Résumé

Après des mois ou des années de développement et de mise en œuvre, naviguant entre les écueils pratiques, théoriques et inférentiels de la recherche expérimentale en sciences sociales, votre expérience est enfin terminée. En comparant les groupes de traitement et de contrôle, vous trouvez un résultat substantiellement et statistiquement significatif pour un résultat d'intérêt théorique. Avant de pouvoir sabrer le champagne pour célébrer une intervention bien évaluée, un collègue sympathique demande : "mais qu'est-ce que cela nous apprend sur le monde ?"

1. Qu'est-ce que la validité externe ?

La validité externe est un autre nom pour la généralisation des résultats, demandant "si une relation causale tient pour une variation des personnes, des paramètres, des traitements et des résultats".[1] Un exemple classique de problème de validité externe est de savoir si les expériences traditionnelles en laboratoire d'économie ou de psychologie menées sur des étudiants universitaires produisent des résultats généralisables au grand public. Dans l'économie politique du développement, nous pourrions considérer comment un programme de développement communautaire en Inde pourrait s'appliquer (ou non) en Afrique de l'Ouest ou en Amérique centrale.

La validité externe devient particulièrement importante lors de la formulation de recommandations politiques issues de la recherche. L'extrapolation des effets causaux d'une ou plusieurs études à un contexte politique donné nécessite un examen attentif à la fois de la théorie et des preuves empiriques. Ce guide des méthodes aborde certains concepts clés, les pièges à éviter et les références utiles à prendre en compte lors du passage d'un effet moyen local du traitement au monde plus vaste.

2. En quoi est-elle différente de la validité interne ?

La validité interne fait référence à la qualité des inférences causales faites pour un groupe de sujets donné. Comme l'a postulé à l'origine Campbell[2], la validité interne pose la question suivante : "le stimulus expérimental a-t-il en fait causé une différence significative dans ce cas spécifique ?" Ce concept concorde avec l'approche contrefactuelle de la causalité que les expérimentalistes utilisent généralement, qui demande si les résultats changent en fonction de la présence ou de l'absence d'un traitement.[3]

Avant de pouvoir extrapoler un effet causal à une population distincte, il est essentiel que l'effet moyen de traitement original soit basé sur un résultat bien identifié. Pour la plupart des expérimentalistes, l'assignation aléatoire fournit la variation d'identification requise, à condition qu'il n'y ait ni attrition, ni interférence, ni débordement ou autres menaces à l'inférence. Pour les études observationnelles, des hypothèses d'identification supplémentaires sont nécessaires, telles que l'indépendance conditionnelle du traitement par rapport aux résultats potentiels.

3. Faire des compromis entre validité interne et externe

Il y a débat au sein des sciences sociales concernant l'importance relative de l'identification de résultats valides en interne, qui par définition s'appliquent à un échantillon local, et la génération de résultats qui peuvent être extrapolés à des populations d'intérêt plus larges. Il est utile de se familiariser avec cette discussion lorsque

[1] Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton, Mifflin and Company.

[2] Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. Psychological bulletin, 54(4), 297.

[3] Plus de détails disponibles dans le guide des méthodes d'inférence causale: http://egap.org/resource/10-things-to-know-about-causal-inference

l'on considère les compromis de design qui surgissent inévitablement dans les interventions à ressources limitées. Des sommités de l'économétrie ont pris partie des deux côtés, ce qui atteste l'importance du sujet.

D'un côté, les partisans de "l'identification d'abord", qui soutiennent que sans des résultats valides en interne, une étude ne fournit tout simplement pas d'informations utiles, qu'il s'agisse d'une population ou d'un contexte local ou général. Comme le dit Imbens,[4] "sans une solide validité interne, les études ont peu à contribuer aux débats politiques, alors que les études [avec validité interne] mais une validité externe très limitée sont souvent, et à mon avis devraient être prises au sérieux dans de telles discussions."

D'autres soutiennent que même sans l'identification complète d'un résultat valide en interne, des informations utiles peuvent être récupérées, surtout si elles sont pertinentes pour des questions importantes qui affectent un large contexte. Manski[5] écrit que "ce qui compte, c'est le caractère informatif d'une étude pour l'élaboration des politiques, qui dépend conjointement de la validité interne et externe". Avec les données d'une étude large mais mal identifiée, soutient Manski, on peut borner l'estimande d'intérêt, ce qui fait toujours avancer la science, même si ce n'est pas aussi utile qu'une estimation ponctuelle précise.

4. Théorie et généralisation

L'extrapolation d'une découverte à un contexte, un résultat, une population ou un traitement distincts n'est pas un processus mécanique. Comme discuté par Samii[6] et Rosenbaum,[7] une théorie pertinente doit être utilisée pour guider la généralisation, en prenant les preuves existantes pertinentes et en faisant des prédictions pour d'autres contextes d'une manière fondée sur des principes. Les théories font d'un problème complexe une représentation plus parcimonieuse, aidant ainsi à élucider les facteurs qui comptent. Tout comme la théorie guide le contenu des interventions et des designs de recherche, les propositions théoriques peuvent vous indiquer quelles conditions sur la portée de l'expérience sont pertinentes pour extrapoler un résultat. Quelles covariables importent ? Quelles informations contextuelles sont importantes ?

5. Comment puis-je déterminer où mes résultats s'appliquent ?

Il existe deux principaux moyens de généraliser les résultats, l'un basé sur les covariables des unités de l'étude et l'autre basé sur la manipulation expérimentale réelle des variables modératrices. L'observation de la variation d'un effet de traitement sur une variable de pré-traitement non randomisée peut décrire l'hétérogénéité de l'effet de traitement, qui peut suggérer où ou pour qui l'intervention est susceptible d'être la plus efficace, au-delà de l'échantillon d'origine. Notez, cependant, que ce type d'analyse ne peut pas déterminer si l'hétérogénéité de l'effet de traitement est causée par cette variable de pré-traitement. Le problème, endémique à la recherche observationnelle, est que la covariable non randomisée peut être corrélée à une variable non observée, et c'est ce facteur "invisible" qui est en fait responsable des impacts hétérogènes du traitement.[8] Idéalement, par conséquent, nous voulons tirer parti de la variation exogène du modérateur d'intérêt, excluant ainsi la possibilité d'une telle confusion. Un design expérimental factoriel dans lequelle le chercheur assigne le modérateur indépendamment du traitement principal d'intérêt peut générer des preuves particulièrement convaincantes sur le rôle d'un modérateur ; bien que des considérations de coût et de puissance statistique puissent empêcher cette approche dans la pratique.

Étant donné que la généralisation est avant tout un exercice de prédiction, se demander où s'attendre à une relation causale similaire à celle observée localement, extrapoler des effets hétérogènes sur la base

[4]Imbens, G. (2013). Book Review Feature: Public Policy in an Uncertain World: By Charles F. Manski. The Economic Journal,123(570), F401-F411.

[5]Manski, C. F. (2013). Response to the review of 'public policy in an uncertain world'. The Economic Journal 123: F412–F415.

[6]Samii, Cyrus. (2016). "Causal Empiricism in Quantitative Research." Journal of Politics 78(3):941–955.

[7]Rosenbaum, Paul R. (1999). "Choice as an Alternative to Control in Observational Studies" (avec discussion). Statistical Science 14(3): 259–304.

[8]Gerber, A. S., & Green, D. P. (2012). Field experiments: Design, analysis, and interpretation. WW Norton.

de covariables similaires est souvent raisonnable, à condition que la théorie n'indique pas de sources de confusion.[9] Néanmoins, la plus forte la preuve de la généralisabilité d'un résultat provient d'une interaction bien identifiée entre un modérateur exogène et le traitement, qui sera ensuite projetée à travers le profil de covariable d'une population cible. En effet, avec certaines hypothèses fortes, l'extrapolation peut fournir des résultats aussi bons ou meilleurs que la réalisation d'une deuxième expérience in situ.[10] Le calcul d'une estimation extrapolée peut souvent être mieux effectué à l'aide de l'apprentissage automatique, bien que la régression linéaire fonctionne également assez bien. [11]

6. Un comportement stratégique peut faire échouer vos extrapolations

L'extrapolation d'un résultat local à un contexte différent peut s'avérer difficile, même avec un profil de covariable convaincant auquel vous souhaitez généraliser les effets. Une manipulation expérimentale randomisée dans une zone locale génère un "effet d'équilibre partiel". Les dynamiques stratégiques, y compris les comportements compensatoires ou les contrecoups, en dehors du contexte local d'une intervention expérimentale peuvent compliquer les efforts de généralisation d'un résultat. Supposons, par exemple, qu'une intervention de transfert monétaire inconditionnel augmente le bien-être, l'entrepreneuriat et l'emploi dans un échantillon de 200 villages. Que se passerait-il si l'intervention était étendue à 1000 villages ? À ce stade, on pourrait imaginer que les régions exclues du programme sont plus susceptibles d'en prendre connaissance. Les unités non traitées peuvent commencer à exiger d'autres types de transferts de la part du gouvernement, provoquant des effets similaires à ceux produits par les transferts monétaires directs. Dans le même ordre d'idées, les relations causales ne fonctionnent parfois que lorsqu'elles sont appliquées à certaines personnes. Par exemple, imaginez un programme de compétences professionnelles qui fonctionne très bien (par rapport à ceux qui ne l'ont pas reçu), que se passerait-il s'il était étendu à tous les travailleurs ? Même s'il y a des effets positifs sur tous les participants, les effets moyens pourraient être réduits ou nuls, car les emplois les plus qualifiés sont déjà pourvus par le premier lot et le deuxième lot est contraint de conserver son emploi précédent, en étant désormais surqualifié. En bref, dans des conditions d'équilibre général, nous pourrions nous attendre à des résultats différents même lorsque le profil de covariable correspond.

7. Ne pas confondre la validité externe avec la validité de construit ou la validité écologique

La validité interne et externe ne sont pas les seules préoccupations de "validité" qui peuvent être soulevées dans le travail expérimental, et bien que pertinentes, elles sont également distinctes. La validité écologique, telle que définie par Shadish, Cook et Campbell[12], concerne le fait qu'une intervention semble artificielle ou déplacée lorsqu'elle est déployée dans un nouveau contexte. Par exemple, un atelier d'information dans une ville rurale mené par des expérimentalistes ressemble-t-il à la manière habituelle de partager l'information pour cette population ? De même, si le même atelier se tenait dans une grande ville, semblerait-il déplacé ?

La validité de construit considère si un concept théorique testé dans une étude est correctement restitué par le(s) traitement(s). Si votre expérience teste l'effet de la colère sur la réciprocité politique et que vous manipulez en fait la peur ou la confiance dans votre traitement, la validité de construit peut être violée. La validité de construit et la validité écologique sont toutes deux pertinentes pour les généralisations, et donc utiles pour affirmer la validité externe.

[9]Bisbee, James; Rajeev Dehejia; Cristian Pop-Eleches & Cyrus Samii. (2016). "Local Instruments, Global Extrapolation: External Validity of the Labor Supply-Fertility Local Average Treatment Effect." Journal of Labor Economics

[10]Bisbee, James; Rajeev Dehejia; Cristian Pop-Eleches & Cyrus Samii. (2016). "Local Instruments, Global Extrapolation: External Validity of the Labor Supply-Fertility Local Average Treatment Effect." Journal of Labor Economics

[11]Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. Journal of Research on Educational Effectiveness, 9(1), 103-127.

[12]Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton, Mifflin and Company.

8. Extrapolation entre les traitements et les résultats

Bien qu'une grande partie de ce guide se soit implicitement centrée sur le transfert d'un traitement donné à un nouveau lieu ou moment, la validité externe prend également en compte les variations dans les traitements et les résultats. C'est-à-dire, imaginez que nous ayons mené la même expérience sur le même échantillon, mais avec une variation sur le traitement, prédirions-nous que l'effet causal local sera similaire ? De même, pouvons-nous prédire si un traitement donné produira des effets causaux identiques ou différents sur un résultat différent ? Parfois, nous pouvons répondre à ces préoccupations en menant des expériences qui évaluent les traitements alternatifs et leurs résultats. Lorsque les expériences de suivi sont rares, ces problèmes doivent être réglés analytiquement. Plutôt que de considérer les caractéristiques des sujets, l'extrapolation dans ce cas nécessite de réfléchir, à l'aide de la théorie, aux caractéristiques des traitements ou des résultats et de faire des prédictions raisonnables.

9. La reproductibilité est importante

Aucune étude n'a le dernier mot sur une question scientifique. Suivant la logique de la mise à jour bayésienne, des preuves supplémentaires en faveur ou contre une théorie donnée permettent à la communauté scientifique et politique de mettre à jour leurs croyances sur la force et la validité d'une relation causale.

La reproductibilité des études en est une partie importante : les chercheurs doivent pouvoir reproduire des études dans des contextes qui semblent très différents, mais aussi dans certains contextes qui semblent très similaires. Le premier nous permet d'identifier des relations causales locales qui peuvent être triangulées avec les preuves existantes et généralisées le cas échéant. Dans le même temps, il est important de répliquer directement les études existantes dans des conditions aussi proches que possible de l'original afin de vérifier que les effets locaux qu'on cherche à extrapoler sont bien fiables. L'Open Science Collaboration[13] a constaté, par exemple, que lors de la reproduction de 100 expériences majeures en psychologie, seulement 47 % des tailles d'effet initialement signalées se situaient dans l'intervalle de confiance de 95 % de la taille d'effet indiquée dans l'étude reproduite.

10. N'oubliez pas le temps

Lorsque l'on réfléchit aux relations causales d'intérêt, il est également important de considérer le temps : les choses que nous apprenons sur le passé s'étendent-elles au futur ? Comment les résultats potentiels d'un individu changent-ils au fil du temps ? Des lois immuables régissent les mondes physique et chimique ; par conséquent, ce que nous apprenons sur ces lois aujourd'hui restera toujours vrai. En revanche, nous comprenons beaucoup moins les moteurs sous-jacents du comportement social et s'ils restent constants de la même manière. Ce pourrait ne pas être le cas. Lors de la prise de décisions concernant la pertinence politique et la généralisabilité des résultats, ces considérations peuvent aider les chercheurs à déterminer un niveau d'incertitude raisonnable et aider les décideurs à s'adapter en conséquence.

[13] Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.