Telephone Surveys, *International Journal of Public Opinion Research, 23 (1)*, 24-49.