L'ÉVALUATION DES POLITIQUES DE L'EMPLOI AU RÉGIME DE L'EXPÉRIMENTATION : LA DÉMONSTRATION D'IMPACT

Olivier BAGUELIN

Université Paris-Saclay, Univ Evry, CEPS, Tepp (FR 2042)

Laurent DUCLOS

Université Paris-Saclay, ENS Paris-Saclay, IDHES (UMR 8533)

Résumé

Dans une perspective d'innovation sociale, après un premier opus consacré à la preuve de concept, la présente discussion s'intéresse à la démonstration d'impact. On part pour cela des difficultés pratiques et méthodologiques rencontrées par les essais avec assignation aléatoire contrôlée (RCT) dans le cadre d'expérimentations de terrain. On examine les approches disponibles pour surmonter ces limites, plus particulièrement l'évaluation basée sur la théorie et l'approche bayésienne. Evaluation basée sur la théorie et réseaux bayésiens offrent un cadre permettant de mettre en communication les enseignements des différentes méthodes d'évaluation, qu'elles soient quali- ou quantitatives, qu'il s'agisse d'enquête monographique, d'études comparative ou statistique. L'enquête qualitative est indispensable pour comprendre les mécanismes des interventions et leurs interactions avec un *environnement*. On montre que l'inférence causale peut s'appuyer sur l'existence de points de rencontre entre conceptions processuelle, logique et probabiliste-contrefactuelle de la causalité. La discussion met en lumière les défis et les opportunités de l'évaluation expérimentale des politiques de l'emploi en insistant sur la portée d'approches intégrées quali-quanti.

Le présent opus est le second d'un couple de numéros de Socio-économie du travail (SET) consacrés à l'expérimentation dans le domaine des politiques de l'emploi et à la façon dont l'approche expérimentale renouvelle leur évaluation. L'ambition est de dresser un état des lieux de la réflexion en la matière qui s'est considérablement développée ces vingt-cinq dernières années à mesure que l'application de l'expérimentation aux politiques publiques se répandait. Accompagnant une ferveur transnationale pour l'*evidence based policy*, la démarche promettait d'adopter telle ou telle mesure non sur la foi d'une doctrine politique ou d'une théorie socio-économique mais de faits probants. En lieu et place d'aprioris théoriques ou idéologiques, une démarche systématique d'évaluation adossée à un instrument essentiel de la méthode scientifique : l'expérimentation. Cette promesse de mieux maîtriser les biais de confirmation pouvant entacher l'évaluation, a reçu un écho particulièrement enthousiaste, en France, dans le domaine des politiques de l'emploi (Zamora, 2011 ; Labrousse et al., 2013).

Sur ces prémices, l'appel à articles lancé par SET posait deux questions principales quant au rôle de l'expérimentation : comment influence-t-elle la fabrique des politiques de l'emploi ? En quoi renouvelle-t-elle la réflexion en matière d'évaluation d'impact ? Traitées en introduction générale (Baguelin et al., 2023) dans une perspective d'innovation sociale, ces questions ont été raccordées à deux enjeux donnant une fonction différente à l'expérimentation : la preuve de concept (est-ce que ça peut marcher ? Si oui, sous quelles conditions ?) d'un côté ; la démonstration d'impact (est-ce que ça marche ? Si oui, comment ? Si non, pourquoi ?) de l'autre. Le présent opus réunit un ensemble de contributions centrées sur le second enjeu et plus précisément sur la façon dont l'expérimentation amène l'évaluation d'impact à diversifier ses méthodes et à tenter de les articuler.

Les essais contrôlés avec assignation aléatoire (*randomized controlled trials* ou RCT selon l'acronyme anglais) se sont imposés comme méthode privilégiée de la démonstration d'impact en raison notamment de doutes persistants sur la portée de la théorie économique standard. Dans le champ académique, cela est intervenu en faveur des *randomistas* (tel qu'Angus Deaton désigne les praticiens des RCT) mais aux dépens des tenants de l'économétrie structurelle, que l'on désignera ici d'*estructuristas* par esprit de symétrie. Les promesses des RCT se sont cependant avérées difficiles à tenir et plus exigeantes qu'attendu en termes de réflexion théorique. La première partie de la présente introduction explique en quoi. Pour le présent exercice éditorial, c'est une nouvelle occasion de pointer la relation ambigüe entre théorie et évaluation[1], cette fois à l'étape de la démonstration d'impact. Prétendre appliquer la méthode scientifique à l'évaluation d'impact met inévitablement en jeu une réflexion théorique. Le cœur de ce second opus est de clarifier le type de théorie requis et son rôle dans la démonstration d'impact. On convoque pour cela deux courants indépendants mais remarquablement cohérents dans leur principe : l'évaluation basée sur la théorie (ou *theory-based evaluation*, TBE) et l'approche bayésienne. Ces courants orientent vers une évaluation de politiques publiques (EPP) usant d'approches et méthodes variées (Stern et al., 2012, p. 15) et articulant étroitement enquête qualitative et quantitative. Les difficultés méthodologiques qui en découlent sont particulièrement discutées par les politistes et les spécialistes de l'aide au développement : la présente introduction s'efforce d'en faire une synthèse adaptée aux enjeux spécifiques de l'évaluation des politiques de l'emploi. On en tire un ensemble de propositions visant une évaluation pluraliste mieux équipées méthodologiquement. Les contributions du numéro[2] ont inspiré l'essentiel de ces propositions. Les trois premières traitent directement de méthode par la voie du témoignage (celui d'Adam Baïz, recueilli dans le cadre d'une journée d'étude), du commentaire critique (Solveig Grimault) et de l'analyse de corpus (Yves de Curraize et Francesco Sergi). Les quatre suivantes sont des travaux d'évaluation prenant en charge les difficultés d'une évaluation pragmatique. Deux exercices portent sur des dispositifs du Plan d'investissement dans les compétences (PIC). L'article d'Agathe Devaux-Spatarkis, Pauline Joly et Thomas Bouget, développe une évaluation croisant TBE et réflexivité des parties prenantes dans le cadre de l'expérimentation de dispositifs d'insertion par la formation : à défaut d'estimer des effets causals moyens, l'enquête évaluative explicite utilement les mécanismes de l'intervention. Carole Beaugendre, Elise Crovella, Jeoffrey Magnier et Isabelle Recotillet s'intéressent à l'évaluation d'un dispositif PIC de pré-qualification en couplant

[1] Dans l'introduction générale proposée au premier opus, l'action publique menée ces vingt dernières années en France est analysée en termes de politiques de l'emploi « au régime de la théorie » pour dire la prééminence des arguments théoriques (coût du travail, employabilité, activation…) aux dépens de l'évaluation *ex post*. Soumettre les politiques de l'emploi au régime de l'expérimentation a en ce sens réhabilité l'évaluation, un rapport plus sceptique à la théorie économique rejoignant une exigence accrue d'évaluation crédible. C'est d'ailleurs cette exigence qui a précipité le regain d'intérêt du monde de la recherche empirique pour l'évaluation, l'exercice lui offrant d'appliquer ses méthodes à un matériau expérimental. En économie du travail, la démarche expérimentale rejoignait une « révolution de la crédibilité » appelée à bousculer les certitudes.

[2] Et celles du numéro précédent sur la même thématique (Baguelin et al., 2023).

mesure d'impact contrefactuel et enquête qualitative. L'article documente les difficultés pratiques posées par l'évaluation en méthode mixte : le constat de divergences entre mesures d'impact et enseignements qualitatifs amène à s'interroger sur les modalités d'une bonne articulation des méthodes. Dans une perspective proche mais appliquée cette fois à un dispositif de la « Stratégie pauvreté », Nicolas Farvaque, Renaud Garrigues, Elise Picon et Carole Beaugendre, s'intéressent à une intervention complexe destinée à favoriser l'accès à l'emploi durable de salariés en insertion. Il s'agit d'une évaluation réaliste en méthodes mixtes, appliquées dans une optique d'enrichissement. Le dernier article du numéro, celui d'Annie Jolivet, propose l'analyse rétrospective d'un matériau qualitatif accumulé au fil des années pour évaluer la portée des incitations à négocier sur l'emploi des « seniors ». Elle tire de l'analyse une série d'enseignements sur la portée de l'enquête monographique pour l'évaluation, entendue comme démarche « formative » (et non seulement normative). Tous ces travaux illustrent la variété des approches et méthodes pouvant participer utilement à la démonstration d'impact. L'enjeu de la présente introduction est d'en définir un cadre unificateur. Le parcours comporte quatre étapes : un retour sur les limites d'une démonstration quantitative d'impact présentée comme athéorique ; une discussion de la proposition unificatrice que constitue la TBE ; un développement sur la manière de traiter l'inférence causale de façon cohérente sous les angles quanti- et qualitatifs ; un état des lieux des promesses des méthodes mixtes pour la démonstration d'impact.

I. LA DÉMONSTRATION QUANTITATIVE D'IMPACT

La statistique est par nature le mode d'accès à la réalité sociale privilégié par les administrations publiques. Celles-ci entendent en effet agir sur des populations et le niveau macrosocial auquel elles opèrent les prédisposent au raisonnement quantitatif. Le concept statistique d'*effet causal moyen* répond bien à leurs préoccupations. Il correspond à la comparaison entre les valeurs moyennes d'un résultat (*outcome*) dans deux situations : l'une où l'intervention est présente, l'autre où elle ne l'est pas, toutes choses égales par ailleurs. Des méthodes s'efforcent d'estimer cet effet causal moyen : de façon paradoxale, ce sont principalement celles de la *micro*-économétrie. La désignation tient à une articulation très étroite, selon un principe cher aux *estructuristas*, entre méthode statistique et théorie microéconomique standard[3]. Il s'avère que ce paradigme abolit les difficultés du passage du micro au macro en permettant de déduire additivement le macro- du micro-social… exactement ce dont la statistique a besoin.

L'évaluation micro-économétrique des politiques de l'emploi est devenue courante en France depuis 30 ans (Erhel, 2020, p. 111) : contrats aidés, indemnisation et accompagnement du chômage, formation… rares sont les interventions n'ayant pas déjà fait l'objet d'une mesure d'impact de ce type (Simonnet, 2014). Les hypothèses théoriques conditionnant la validité de ces évaluations ont pourtant rapidement embarrassé les spécialistes de l'évaluation quantitative (Angrist et al., 2009). Il apparaissait souhaitable de pouvoir démontrer l'impact en se passant d'hypothèses douteuses.

I.1. LES RCT POUR SE LIBÉRER DE LA THÉORIE

Le début des années 2000 est ainsi marqué par un regain d'intérêt pour une méthode ancienne : l'expérimentation par RCT. Empruntés à la médecine, les RCT sont appliqués à l'EPP dès les années 1960 (Lewitt et al., 2008). Après une phase de relative éclipse dans les

[3] Celle-ci représente les situations sociales en termes d'équilibres coordonnant les choix autonomes d'agents rationnels versés au calcul coût-bénéfice.

années 1980, ils trouvent un regain de faveur au début des années 2000 au point d'accéder au titre d'étalon-or de la démonstration d'impact (de Curraize et Sergi dans le présent numéro).

L'idée est d'assimiler les interventions à des traitements et de constituer par tirage au sort deux groupes d'individus : l'un exposé au traitement (groupe traité), le second non exposé (groupe témoin). Au prix certes d'une perte d'efficacité statistique, l'assignation aléatoire est entendue comme procédure d'équi-répartition entre les deux groupes de caractéristiques individuelles susceptibles d'être confondantes[4] (*confounding*). Pour les économistes, cette méthode semble opérationnaliser le raisonnement « toutes choses égales par ailleurs » et pour les critiques de la microéconométrie, libérer l'évaluation d'hypothèses invérifiables.

En France, la méthode est appliquée en 2007 au placement des personnes en demande d'emploi, en confrontant un accompagnement renforcé à un suivi de base (Behaghel et al., 2009)[5]. Le protocole prévoit qu'environ 200 000 personnes en demande d'emploi soient aléatoirement assignées à l'un ou à l'autre, avec l'espoir de quantifier l'effet causal moyen de l'accompagnement. Les groupes traité et témoin étant statistiquement équivalents du fait de cette assignation aléatoire, le supplément de taux d'accès à l'emploi du groupe traité (estimé, six mois après l'assignation, entre 15 et 35 %) est imputé à cet accompagnement. Pas véritablement besoin d'une théorie des processus sociaux sous-jacents : l'accompagnement, ça marche !

Les difficultés des RCT de terrain

Les RCT apparaissent d'un intérêt considérable pour penser la causalité mais leur application hors d'un laboratoire pose des difficultés sous-estimées (Scriven, 2008 ; Deaton et al., 2018).

Premièrement, ce que l'on peut espérer contrôler en laboratoire ne peut l'être en dehors, notamment parce que le déploiement sur le terrain suppose de mobiliser des auxiliaires qui sont eux-mêmes des acteurs sociaux ; dans l'exemple cité, les référents emploi. Le traitement n'est donc jamais administré en double-aveugle (les sujets et les auxiliaires savent qu'ils participent à une expérimentation) et un effet Hawthorne ne peut être exclu[6]. L'effet causal moyen mélange donc l'effet de l'intervention, l'effet psychologique de participer à une expérimentation et l'interaction entre les deux. Conséquence : on ne sait jamais si un effet mesuré dans le contexte d'une expérimentation se manifestera en dehors.

Deuxièmement, toujours dans le registre de la validité externe, la composition de l'échantillon expérimental n'est jamais celle de la population-cible (à laquelle l'intervention est destinée). L'enrôlement d'une personne, en tant que sujet, devant (théoriquement) donner lieu à un consentement de sa part, l'effet estimé vaut pour la sous-population des « consentants », qui n'a pas de raison d'être représentative de la cible.

Troisièmement, contrairement à ce qui est parfois affirmé, les RCT n'écartent pas le risque que des facteurs confondants empêchent l'interprétation causale. L'assignation aléatoire permet d'associer aux effets estimés des erreurs types (écarts types empiriques) correctes[7] mais ne garantit l'équi-répartition des déterminants du résultat qu'en probabilité (Deaton et al., 2018, p.

[4] C'est-à-dire de déterminer le résultat indépendamment du fait d'être traité.

[5] Sur la période 2008-2020, on recense en France moins d'une dizaine d'expérimentations contrôlées avec assignation aléatoire relevant de la thématique travail (Baïz *et al.*, 2022).

[6] Les exigences concernant les RCT de laboratoire ont elles-mêmes été augmentées avec l'idée que l'analyse des données soit menée sans connaître le statut de traitement de chaque groupe : on parle de traitement administré en *triple*-aveugle.

[7] Moyennant cependant quelques précautions qui seraient hélas, selon Deaton *et al.* (2018, p. 8), souvent ignorées dans les études publiées.

5 ; Ravallion, 2020, p. 7). Le hasard peut produire l'équi-répartition d'une multitude de caractéristiques dépourvues d'importance mais, faute de chance, ne pas la produire précisément dans des dimensions confondantes.

Les politiques de l'emploi, des traitements ?

Lorsqu'un effet moyen est mesuré par RCT, à quoi est-il dû ? Quelle est la cause de l'effet ? Réduire l'intervention à un traitement (entendu comme action ponctuelle générique) rend cette question d'attribution triviale : la cause de l'effet, c'est le traitement. Mais l'analogie médicale suggérant cette réduction est-elle pertinente ? Le fait qu'une intervention de politique de l'emploi ne soit pas une pilule que l'on ingère perturbe la portée des RCT[8].

Le problème est notamment celui de l'hétérogénéité du « traitement ». Dans l'exemple de l'aide à la recherche d'emploi, le référent emploi (l'auxiliaire) peut en particulier être plus ou moins performant. Imaginons que les référents auxquels est confié le groupe traité soient en moyenne meilleurs (plus motivés) que les autres : le résultat tient-il à la qualité du dispositif ou à celle des auxiliaires de l'expérimentation ?[9,10] Les sources possibles d'hétérogénéité sont multiples et difficiles à repérer sans aller sur le terrain.

Cet enjeu d'attribution est particulièrement sensible dans le cas d'interventions complexes (Stern et al., 2012, p. 50 ; Desquinabo, 2021) : variété des dispositions (règlement, incitation monétaire, mesure de contrôle, conseil et information, accompagnement…), de leur durée (ponctuelle ou prolongée), des acteurs impliqués (politiques, administratifs, auxiliaires) et des objectifs poursuivis. Ces éléments de complexité sont omniprésents dans les politiques de l'emploi et de lutte contre le chômage et vont probablement croissant. Presque toutes les interventions évoquées dans le présent couple de numéros thématiques sont concernées : la Garantie Jeunes (Gaini, 2023 ; Gautié, 2023), l'expérimentation « Territoires zéro chômeur de longue durée » (Jany-Catrice et al., 2023 ; Tantot, 2023 ; Retsin, 2023), la loi de 2018 pour « la liberté de choisir son avenir professionnel » (Cibois, 2023), le PIC (Devaux-Spatarakis et al., Beaugendre et al., dans ce numéro), la « Stratégie pauvreté » (Farvaque et al., dans ce numéro) ou les dispositifs d'incitation à négocier (Jolivet, dans ce numéro). Dans chacun de ces cas, l'intervention publique articule des dispositions plurielles déployées de façon variée, qu'il est difficile de réduire à un traitement générique.

Il ne s'agit d'ailleurs pas seulement de savoir si un dispositif de politique de l'emploi peut être réduit à un traitement mais à l'inverse si un traitement, entendu comme une intervention évaluable par RCT, peut constituer un dispositif pertinent. En matière d'aide au développement, la prédilection pour les RCT opérerait une sélection en faveur d'interventions génériques, adaptées à une évaluation par RCT (Ravaillon, 2020, p. 35). Dans le champ de l'insertion des jeunes, la formule du Revenu contractualisé d'autonomie (RCA), expérimentée au début des années 2010 (Aeberhardt et al., 2014), a pu tirer une part de son attrait au fait de se prêter aux conditions d'un essai avec assignation aléatoire. Dans la lutte contre le chômage, les traitements

[8] Garantir une estimation sans biais nécessite de n'introduire, après l'assignation aléatoire, aucune autre différence que l'exposition à l'intervention : « facile » pour une pilule, beaucoup moins pour une action personnalisée étalée dans le temps.

[9] Il s'agit d'une des difficultés relatées par Behaghel *et al.* (2013, p. 143). On peut aussi citer l'exemple d'une expérimentation de prospection d'employeur menées par Pôle emploi dont les résultats positifs ont probablement tenu à la qualité des auxiliaires chargés de mettre en œuvre le traitement : ceux-ci étaient plus expérimentés (Aventur *et al.*, 2016, p. 11).

[10] Techniquement, on peut certes répondre à cette question d'hétérogénéité en croisant assignation au traitement et, en admettant qu'elle soit observable, qualité de l'auxiliaire. À échantillon expérimental de taille donnée, cela réduit la précision de l'estimation d'impact. Si au contraire, on ajuste à la hausse la taille de l'échantillon expérimental pour maintenir la précision, il faut aussi augmenter le nombre d'auxiliaires. Problème : plus ces auxiliaires sont nombreux, plus leur « qualité » varie de sorte que la difficulté prend de l'ampleur à mesure qu'on cherche à la surmonter.

inspirés par la psychologie comportementale[11] (effort de recherche, rationalité de la prospection, réalisme des attentes…) se prêtent particulièrement bien à l'évaluation par RCT. Ce ne sont alors pas les RCT en tant que tels qui sont en question mais l'effet de sélection que leurs conditions de validité opèrent sur les diagnostics, les priorités et/ou les paris de politique publique.

I.2. ATHÉORIQUES, VRAIMENT ?

Lorsque tout va bien, les RCT livrent une estimation sans biais de l'effet de l'intervention assortie d'une erreur type correcte. Mais à partir d'un seul essai, on n'a aucune idée de la distance de l'estimation obtenue à la valeur de population (*estimand*)[12]. Conscients du problème, les praticiens inspectent la composition des groupes traité et témoin pour repérer d'éventuels cas aberrants susceptibles de tirer l'estimation vers le haut ou vers le bas[13]. La portée d'une estimation d'impact par RCT n'est en fait jamais indépendante de la taille de l'échantillon et plus le nombre de facteurs d'un résultat est grand (pour autant qu'on le sache) plus cet échantillon doit l'être pour associer un degré de confiance correct à une estimation.

Mais comment juger qu'un échantillon est de taille suffisante[14], du caractère aberrant d'une observation ?[15] Comment diagnostiquer l'hétérogénéité d'un traitement[16]… sans se référer à un a priori sur les processus à l'œuvre, c'est-à-dire à une théorie ? Les difficultés des RCT semblent immanquablement ramener à la théorie.

Sans surprise au fond, le projet d'appliquer la méthode scientifique à l'EPP implique une réflexion théorique. On peut même penser que plus c'est explicite mieux ça vaut. Deaton et al. (2018) illustrent l'utilité de la théorie pour exploiter efficacement les données dans un but de démonstration causale en invoquant la posture bayésienne. Les RCT seraient appropriés lorsqu'on n'a aucune idée des processus causaux à l'œuvre et qu'on peut disposer d'un large échantillon[17]. Il suffit d'un peu de connaissance a priori pour qu'il soit préférable, en termes de rythme d'apprentissage et/ou de précision estimative[18], de cantonner l'assignation aléatoire (*controlled assignment*). Si, pour une relation d'intérêt, on soupçonne un facteur d'être confondant, il est toujours préférable d'en tenir compte (en stratifiant ou, dans le cadre d'une expérimentation contrôlée, en forçant son équi-répartition).

---

[11] Voir l'engouement suscité par les *nudges*.

[12] En présence de réplications, on peut certes procéder à une méta-analyse… mais disposer de RCT de terrain répliqués largement pour une intervention donnée reste exceptionnel. Au demeurant, quand elles sont possibles, les méta-analyses ne valident pas nécessairement le statut d'étalon-or octroyé aux RCT. En médecine, une méta-analyse montre, pour un même traitement, que les effets estimés à partir d'études observationnelles (études de cohortes) sont en moyenne similaires à ceux des RCT… en étant moins dispersés (Concato *et al.*, 2000). Cela tient vraisemblablement à la perte d'efficacité statistique associée à l'assignation aléatoire.

[13] L'usage est d'éprouver la robustesse de l'estimation au retrait de ces cas.

[14] Moins les causes déterminantes seront variées moins un échantillon de taille réduite sera problématique (l'assignation aléatoire ayant plus de chances de réaliser une équi-répartition sur deux ou trois déterminants que sur une dizaine). Surmonter la perte de précision associée à l'assignation aléatoire peut passer par une étape de stratification qui, pour être utile, suppose une compréhension préalable des déterminants qu'il importe d'équi-répartir.

[15] Qu'est-ce qu'un cas aberrant sinon une déviation par rapport à un *a priori* ?

[16] La spécification correcte d'un traitement hétérogène (effet croisé) nécessite une théorie des déterminants du résultat.

[17] Une étude drolatique parue dans le *British Medical Journal* (Smith *et al.*, 2003) le démontre par l'absurde. Les auteurs proposent une revue systématique des résultats disponibles sur l'utilité du parachute et déplorent l'absence de RCT sur cette question.

[18] Efficacité dans l'utilisation de la variabilité.

Cette posture bayésienne donne peut-être la priorité à l'apprentissage (cumul des savoirs) sur l'exigence d'absence de biais, ce qu'une perspective d'EPP pourrait contester[19]. Les *estructuristas* soutiennent exactement l'inverse lorsqu'ils fixent à l'EPP l'agenda suivant (Heckman et al., 2024, p. 4)[20] : (i) évaluer l'impact d'interventions mises en œuvre dans un contexte donné, y compris en termes de bien-être des personnes traitées et de la société dans son ensemble ; (ii) comprendre les mécanismes produisant les effets des traitements et les résultats de politiques ; (iii) prévoir les impacts (construction d'états contrefactuels) d'interventions mises en œuvre dans un contexte donné, lorsqu'elles sont appliquées à d'autres contextes, y compris leurs impacts en termes de bien-être ; (iv) prévoir les impacts d'interventions (construction d'états contrefactuels associés aux interventions) jamais mises en œuvre dans divers contextes, y compris leurs impacts en termes de bien-être.

La portée des RCT se réduisant essentiellement au premier objectif, ce serait au nom même des exigences de l'EPP qu'on ne pourrait s'en satisfaire. Là où les *randomistas* proposaient de strier les enjeux de l'EPP, en séparant la mesure de l'impact (combien ?) de sa compréhension (comment ?), les *estructuristas* promeuvent une EPP entièrement intégrée à la production de connaissances, qui articulerait tous les enjeux : quantifier, comprendre, prévoir.

I.3. UN RETOUR DE CONFIRMATIONNISME

Difficile pourtant de prendre tout-à-fait au sérieux ces promesses merveilleuses d'une exploitation des données étroitement structurée par la théorie, débouchant sur une EPP toute puissante. C'est que la théorie sous-tendant ces promesses n'est pas n'importe laquelle. Les *estructuristas* ont beau s'efforcer de dissocier leur programme de celui de la microéconomie standard[21], leurs évaluations de politiques de l'emploi finissent généralement par invoquer quelque élasticité de substitution (consommation-loisir, capital-travail, intertemporelle…). Réflexion théorique certes mais dans un cadre un peu restreint. Il n'est d'ailleurs pas tout à fait juste de reprocher aux *randomistas* une absence de préoccupation théorique explicite et d'achopper sur le second point de la glorieuse liste précédente : la théorie fait souvent son retour à l'étape de l'interprétation de l'effet causal moyen… dans des termes pas si différents de ceux des *estructuristas*. Que ce soit pour spécifier un modèle économétrique ou pour interpréter un effet moyen, la microéconomie standard est de rigueur.

Cela tient au caractère abstrait du narratif qu'elle propose. Interpréter un effet moyen c'est faire le narratif d'évènements qui, le plus souvent, ne se sont produits pour aucune des personnes soumises à l'observation[22]. Circonstancier de tels évènements ne peut alors relever que de l'imagination, par exemple en reconstruisant le monde social en termes d'arbitrages rationnels désencastrés. En l'absence d'observations ou d'hypothèses anthropologiques réalistes, il est en effet toujours possible de définir un système de coûts et de bénéfices permettant d'interpréter l'effet causal moyen comme le produit de choix rationnels[23]. Derrière la controverse entre *randomistas* et *estructuristas* s'affirment finalement les mêmes représentations fondées sur des arbitrages coûts-bénéfices abstraits définis pour rendre compte de comportements synthétiques. On est alors porté à penser que l'effet d'une intervention a tenu à des incitations (des prix) et à des paramètres fondamentaux (préférences, technologies) que

[19] Ce n'est en fait pas le cas lorsque l'inconvénient d'un biais est compensé par une estimation plus précise (Ravaillon, 2020, p. 10).

[20] Notre traduction.

[21] Du moins de la théorie du choix rationnel (Heckman et al., 2024, p. 4, note de bas de page).

[22] Contrairement à la formule du *parangon* à laquelle recourt la statistique exploratoire.

[23] Certains auteurs insistent d'ailleurs la fonction performative (plutôt qu'explicative) de l'analyse coût-bénéfice des comportements sociaux (Callon et al., 1997).

l'on s'attend à voir jouer identiquement lors du « passage à l'échelle » ou d'un changement de contextes[24] (puisque ceux-ci doivent être retirés des fondamentaux de l'explication). Les marques de ces représentations fabriquées sont nombreuses dans les politiques du marché du travail menées ces vingt dernières années (Baguelin et al., 2023, p. 19).

Le problème n'est donc pas seulement d'inscrire l'EPP dans une réflexion théorique. C'est aussi de définir le type de théorie utile et les modalités de sa mobilisation. La proposition du présent développement est qu'une théorie compréhensive (*interpretive*), ancrée (*grounded*) dans une enquête préalable (Glaser et al., 1967), ainsi qu'une réflexion explicite sur l'effet du contexte sont souhaitables. C'est peu ou prou ce que propose l'évaluation basée sur la théorie.

II. LA DÉMONSTRATION BASÉE SUR LA THÉORIE

L'évaluation basée sur la théorie (*theory-based evaluation*, TBE – Weiss, 1997) inscrit l'ensemble de la démarche évaluative dans une réflexion préalable sur les processus animant la situation sociale à laquelle s'applique l'intervention. Trois des exercices d'évaluation du numéro s'y réfère explicitement : Devaux-Spatarakis et al., Beaugendre et al., Farvaque et al. La TBE rejoint en un sens la posture évaluative des *estructuristas* selon laquelle caractériser une relation causale nécessite de l'inscrire dans un ensemble de déterminations collatérales… mais sans exclure les effets de contexte (ce qui place souvent la TBE dans une perspective dite d'évaluation réaliste), ni figer les comportements dans le calcul coûts-bénéfices. En TBE, le premier moment de l'évaluation, avant même d'en définir la méthode (qui peut être qualitative, quantitative ou mixte), est l'explicitation d'un référentiel – diversement désigné (et délimité) dans la littérature[25] – que l'on appellera ici *théorie de l'intervention*[26].

II.1. THÉORIE DE L'INTERVENTION ET DIAGRAMME LOGIQUE D'IMPACT

La théorie de l'intervention et son pendant graphique, le « diagramme logique d'impact » (DLI), constituent une représentation *partagée* (entre parties prenantes) de la situation sociale à laquelle s'applique l'intervention et des modifications qu'en attend sa maîtrise d'ouvrage. L'élaboration de cette théorie s'appuie sur une revue de la littérature et implique les parties prenantes (Devaux-Spatarakis et al., dans le présent numéro) ; c'est donc une combinaison de SHS, de dires d'experts et d'intuitions des concepteurs de l'intervention. Lorsque la complexité de celle-ci ou la diversité des cas ne permettent pas de synthétiser théoriquement les processus en jeu, la formulation de la théorie de l'intervention peut nécessiter une expérience pilote (preuve de concept, Baguelin et al., 2023) ou une enquête ethnographique.

À propos de la théorie de l'intervention, certains parlent de théorie « avec un petit t » (Chen et al., 1987) pour en suggérer le caractère situé et centré sur des causes proximales. Rien à voir, donc, avec la « Théorie générale » d'une société de marché. La théorie de l'intervention a une vocation pratique : fixer des questions évaluatives précises (et leur degré de priorité) ; structurer l'enquête évaluative ; définir des variables opérationnelles de résultat, de processus et de contexte ; formuler des hypothèses causales clarifiant comment et pourquoi des résultats

[24] La notion de *contexte* à laquelle se réfèrent généralement les économistes renvoie à un donné, un « déjà-là ». La philosophie pragmatique lui oppose un concept d'*environnement* (Cf. *infra*).

[25] Théorie de/du changement, théorie de/du programme, théorie de l'action, modèle logique, logique de l'intervention (Rey *et al.*, 2022).

[26] Conformément à la proposition francophone de Marceau et al. (2022).

pourraient être obtenus. C'est là sa fonction essentielle : expliciter une série d'évènements conduisant à un résultat (Funnell et al., 2011, p. 13), énoncer, dans une logique causale, « les conditions et les hypothèses nécessaires pour que des changements se produisent » (Gertler et al., 2011, p. 22).

L'élaboration d'une théorie de l'intervention peut faire l'objet d'une démarche systématique (Mejia et al., 2022, p. 195)[27]. On explicite d'abord les ressources, *activités*, produits et résultats de l'intervention. Les activités renvoient aux opérations (formulées en verbes d'action) transformant les ressources en produits (biens ou services mis à disposition du public cible). Les résultats sont les influences exercées sur le public cible en termes de capacités, connaissances, aptitudes… La théorie s'élabore alors de façon récursive : définition du public cible puis des résultats puis des produits, des activités et enfin des ressources. La distinction entre produits et résultats correspond à la frontière entre ce qui est sous le contrôle de la maîtrise d'ouvrage (produits) et ce qui traduira une éventuelle influence (résultats) pouvant impliquer l'action de facteurs externes. Ceux-ci relèvent en particulier du contexte de l'intervention : ils font l'objet d'une réflexion spécifique et apparaissent dans le DLI là où s'exerce leur action présumée.

L'élaboration de la théorie de l'intervention se poursuit en explicitant des hypothèses (et risques) de l'intervention. Les hypothèses portent sur la manière dont celle-ci est censée susciter les résultats ; elles constituent une liste de causalités et conditions des résultats. Quatre classes d'hypothèses sont parfois distinguées (Mayne, 2017, p. 176) : (i) de portée (conditions nécessaires pour que les produits découlent des activités) ; (ii) de changements en capacité (conditions nécessaires de résultats en termes de connaissances), en comportement (conditions nécessaires de résultats en termes de comportements) et en bien-être (conditions nécessaires de résultats en termes de conditions de vie). Ces conditions s'inscrivent dans la théorie de l'intervention à l'aide de clauses « si… alors… » sous la forme générique : sous l'hypothèse $H_{\text{portée}}$, si l'activité A est mise en œuvre alors Y_0 est produit ; sous l'hypothèse $H_{\text{capa.}}$, si Y_0 a été produit alors le résultat Y_1 en termes de connaissances se réalise ; sous l'hypothèse $H_{\text{comp.}}$, si Y_1 s'est réalisé alors le résultat Y_2 en termes de comportement se réalise ; etc. Les risques de l'intervention correspondent alors à l'invalidité de l'une ou l'autre de ces hypothèses.

Trois des exercices d'évaluation du numéro partent d'une théorie d'intervention explicite : c'est devenu un procédé standard, même si les règles listées ci-dessus peuvent être suivies plus ou moins rigoureusement. La suite de la présente introduction entend montrer que cet outil a une portée dépassant sa fonction opérationnelle. Cette démonstration commence par un détour par la diagrammatique.

II.2. DIAGRAMME, ALGORITHME ET RÉSEAU BAYÉSIEN

En représentant méthodiquement une théorie partagée, un DLI conforme aux exigences précédentes a une portée heuristique considérable[28]. On peut le montrer en en rapprochant le principe de ceux de deux formalismes graphiques élaborés indépendamment : l'algorigramme de politique publique (Baïz et al., 2018) et le graphe probabiliste (Pearl, 2000).

---

[27] Des applications permettent l'élaboration de DLI : https://www.betterevaluation.org/tools-resources/theory-change-software (consulté fin 2024).

[28] L'attention accordée aux diagrammes par la TBE ne doit pas surprendre : « le diagramme ne représente pas son objet mais il le construit au sens où il montre ses relations constitutives ; ce faisant, il formule une hypothèse sur lui, ce qui le rend du même coup capable de révéler une vérité inattendue le concernant. L'opérativité du diagramme est étroitement liée à son organisation topologique qui donne à voir des hypothèses inscrites graphiquement dans son réseau de lignes, hypothèses qui s'offrent à l'observation, à la manipulation, voire à l'expérimentation » (Dahan-Gaida, 2023).

Le premier de ces formalismes est destiné à décrire avec précision les interventions publiques (l'article d'Annie Jolivet dans le présent numéro en illustre l'importance à propos d'une incitation à négocier). Il s'agit d'un langage diagrammatique destiné à caractériser de façon univoque tout *instrument* de politique publique, entendu comme « chaîne de causalité préfigurée d'une action collective » (Baïz et al., 2018, p. 157). Ce langage repose sur une liste d'éléments nécessaires et suffisants : trois variables-types (acteurs, actions et « conditions »), six vecteurs d'impact et cinq opérateurs. Les « conditions » sont des circonstances conditionnant le développement de la chaîne causale. Les vecteurs d'impact décrivent l'influence exercée par la réalisation d'actions, influence déterministe (rendre certain/impossible) ou probabiliste (rendre plus/moins probable, rendre possible, rendre facultatif). Les opérateurs sont empruntés à l'algorithmique : « si... alors... / sinon » ; « tant que... faire... » ; « faire... jusqu'à... » ; « et, ou inclusif, ou exclusif » ; « non ». Un instrument de politique publique articule ces éléments selon trois règles de conception : la complétude, la cohérence et la connexité[29]. Ce langage permet de caractériser, avec un degré quelconque de précision, n'importe quel instrument. En supprimant toute ambigüité dans leur définition, l'algorigramme obtenu instille clarté et transparence dans l'action publique, favorise la conception de nouveaux dispositifs, et… permet de guider l'évaluation. Cela tient à ce qu'il confère une certaine épaisseur à la *démarche évaluative* elle-même (par différence à la simple considération des méthodes[30]).

Le second formalisme sert à représenter la structure des dépendances probabilistes d'un jeu de variables aléatoires. Chaque nœud d'un graphe probabiliste est associé à une variable ; tout arc (ou arête) reliant deux nœuds traduit une dépendance probabiliste entre les variables correspondantes. Les graphes probabilistes peuvent être orientés ou non ; les arcs orientés sont des flèches indiquant le sens d'une causalité. Les graphes probabilistes orientés débouchent, en l'absence de cycle (de rétro-causalité indirecte), sur le concept de réseau bayésien (Pearl, 1985). Un graphe orienté acyclique (ou DAG pour directed acyclical graph) représente une croyance qualitative sur la structure causale du jeu de variables considéré : quelles variables sont présumées exogènes (i.e. déterminées hors du jeu de variables) ? Pour chaque variable présumée endogène, quelles seraient ses causes directes ? ses causes indirectes ? Quelles seraient les variables indépendantes en probabilité ? De façon moins évidente, l'utilité de ces graphes tient surtout au fait qu'expliciter une structure causale c'est, incidemment, faire apparaître les dépendances non causales du jeu de variables. À la notion purement qualitative de DAG, le réseau bayésien ajoute une dimension quantitative. En présence d'une distribution jointe pour le jeu de variables considéré, le graphe constitue une carte de calcul des probabilités conditionnelles qui permet notamment, en présence de variables réalisées, d'estimer la probabilité des variables qui ne le sont pas, selon le principe de la révision bayésienne.

Conçus comme objets adaptatifs, ces graphes forment le support visuel, donc aisément partageable, adéquat à un tâtonnement bayésien consistant en allers-retours entre théorie (croyance) et collecte de données. L'assimilation du DLI à un DAG est en ce sens immédiate (Powell, 2018) et la TBE s'interprète aisément comme application de ce principe à l'enquête évaluative. Le DLI est un DAG qui initialise l'enquête évaluative. Celle-ci pourra alors se déployer comme un tâtonnement bayésien indifféremment qualitatif (caractérisation de causalité) et/ou quantitatif (mesure de dépendance causale). Souligner la proximité conceptuelle

[29] La règle de complétude pose que la chaîne causale comporte les éléments nécessaires à sa compréhension (au moins un acteur, une action par acteur et un vecteur d'impact par action). La règle de cohérence impose que la chaîne causale ne comporte pas de contre-sens logique. Enfin, la règle de connexité s'énonce ainsi : « Toute condition et tout couple acteur-action doit pouvoir être relié, directement ou non, à tout autre condition ou autre couple acteur-action de la chaîne de causalité, à travers les vecteurs d'impact et les opérateurs logiques que la chaîne mobilise » (ibid., p. 162).

[30] Cf. encadré 2 et contribution de Solveig Grimault dans ce numéro.

du DLI avec l'algorigramme et le DAG, c'est donc définir le statut de la théorie dans l'exercice d'évaluation : le DLI permet d'orienter la production de connaissance et de mettre en œuvre un apprentissage bayésien (encadré 2). Dans la controverse entre *estructuristas* et *randomistas*, la proximité entre DLI et DAG permet à l'évaluation basée sur la théorie de se recommander de la synthèse procurée par l'apprentissage statistique (Pearl, 2000), moyen d'aborder la démonstration d'impact avec les ambitions des *estructuristas* et la prudence des *randomistas*.

III. L'INFÉRENCE CAUSALE

Dans le cadre d'une TBE, la démonstration d'impact n'a ainsi pas à se cantonner aux perspectives théoriques étroites des *estructuristas*. TBE et réseaux bayésiens offrent un cadre dans lequel mettre en communication l'ensemble des approches mobilisées par l'EPP : l'étude monographique (*single case study*), l'étude comparative (*multiple case study*), l'étude statistique. Pour s'en convaincre, il faut cependant expliciter l'articulation des concepts propres à chacune : le propos de la présente introduction n'est en effet pas d'en plaider l'identité (Beach, 2019) mais d'en opérationnaliser la complémentarité (Bamberger, 2012). L'inférence causale est le lieu à partir duquel établir cette articulation (Johnson et al., 2019). Qu'elle mobilise des méthodes qualitative ou quantitative, l'EPP d'intention scientifique ne peut en effet contourner cette exigence.

III.1. CAUSALITÉ ET CAUSATION

En la matière, la démonstration d'impact se réfère de façon privilégiée à la conception contrefactuelle (Lewis, 1973) : A cause Y si une modification contrefactuelle de la valeur de A dans un monde dont le passé et les déterminismes seraient ceux du monde factuel, modifie la valeur de Y. Pour revenir à l'exemple de l'accompagnement des personnes en demande d'emploi[31], l'accompagnement cause l'accès à l'emploi si, à la seconde où cet accompagnement a factuellement débuté, quelque chose l'a empêché[32] (quelque chose qui n'a modifié le factuel d'aucune autre façon) et que l'accès à l'emploi n'a pas lieu. La conception contrefactuelle intègre adéquatement l'asymétrie causale voulant que le futur, non le passé, dépend du présent. Cette référence est réputée la plus ouverte à la diversité des approches (Johnson et al., 2019 ; Rohlfing et al., 2021). De fait, si le concept a été élaboré à propos de causalités singulières (« l'accompagnement dont a bénéficié Bernard lui a permis d'accéder à l'emploi »), il s'avère adapté aux causalités générales auxquelles s'intéresse l'évaluation quantitative (« l'accompagnement accroît l'accès à l'emploi »). On lui reproche pourtant typiquement trois choses : un caractère métaphysique (Dawid, 2000) ; d'être difficile à opérationnaliser (Blanchard, 2018) ; une sophistication décalée au regard de la banalité des expériences causales (Huneman, 2020).

Il existe une variété d'autres conceptions de la causalité empruntant alternativement à la physique, à la logique ou aux probabilités. Le problème que cette variété pose à la métaphysique (Blanchard, 2018) ne concerne pas vraiment les SHS (Johnson et al., 2019)[33]. Pour celles-ci, il peut même s'agir d'un atout pour articuler les enseignements d'enquêtes empiriques opérant de différents points de vue ou à différentes échelles : celle du cas, du petit nombre de cas, de la population. De fait, méthodes quantitative et qualitative se distinguent moins par des espèces

[31] Et dans une perspective de « treatment effect on the treated ».

[32] Et donc empêché la réalisation de l'ensemble de ses effets « propres ».

[33] Pour un avis plus nuancé issu de la *Comparative politics* : Beach (2019).

différentes de causalité[34] que par des stratégies d'inférence différentes. L'étude monographique (*single case study*) infère la causalité de la causation en décrivant directement, en contexte et à partir d'hypothèses anthropologiques réalistes (Rao et Woolcock, 2003, p. 172), un processus causal (conception processuelle). L'étude statistique infère indirectement la causalité de la dépendance probabiliste (conception probabiliste). Pour la démonstration d'impact, études statistique et monographique ne sont pas concurrentes (Johnson et al., 2019, p. 152). La causation établit la causalité mais, dans le monde social, un processus causal est toujours singulier ; la dépendance probabiliste définit un macro-concept de causalité utile pour penser l'impact à l'échelle d'une population, mais reste incapable d'inférer la causation[35]. L'inférence causale dans l'étude comparative (*multiple case study*), qui procède en termes de condition logique, établit un pont permettant d'articuler « macro-causalité » et causation (Mahoney, 2008).

III.2. UNE PLURALITÉ DE CONCEPTIONS… NON CONCURRENTES

Trois conceptions de la causalité irriguent les SHS dont on montre ici que l'une, la conception logique, met en communication les deux autres : les conceptions probabilistes et processuelles.

Conception processuelle

La conception processuelle[36] infère la causalité de la causation. Deux critères essentiels sont attachés à la causation : la continuité d'un processus et le transfert de propriété de la cause dans l'effet (Kistler, 1998). En SHS, la continuité processuelle renvoie au concept de *mécanisme* (Glennan, 2009) entendu comme « système composé de différentes parties interagissant pour produire de manière régulière un certain phénomène » (Blanchard, 2018). L'interaction étant elle-même une notion causale, le mécanisme est un ensemble de « lois causales » gigognes allant des plus fondamentales (lois physiques) aux plus composites (lois psychologiques). Le mécanisme correspond au niveau minimal permettant de fonder la causation, il convoque un déterminisme (Mahoney, 2008, p. 420). Par exemple, le mécanisme de l'accompagnement peut tenir dans la fixation de rendez-vous réguliers, la tenue d'un journal de recherche d'emploi, la définition d'une méthode de prospection… Le critère de transfert de propriété signifie que la cause injecte une propriété dans l'effet : toujours en matière d'accompagnement, on peut imaginer la transmission d'une information qui oriente la recherche d'emploi, le soutien psychologique qui permet de solliciter une capacité d'agir, l'octroi d'une certification qui améliore un signal sur le marché du travail... On voit que la description d'un mécanisme et le repérage d'un transfert de propriété passe par l'observation fine de l'intervention et ne peut se satisfaire de l'intention formulée par ses concepteurs. En médecine, l'établissement de la causation procède aussi de l'examen clinique, la détection de symptômes. En SHS, l'exercice évoque le « paradigme indiciaire » où l'on part des indices présents dans l'effet pour remonter à la cause (Ginzburg, 1980). L'interrogation d'une personne sur les circonstances de sa sortie du chômage peut l'amener à mentionner des éléments de l'accompagnement dont elle a bénéficié : « j'ai eu connaissance de l'opportunité par… », « j'ai eu l'audace de candidater grâce au soutien de… », « disposer d'un certificat de qualification m'a permis d'accéder à un entretien ». L'approche compréhensive est spécialement requise :

[34] Dans les deux cas, savoir si la causalité est une propriété du réel n'est pas vraiment l'enjeu : il suffit de tenir la causalité pour une catégorie de l'entendement humain, un concept utile pour ordonner les données de l'observation.

[35] D'où la tentation de l'inventer (interprétation controuvée) selon les catégories analytiques de la microéconomie standard.

[36] Certains parlent de causalité générative et/ou productive pour insister sur le rôle des intentions des acteurs (Gagnon, 1975).

respect de la diversité des types d'action (rationnelles en but ou en valeur, affectuelles, habituelles – Weber, 1921, 2019), cohérence narrative, expérience située, etc.

Conception probabiliste

Le questionnement causal en matière d'EPP prend la forme d'énoncés généraux du type « l'accompagnement cause-t-il l'accès à l'emploi ? ». Plutôt qu'à l'échelle du cas, ce type d'énoncé s'entend à celle d'une population. L'énoncé pourrait toutefois recevoir une interprétation binaire du type « l'accompagnement cause toujours/jamais l'accès à l'emploi ». Ce serait très restrictif au regard des limites de tout dispositif d'observation de population : erreurs de mesure, hétérogénéité inobservée, interdépendances insoupçonnées… La question sera donc plutôt interprétée comme « l'accompagnement cause-t-il plus souvent l'accès à l'emploi ? », c'est-à-dire en termes probabilistes. La causalité probabiliste est conceptualisée par Reichenbach (1956) : partant d'un monde assimilable à un jeu de variables aléatoires (dichotomiques pour faire simple) liées par une distribution de probabilités, $A = 1$ cause $Y = 1$ si (i) la réalisation de A est antérieure à celle de Y (asymétrie), (ii) la probabilité de $Y = 1$ sachant $A = 1$ est supérieure à celle de $Y = 1$ sachant $A = 0$, (iii) A et Y n'ont pas de cause commune. Le concept de Reichenbach se systématise en associant au jeu de variables aléatoires évoqué une structure causale représentée par… un réseau bayésien (Blanchard, 2018)[37].

Conception logique

La réduction de la causalité à un critère logique (méthodes de Mill, 1843) correspond au mode d'inférence causale de l'étude comparative (*multiple case study*). On a vu que cette conception est aussi centrale en TBE avec le DLI. Moyennant une formulation adaptée aux SHS, cette réduction dresse surtout un pont entre conceptions probabiliste et processuelle de l'inférence causale (Mahoney, 2008).

La cause peut d'abord être définie comme condition nécessaire et suffisante (NS) de l'effet. Dans l'exemple de l'accompagnement, celui-ci sera tenu pour cause d'accès à l'emploi si : (N) l'accès à l'emploi implique l'accompagnement[38] et (S) l'accompagnement implique l'accès à l'emploi. Empiriquement, un cas d'accès à l'emploi sans accompagnement ou d'accompagnement sans accès à l'emploi suffit à conclure que l'accompagnement n'est pas cause d'accès à l'emploi. La condition NS ramène donc à l'embarras suscité par une interprétation binaire (toujours/jamais) de la causalité. Lorsque, mis à part celui amenant à réfuter la condition NS, on n'a que des cas de non-emploi sans accompagnement ou d'accès à l'emploi avec accompagnement, difficile de renoncer à l'interprétation causale. On peut donc affaiblir le critère en définissant la cause comme condition nécessaire ou (inclusif) suffisante. Dans l'enquête empirique, la réfutation sera un peu plus exigeante puisqu'elle réclamera au moins deux cas : l'un avec accès à l'emploi sans accompagnement, le second avec accompagnement sans accès à l'emploi. Mais dans le cadre d'une enquête de terrain, le critère apparaît encore excessivement restrictif.

Opérationnaliser la réduction logique de la causalité amène donc à reconnaître que les causes n'agissent qu'exceptionnellement de façon isolée. Pour inférer la causalité, l'étude comparative sépare donc ce qu'il y a de commun à tous les cas, de facteurs prenant des valeurs

[37] Le critère *(iii)* débouche notamment sur la *condition causale de Markov* (Geiger et al., 1990) : à causes directes de A données, A est indépendant de toutes les variables autres que ses effets. Cette condition joue comme un détecteur de relations de cause à effet et peut donc servir à la démonstration du caractère causal du lien entre deux variables : sous certaines hypothèses, l'application systématique de la condition causale de Markov permet même d'inférer entièrement la structure causale (au sens probabiliste) du jeu de variable (Verma et al., 1990).

[38] Sans accompagnement pas d'accès à l'emploi.

qui varient d'un cas à l'autre. L'effet éventuel de l'un de ces facteurs ne peut s'entendre qu'en interaction avec les autres, ce que doit traduire le critère logique appliqué. C'est ce que propose la condition INUS pour « Insufficient but Necessary part of a condition itself Unnecessary but Sufficient » (Mackie, 1965). Cette condition est remarquable : elle correspond à la manière dont opère la fonction de régression dans le cas d'une spécification saturée qui, elle-même, ajuste exactement la fonction d'espérance conditionnelle (Angrist et al., 2009, p. 49). Il s'avère qu'en procédant à l'inférence causale par caractérisation de conditions INUS, les praticiens de l'étude comparative sur quelques cas (*small n multiple case study*) réalisent la même opération que l'espérance conditionnelle. L'encadré 1 pose formellement la relation entre les deux.

La proposition unificatrice de la présente introduction n'est donc pas un mirage : pour autant qu'ils portent sur un même objet, les enseignements d'études monographique, comparative et statistique, peuvent (et doivent ?) s'articuler. En matière d'évaluation, la TBE offre la démarche dans laquelle réaliser cette articulation mais l'opérationnaliser pleinement suppose d'être au clair sur la répartition des rôles entre formes d'enquête, notamment entre approches quanti- et qualitative. La pleine exploitation de leur complémentarité suppose de respecter le « fonctionnement » de chacune et de lui appliquer des exigences adéquates. En évaluation des politiques de l'emploi, cette condition n'est pas toujours satisfaite, notamment concernant l'enquête qualitative.

1. Le lien formel entre conceptions logique et probabiliste de la causalité

Le lien entre les conceptions logique et probabiliste de la causalité peut être illustré dans l'exemple de la relation entre accompagnement ($A \in \{0; 1\}$) et accès à l'emploi ($Y \in \{0; 1\}$) en considérant deux autres facteurs d'accès à l'emploi : l'absence de discrimination ($W \in \{0; 1\}$) et la (bonne) santé ($X \in \{0; 1\}$). La spécification dichotomique permet de formuler simplement les différents critères logiques d'inférence causale.

Pour une condition NS, l'accompagnement cause l'accès à l'emploi si et seulement si $A = Y$. Les autres facteurs d'accès à l'emploi ne jouent aucun rôle. Alternativement, imaginons que l'accompagnement soit nécessaire pour qu'une personne discriminée ($W = 0$) accède à l'emploi mais pas pour les autres. L'accompagnement est alors une condition suffisante (SU pour *Sufficient Unnecessary*) d'accès à l'emploi et le critère peut s'écrit $\max(A, W) = Y$. Peut-être, au contraire, l'accompagnement ne cause-t-il l'accès à l'emploi qu'en l'absence de discrimination ($W = 1$) : c'est une condition nécessaire non suffisante (NI pour *Necessary Insufficient*) de l'accès à l'emploi et le critère s'écrit $\min(A, W) = Y$.

Sous les deux dernières hypothèses, le terme W peut dissimuler une influence causale de A. Selon un principe analogue, cette influence peut exister pour $A = 1$ ni nécessaire ni suffisant à $Y = 1$. L'effet de l'accompagnement peut s'articuler avec l'état de santé. Première possibilité, l'accompagnement est une condition suffisante d'accès à l'emploi pour les personnes en bonne santé ($X = 1$) mais pas pour les autres : on a alors affaire à une condition SUIN (pour *Sufficient but Unnecessary part of a condition itself Insufficient but Necessary*) dont le critère s'écrit $\min\left(\max(A, W), X\right) = Y$. Seconde possibilité, l'accompagnement est une condition nécessaire d'accès à l'emploi pour les personnes en mauvaise santé ($X = 0$), les autres n'en ayant pas besoin. C'est la condition INUS dont le critère s'écrit $\max\left(\min(A, W), X\right) = Y$.

L'analyse des données à partir de cette variété de conditions logiques raffine l'enquête évaluative qui peut conclure diversement au caractère NS, NI, SU, SUIN ou INUS de l'accompagnement dans l'accès à l'emploi. Cette conception de l'inférence causale permet

surtout de relier enquêtes comparative (*small-n*) et statistique. Pour un jeu de données $(A_i, W_i, X_i, Y_i)_{i\in I}$, ajuster la spécification saturée

$$Y_i = \alpha + \beta_A A_i + \beta_W W_i + \beta_X X_i + \gamma_{AW} A_i W_i + \gamma_{AX} A_i X_i + \gamma_{WX} W_i X_i + \delta_{AWX} A_i W_i X_i + \varepsilon_i$$

permet de détecter le type de condition logique reliant accompagnement et accès à l'emploi selon le tableau de correspondances suivant :

$$\left(\hat{\beta}_A = 1\right) \wedge \left(\hat{\beta}_W = 0\right) \wedge \left(\hat{\beta}_X = 0\right) \wedge \left(\hat{\gamma}_{AW} = 0\right) \wedge \left(\hat{\gamma}_{AX} = 0\right) \wedge \left(\hat{\gamma}_{WX} = 0\right) \wedge \left(\hat{\delta}_{AWX} = 0\right) \rightarrow NS$$

$$\left(\hat{\beta}_A = 1\right) \wedge \left(\hat{\beta}_W = 1\right) \wedge \left(\hat{\beta}_X = 0\right) \wedge \left(\hat{\gamma}_{AW} = -1\right) \wedge \left(\hat{\gamma}_{AX} = 0\right) \wedge \left(\hat{\gamma}_{WX} = 0\right) \wedge \left(\hat{\delta}_{AWX} = 0\right) \rightarrow SU$$

$$\left(\hat{\beta}_A = 0\right) \wedge \left(\hat{\beta}_W = 0\right) \wedge \left(\hat{\beta}_X = 0\right) \wedge \left(\hat{\gamma}_{AW} = 1\right) \wedge \left(\hat{\gamma}_{AX} = 0\right) \wedge \left(\hat{\gamma}_{WX} = 0\right) \wedge \left(\hat{\delta}_{AWX} = 0\right) \rightarrow NI$$

$$\left(\hat{\beta}_A = 0\right) \wedge \left(\hat{\beta}_W = 0\right) \wedge \left(\hat{\beta}_X = 0\right) \wedge \left(\hat{\gamma}_{AW} = 0\right) \wedge \left(\hat{\gamma}_{AX} = 1\right) \wedge \left(\hat{\gamma}_{WX} = 1\right) \wedge \left(\hat{\delta}_{AWX} = -1\right) \rightarrow SUIN$$

$$\left(\hat{\beta}_A = 0\right) \wedge \left(\hat{\beta}_W = 0\right) \wedge \left(\hat{\beta}_X = 1\right) \wedge \left(\hat{\gamma}_{AW} = 1\right) \wedge \left(\hat{\gamma}_{AX} = 0\right) \wedge \left(\hat{\gamma}_{WX} = 0\right) \wedge \left(\hat{\delta}_{AWX} = -1\right) \rightarrow INUS$$

Cela signifie qu'une étude comparative permet de motiver un choix de spécification de modèle (gain d'efficacité statistique) ou, réciproquement, que l'estimation de modèles statistiques saturés permet de guider l'enquête qualitative.

IV. LA DÉMONSTRATION QUALITATIVE D'IMPACT

Il est fréquent que le choix d'une « évaluation qualitative » soit considéré comme un pis-aller. Cela se ressent notamment dans la façon dont la commande publique tend à faire peser sur les projets de monographie évaluative des attentes empruntant aux critères parfaitement hétéronomes de la statistique[39] : langage de variables déterminées avant l'enquête ; contrainte de représentativité pour le choix des terrains ; notion d'échantillon et de biais ; conception purement « formulaire » de l'enquête par entretien, etc. Une fois les terrains déterminés, il n'est pas rare non plus que la commande publique fasse montre d'un certain laxisme quant à ce que la production monographique satisfasse in fine aux exigences générales d'une évaluation (pour le coup assorti de critères) en termes de reconstitution du chemin causal (Grimault, dans ce numéro). Un processus d'évaluation et son schéma d'ordonnancement (encadré 2) ont pourtant une consistance propre qui soumet nécessairement l'enquête, qu'elle soit quanti- ou qualitative, à des exigences. Celles-ci concernent une délibération préalable sur les critères de l'évaluation mais surtout la prise en compte ex ante des résultats projetés d'une intervention (théorie de l'intervention).

IV.1. LA MISE EN CAS DES CAUSES

Il ne faut pas « confondre monographie [narrative] et cas explicitement choisis et analysés aux fins d'une démonstration » (Passeron et al., 2020, p. 9). Dans une perspective de

[39] « Comme si, à budget d'enquête donné, il s'agissait de choisir entre une connaissance intensive (beaucoup d'informations sur un petit nombre de cas) et une autre extensive (peu d'information mais sur un grand nombre de cas) [laissant] croire que les deux types d'enregistrement sont de même nature, et que si, par exemple, les budgets d'enquête étaient par magie multipliés, ce choix angoissant pourrait être évité » (Desrosières, 2008, chap. 8).

démonstration d'impact, l'observation est préordonnée par les questions évaluatives que rassemble le DLI : ces questions déterminent la « mise en cas » (*casing*) du réel (Ragin, 2018). Cette opération engage l'ensemble des actes par lesquels un agencement de faits[40] (et de normes) est constitué en cas d'un *mécanisme*. Le cas se constitue autour de la notion d'*activité* (supra) qui joue dans l'ordre qualitatif le rôle de la notion de traitement dans l'ordre quantitatif. La mise en cas doit répondre à trois questions : de quoi tel agencement de faits est-il le cas ? de quelles variables[41] ce cas se compose-t-il ? que peut produire le cas en question ? (Ragin et al., 1992). Si le plan d'évaluation permet de structurer la réponse à la première question, l'analyse de causation répond normalement à l'examen de ce que produisent les activités constituant le cas. Cet examen part lui-même de la composition du cas (deuxième question). Relativement à une enquête monographique dépourvue d'intention démonstrative, la composition du cas découle en TBE de la théorie de l'intervention. C'est cette théorie qui institue ex ante les entités/catégories composant le cas et qui dirige l'attention que l'on accordera à certains éléments du réel, distinguant ces derniers comme classe d'objets pertinents et les érigeant au statut de variables du cas. L'analyse de causation consiste alors à examiner si les instances observées (configurations de valeurs de variables et résultats) confirment ou non les propositions ou hypothèses de la théorie de l'intervention. L'enquête correspondante amène notamment à repérer des variables « insues » qu'il convient d'ajouter à l'analyse de causation. La mise en cas consiste donc à traduire en « langage de variables » la perspective narrative dont part généralement l'enquête monographique[42].

L'établissement d'une collection de cas pour l'exploration et la qualification de la causalité peut répondre à une perspective comparative (*multiple case study*). Il s'agit alors de disposer de quoi faire « fonctionner » les conditions minimales à la production d'un effet causal (au sens par exemple d'une condition logique) : à effet donné, quels sont les éléments dont la (co-)présence est une condition du résultat ? L'intérêt de l'enquête monographique (*single case study*) est, plus exclusivement, de caractériser directement la causation. L'examen porte alors sur les « processus continus » transmettant « l'influence causale d'une région à l'autre de l'espace-temps » par la reconstitution de « lignes causales » (Fagot-Largeault, 1992 ; Salmon, 1994). En cohérence avec ce clivage entre monographie et analyse comparative, on distingue, selon l'objet et les conditions des interventions considérées, deux groupes de méthodes d'évaluation basée sur la théorie (Stern et al., 2012, p. 24). L'un s'intéresse aux raisons du résultat et aux mécanismes par lesquels l'intervention le détermine : pourquoi et comment l'intervention produit un résultat donné ? Le second s'intéresse à la variabilité du résultat de l'intervention selon la configuration de sa mise en œuvre : dans quelle configuration le résultat est-il le meilleur ?

IV.2. LA DÉMONSTRATION MONOGRAPHIQUE (SINGLE CASE STUDY)

Elle peut emprunter à l'analyse de contribution et/ou à la reconstitution de processus.

L'analyse de contribution consiste à associer à la théorie de l'intervention un récit de contribution (Mayne, 2012) et à le tester itérativement pour en réfuter/consolider les hypothèses. Les questions causales doivent rester peu nombreuses et être formulées le plus

[40] Entendus comme « un ensemble de données prélevées dans le flux continu du réel et organisées en une configuration, en raison même de la signification que lui confère l'observateur » (Prairat, 2014). Désigner un fait sera « souligner l'attention privilégiée que l'on accorde à certains éléments du réel » (ibid.).

[41] Construites à des fins d'analyse et non projetées sur le réel.

[42] On comprend alor que les chercheurs en SHS revendiquent « la substitution partielle de la redescription à l'inférence [causale] » (Rorty, 1993, p. 118) quitte à « dissoudre les problèmes […] plutôt que les résoudre » (*ibid.*, p. 43).

précisément possibles. Le récit de contribution dresse une liste des facteurs présumés contribuer au résultat. Ce récit évolue au fil de la collecte d'éléments de preuve. Ceux-ci peuvent relever de recherches liées aux questions causales d'intérêt, de quantifications ou d'évaluations antérieures ; on applique un principe de triangulation en croisant des indices tirés de sources indépendantes. Tout cela débouche sur la formulation d'une hypothèse sur la contribution de l'intervention au résultat ; cette hypothèse est ensuite testée à partir d'indices supplémentaires. Un récit de contribution est finalement stabilisé en indiquant sa portée et ses limites.

Lorsque l'intervention implique une chaîne causale courte (ne comportant que quelques relations de cause à effet[43]) ou lorsqu'on entend se concentrer sur un point précis de la théorie de l'intervention, la reconstitution de processus (*process tracing*) consiste à décrire minutieusement (dans l'esprit d'une enquête judiciaire) la séquence temporelle d'évènements jalonnant le processus causal (Bezes et al., 2018). Il s'agit de suivre la piste reliant l'intervention au résultat en distinguant des variables intermédiaires (médiateurs) et en collectant les indices laissés par le processus causal. Collier (2011) préconise de procéder de façon itérative à la lumière d'hypothèses explicites : en les révisant selon les indices collectés et en renouvelant la recherche d'indices à partir d'hypothèses révisées. Des tests formels, inspirés de la logique bayésienne, peuvent être mis en œuvre. La reconstitution de processus se distingue d'autres approches par le fait que l'analyse opère à l'échelle des évènements que comporte un cas et non par comparaison de cas distincts (Farvaque et al., dans ce numéro).

IV.3. LA DÉMONSTRATION COMPARATIVE (MULTIPLE CASE STUDY)

La démonstration qualitative d'impact peut alternativement porter sur les facteurs associés aux résultats de l'intervention.

Lorsque l'intervention et son action sont complexes et/ou les facteurs participant au résultat potentiellement nombreux, comprendre la variabilité du résultat obtenu requiert une approche de « tri et examen » (*screening, scoping*) procédant cas par cas. La variété des configurations d'intervention est documentée et réduite à quelques cas-type. Chacun de ces cas-type est examiné selon la même procédure (*scoping*) pour proposer une explication du résultat de l'intervention. On tire de l'analyse des hypothèses quant aux conditions sous lesquelles l'intervention a des résultats satisfaisants.

Lorsque certains facteurs pouvant déterminer le résultat sont connus et qu'ils ne sont pas trop nombreux, il est utile que l'enquête évaluative en enregistre systématiquement les valeurs. Cela passe par une phase de « mise en cas » (*casing*) pour constituer une collection de cas suffisamment homogènes. L'analyse quali-comparative[44] (Qualitative comparative analysis – Ragin, 1987) consiste à étudier les configurations de valeurs associées à un résultat plus ou moins satisfaisant. Les prédicteurs considérés peuvent tenir aux caractéristiques des personnes ou aux modalités de mise en œuvre de l'intervention. Il s'agit, dans l'esprit de l'encadré 1, de caractériser les conditions nécessaires et/ou suffisantes d'un résultat satisfaisant. Une telle approche a été appliquée à la relation entre formation et/ou accompagnement et accès à l'emploi (Duclos, 2017). De façon générale, quatre groupes de facteurs peuvent être distingués selon leur association au résultat : les facteurs ni nécessaires ni suffisants ; les facteurs nécessaires mais pas suffisants ; les facteurs suffisants mais pas nécessaires ; les facteurs nécessaires et suffisants. Ces derniers sont parfois assimilés à des causes du résultat. Passé le seuil de la trentaine de cas, le rôle de ces configurations de facteurs dans la détermination du résultat, peut

[43] Ce qui permet notamment l'application du critère INUS.

[44] Ou quali-quanti comparée.

faire l'objet d'une analyse de variance et/ou d'un ajustement par régression pour quantifier les relations causales présumées[45].

2. Schéma général d'ordonnancement d'une démarche évaluative

Tout programme d'évaluation opère selon trois plans articulés, chacun établissant un niveau de connaissance spécifique : un plan logique, un plan d'analyse causale et un plan d'effectuation.

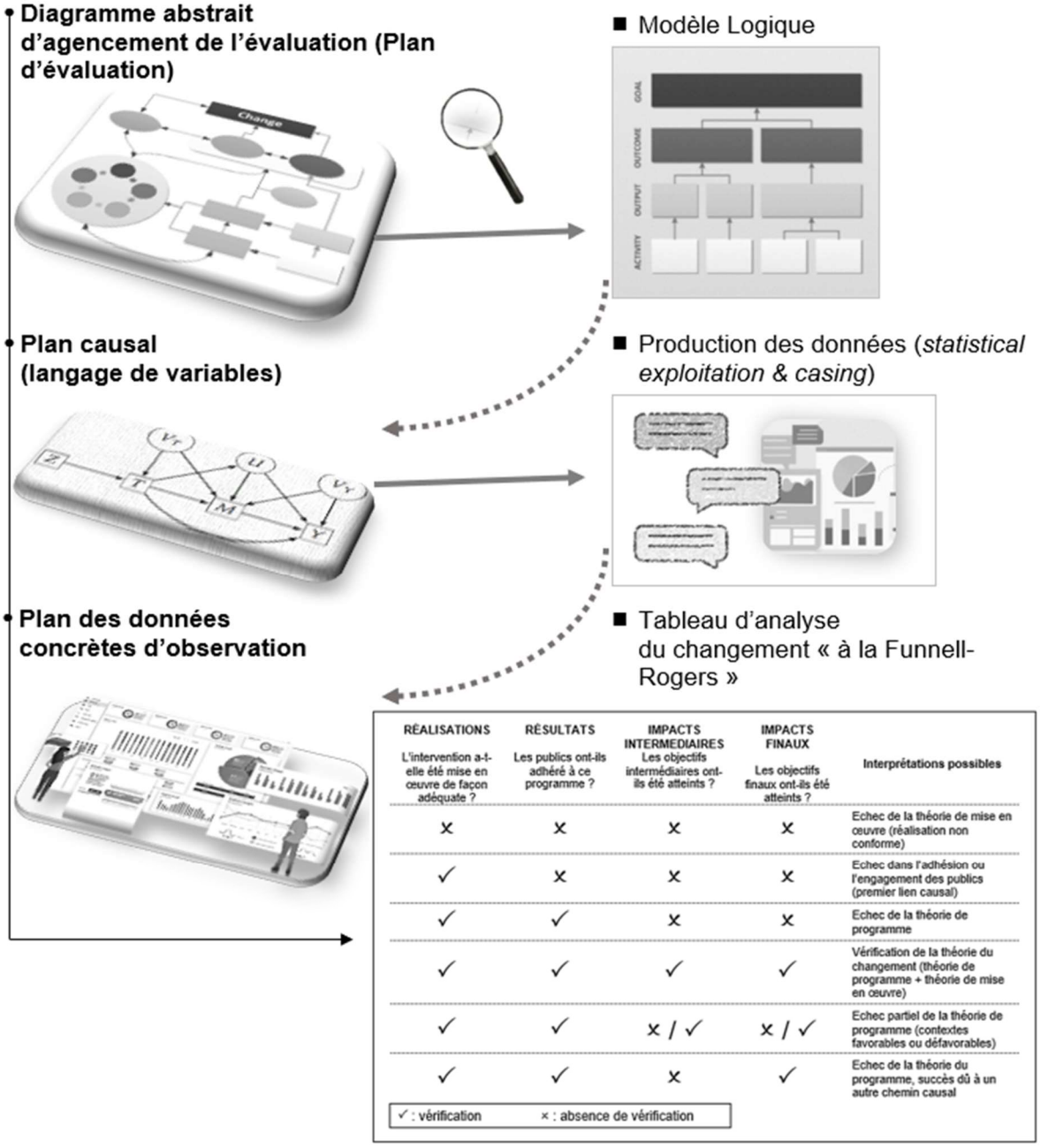
i. Le *plan logique* sur lequel est représentée la stratégie d'intervention et la théorie du processus de mise en œuvre afférente, donne à connaître les hypothèses du programme. Ce plan – qui comprend l'établissement du DLI – est parfois qualifié de « matrice d'évaluation » (Marceau, 2022) ce qui permet notamment de pointer sa portée exécutive ou programmatique concernant le processus d'évaluation. Les relations des parties dans cette diagrammatique présentent évidemment une analogie avec celles des parties de l'objet soumis à évaluation. Le diagramme des « relations » permet de représenter le processus par lequel une intervention est susceptible de produire un effet, mais aussi d'y « superposer » le cas échéant les contraintes et les conditions qui concernent, cette fois, le processus d'évaluation lui-même : en quoi, par exemple, les parties prenantes à l'intervention ou les bénéficiaires peuvent-ils et/ou doivent-ils, selon les cas, être impliqués dans ce processus d'évaluation, *a fortiori* dans un cadre d'expérimentation ? Comme dans l'illustration ci-dessous, le contenu de ce plan peut être décomposé et détaillé par « briques » s'agissant notamment des activités déployées par hypothèse pour la mise en œuvre du programme, des produits, des résultats, etc. Toute la difficulté est de symboliser proprement l'articulation entre processus d'intervention et processus d'évaluation, *a fortiori* lorsque ces processus sont mis au régime de l'expérimentation et nécessitent, à ce titre, l'enrôlement de tout un ensemble d'acteurs et d'actants.

ii. Le *plan d'analyse causale* conditionne le traitement des données factuelles dans le cas des démarches en observation et des analyses de contribution, la solution d'un contrefactuel dans le cas des analyses d'attribution. Le passage au plan de l'analyse causale nécessite que les choses puissent être exprimées en langage de variables.

iii. Le *plan d'effectuation* concerne l'établissement de la base informationnelle et la production des données, laquelle est elle-même un processus orienté par la théorie de l'intervention au sens où le modèle logique sélectionne *a priori* les données réputées pertinentes qu'il convient d'obtenir à travers les enquêtes.

[45] Cela ne doit toutefois pas donner lieu à confusion. Il n'est alors évidemment pas question d'estimer un paramètre de population, la collection de cas n'étant pas constituée dans un but de représentativité statistique mais d'inférence causale.

Figure 1 – Schéma d'ordonnancement d'une évaluation



RÉALISATIONS L'intervention a-t-elle été mise en œuvre de façon adéquate ?	RÉSULTATS Les publics ont-ils adhéré à ce programme ?	IMPACTS INTERMÉDIAIRES Les objectifs intermédiaires ont-ils été atteints ?	IMPACTS FINAUX Les objectifs finaux ont-ils été atteints ?	Interprétations possibles
×	×	×	×	Echec de la théorie de mise en œuvre (réalisation non conforme)
✓	×	×	×	Echec dans l'adhésion ou l'engagement des publics (premier lien causal)
✓	✓	×	×	Echec de la théorie de programme
✓	✓	✓	✓	Vérification de la théorie du changement (théorie de programme + théorie de mise en œuvre)
✓	✓	× / ✓	× / ✓	Echec partiel de la théorie de programme (contextes favorables ou défavorables)
✓	✓	×	✓	Echec de la théorie du programme, succès dû à un autre chemin causal

✓ : vérification × : absence de vérification

Source : Duclos (2024).

IV.4. UNE AUTRE CONCEPTION DU PASSAGE À L'ÉCHELLE

Les démonstrations d'impact mobilisant des approches qualitatives sont, par nature, « *moins concernées par la question de savoir "dans quelle mesure" une intervention a eu un impact sur un résultat souhaité* » que par la question de savoir « *"dans quelles conditions" [et à quelles conditions] une intervention a fonctionné dans le monde réel* » (Raimondo, 2023, p. 198). Une fois éclairées, ces conditions sont autant de « *points de passage obligé* » pour la production de l'effet attendu, et donc pour l'action. Si elles peuvent être posées dans les contours de l'intervention elle-même, et s'attribuer en propre à la « solution expérimentale », la reconstitution du chemin causal dans les approches qualitatives du type *process tracing* ne s'effectue jamais *ceteris paribus*. Les conditions de félicité d'une intervention peuvent

également, et à l'inverse, s'attribuer aux contextes changeants dans lesquels une action est amenée à se déployer.

La sensibilité des résultats d'une intervention au contexte invite alors à documenter les effets propres de ce « contexte », entre guillemets. Car si les « *caractéristiques d'un contexte sont pensées comme indépendantes des conduites que l'on y réfère* » (Zask, 2008, p. 314), sans pouvoir être affectées par ces conduites en retour, l'analyse reste, pour penser les perspectives de « généralisation », prisonnière de la solution de continuité – c'est-à-dire du *hiatus* – introduits par le jeu des critères de validité « interne » *versus* « externe » typique des analyses quantitatives contrefactuelles.

Le passage des analyses quantitatives aux analyses qualitatives nécessite de s'appuyer sur une autre compréhension de ce qu'est un « contexte » (Grimault, dans ce numéro). Alors que dans le premier cas, la solidité d'un lien de causalité formel doit manifester la puissance à s'affranchir des contextes, dans le *process tracing* la conception qu'on peut avoir d'un « contexte » doit procurer le moyen d'intégrer ce dernier pour la reconstitution du chemin causal, un point que la littérature ne parvient pas toujours à éclairer.

Pour mieux comprendre ce qui se joue entre une intervention et son « contexte », et peut s'effectuer sous la forme d'une action réciproque « dans le monde réel », Solveig Grimault invite à resituer l'évaluation par étude de cas dans une perspective pragmatiste (celle de John Dewey) établissant une distinction entre « contexte donné », ou antécédent, et *environnement*. À la différence d'un contexte, un environnement serait « *constitué par l'ensemble des conditions qui interviennent dans le développement [de quelque chose] au titre de moyen ou de ressources* » (Zask, 2008, p 314) : l'action ne se déploie pas « dans » un environnement dont elle serait séparée par une ligne de démarcation nette, et dont elle ne pourrait que subir l'effet – sous un régime d'externalités –, mais par le moyen des ressources que l'environnement est susceptible *ou non* de lui procurer en marchant.

Les conditions de félicité d'une intervention seraient alors réunies, non pas tant parce que le processus qu'elle instaure serait « résistant » à l'influence de tous les contextes, mais parce que ce processus aurait, à l'inverse, la capacité de puiser une ressource dans les environnements avec lesquels il saurait entrer en transaction, jusqu'au point où l'environnement pourrait « *se confond(re) avec les conséquences* » des activités ainsi déployées (*ibid*., p. 315). Une intervention, de ce point de vue, ne conserverait sa qualité constitutive de processus « orienté », à travers sa mise en œuvre, que parce qu'elle serait ouverte par principe *« à ce qui ne concerne pas déjà les conditions qui constituent son état initial »* (Zusman, 2021, p. 56) : *« constamment nourri par ses désorientations locales constantes »,* elle serait alors définie et ne pourrait persévérer dans son être que par la capacité qu'elle acquiert de « *restructuration continue (…) de la division entre intériorité et extériorité »* (*ibid*., p.57).

Le concept d'environnement paraît particulièrement congru à la prétention élevée par le *process tracing* (Mahoney, 2012 ; Palier *et al.*, 2018 ; Raimondo, 2023) ou l'analyse par cas à introduire une certaine parité dans la considération des concours permettant d'expliquer les résultats observés « par-dessus » les frontières de l'interne et de l'externe. Ce faisant, l'évaluation des conditions dans lesquelles des interventions complexes – amenées à négocier continûment leur existence avec des forces extérieures – peuvent produire leurs effets, fait émerger une autre conception du passage à l'échelle et des modalités qui s'y attachent.

Alors que le schéma probabiliste emporte avec lui un *a priori* de « scalabilité » (Tsing, 2012) constituant précisément ce que recherchent les porteurs d'interventions expérimentales (d'abord développées à petite échelle pour être « généralisées ») tout se passe comme si les approches processuelles préféraient prudemment se retrancher derrière une hypothèse *a priori* de « non scalabilité ». Il s'agirait moins de chercher à généraliser un « traitement » à base

d'activités largement préfigurées (*prefigured activities*) que d'identifier, par l'exploration des différentes contributions à l'effet recherché, une capacité de fonctionnement d'un dispositif avec ses environnements dans une perspective interactive de « validité écologique ».

Pour revenir à l'exemple de l'accompagnement, des auteurs critiques de l'application du principe d'*evidence-based practice* au travail social, remarquent que c'est le plus souvent « *la situation qui est probante, les données et les résultats pouvant s'y dissoudre ou y être investis selon les circonstances* » (Couturier et al., 2003, p. 75). L'enjeu de l'évaluation ne serait plus tant de mettre à l'épreuve des traitements génériques que des protocoles ouverts à un dialogue avec leur environnement (Guérin, 2021). Ces protocoles d'intervention auraient précisément pour cible des *situations* (par exemple, des « situations d'accompagnement ») et l'évaluation porterait sur la capacité de tels protocoles à appréhender dans leur généricité toute une gamme de « situations probantes » et de les structurer. Les situations « configurées »[46] par ces protocoles doivent permettre, du fait précisément de leur configuration, de satisfaire aux conditions de « re-production » d'un effet. L'expérimentation sociale, dans cette approche, répond davantage au modèle expérientiel (Baguelin et al., 2023) qu'aux présupposés diffusionnistes qui peuvent s'attacher à de réputées « bonnes pratiques » figées (*prefigured activities*). Dans cette perspective, l'évaluation aura pour charge de spécifier *mutatis mutandis*, les briques du protocole qui soutiennent le mieux – à travers leur appropriation par les acteurs – la perspective d'un essaimage soit, dans la métaphore apicole, d'un passage à l'échelle.

V. L'ÉVALUATION EN MÉTHODE MIXTE

Si l'évaluation apparait propice au rapprochement des approches quanti- et qualitatives, notamment dans une perspective d'inférence causale, il reste à en discuter les modalités. En évaluation, la combinaison des approches prend classiquement trois formes (White, 2008) : l'intégration des méthodes ; la comparaison des résultats ; leur synthèse dans une présentation unifiée. Quoi que déjà féconde (Gautié, 2023), cette dernière possibilité constitue une formule minimaliste. La deuxième, un peu plus ambitieuse, peut intervenir dans un but de confirmation/réfutation, d'enrichissement, d'explication[47] (Baïz, dans ce numéro). On propose ici de se concentrer sur la plus exigeante des trois formes : l'intégration des méthodes. Des exemples existent en évaluation de politiques de l'emploi (Osiander, 2021). Dans une optique séquentielle (Bamberger, 2012), il peut s'agir de : partir de l'enquête statistique pour cibler l'enquête qualitative sur des publics d'intérêt ou concevoir un guide d'entretien adéquate ; partir de l'enquête qualitative pour procéder à un échantillonnage stratifié ou concevoir un questionnaire adéquat ; passer par une étape qualitative pour tester un questionnaire. L'EPP (et l'exigence d'inférence causale qu'elle comporte) réclame d'aller plus loin.

V.1. L'ÉCONOMÉTRIE PARTICIPATIVE

Quoi qu'elle donne peut-être au quantitatif un rôle prépondérant, la formule de « l'économétrie participative » constitue un exemple d'intégration forte. Sans se référer à la TBE, elle repose sur des principes étonnamment cohérents (Rao et Woolcock, 2003, p. 173) : partir de questions et d'hypothèses définies mais révisables ; ne pas séparer les tâches de

[46] Configured situations vs prefigured activities.

[47] On pourra par exemple vérifier au moyen d'une enquête qualitative une association statistique, enrichir l'analyse en identifiant des problèmes ou en obtenant de l'information sur des facteurs non couverts par l'enquête statistique. L'enrichissement pourra aussi être théorique lorsque l'enquête qualitative permet de formuler des hypothèses testables statistiquement. Enfin, l'enquête qualitative peut permettre de comprendre des résultats statistiques inattendus.

collecte et d'analyse des données ; ne pas séparer la mesure d'impact de l'analyse des processus ; impliquer les enquêtés dans l'analyse et l'interprétation des résultats ; centrer l'analyse sur les dépendances statistiques et mécanismes généralisables. Dans ce cadre, l'enquête qualitative génère des hypothèses ancrées dans un monde vécu, échappant aux préjugés de l'analyste. À l'étape du raisonnement économétrique, elle aide à intuiter le sens des causalités, à trouver des variables instrumentales et à exploiter des configurations quasi-expérimentales. Elle permet d'améliorer la mesure des résultats (choix d'indicateurs), de repérer le rôle de confondants potentiels, de trouver les moyens de mesurer d'éventuelles confondants « inobservables ». À l'étape de l'estimation, l'information tirée de l'enquête qualitative aide à réduire les erreurs de mesure, à diagnostiquer les biais d'estimation et leur direction. À l'étape de l'interprétation, elle apporte des exemples de causation suggestifs, qui guident l'élaboration de narratifs génériques.

V.2. L'ENQUÊTE QUALITATIVE EN RCT

L'enquête qualitative est a fortiori pertinente dans le cadre d'une expérimentation avec assignation aléatoire (Prowse et Camfield, 2013). Au regard des limites évoquée supra, elle permet de répondre à plusieurs difficultés : la compréhension du traitement, l'analyse des effets de sélection liés à l'assignation au traitement, l'analyse de l'hétérogénéité de l'impact, l'interprétation des variables et l'inférence de causation. Comprendre comment une intervention s'est inscrite dans un « contexte », ce qu'elle a effectivement changé dans la vie quotidienne des sujets et comment ceux-ci l'ont vécu ne va pas de soi. De même, la question de fidélité d'une intervention mise en œuvre à sa définition nominale (*program fidelity*) est toujours posée. L'enquête qualitative permet de répondre à cette question faussement triviale (« en quoi a effectivement consisté le traitement ? ») et d'éviter les erreurs d'imputation. Le deuxième aspect concerne les risques d'assignation « défiée » (*defiers*) mais aussi les configurations quasi-expérimentales (Harding et al., 2013) où comprendre le processus d'assignation des sujets au traitement est essentiel[48]. Documenter les processus d'assignation par une enquête qualitative ad hoc évite de s'en remettre à des hypothèses incertaines. Le troisième aspect (effets hétérogènes) répond à la limitation que constitue le fait de devoir s'en remettre à un effet causal moyen. L'enquête qualitative est particulièrement adéquate pour dévoiler les dimensions et sources d'effet hétérogène, particulièrement quand on n'en a pas de préconception. Cela tient au caractère ouvert et non-structuré de la collecte de données qualitative et à la posture compréhensive que l'enquête qualitative rend possible. Enfin, le quatrième aspect (l'interprétation des variables) concerne la fiabilité de l'information contenue dans une variable statistique et/ou la validité du sens qu'on lui prête[49]. L'enquête qualitative peut opportunément porter sur le système d'information produisant la donnée et sur les agents contribuant à la fabrication de la « donnée » (Thévenot, 1983).

[48] L'enjeu est alors de motiver une comparaison identificatrice (*i.e.* pouvant justifier une interprétation causale) en distinguant des facteurs actifs dans l'assignation et corrélés aux (variables de) résultats. Deux possibilités. Si la corrélation prévaut indépendamment du traitement, ces facteurs sont des confondants : la comparaison doit être ajustée selon leurs valeurs. Si la corrélation tient *exclusivement* à l'action du traitement, ces facteurs sont des instruments : la comparaison *ne doit pas* être ajustée selon leurs valeurs (Cinelli et al., 2022).

[49] Dans une enquête par questionnaire, il existe toujours le risque qu'une même question soit l'objet d'interprétations différentes de la part de l'enquêté et du statisticien. L'écart d'interprétation peut aussi survenir d'un enquêté à l'autre, notamment lorsqu'une formulation comporte une part d'ambiguïté. L'enquête qualitative permet d'apprécier ce risque voire de le prévenir dans le cadre d'un test de questionnaire. Un même enjeu peut concerner des données administratives. La signification d'un enregistrement lié à un acte administratif est parfois plus subtile que ce qu'en dit la documentation adjointe : assister concrètement à l'opération administrative peut s'avérer précieux pour saisir l'information contenue dans une variable. Cela peut aussi concerner une étape de codage d'un enregistrement ou de regroupement de valeurs.

V.3. LE RÉSEAU BAYÉSIEN COMME INTÉGRATEUR QUALI-QUANTI

En tant que procédure consistant à initier l'enquête évaluative à partir d'aprioris probabilisés sur la structure causale du monde empirique et à réviser ces croyances au gré de la collecte d'information, l'approche bayésienne est invoquée depuis longtemps par les praticiens des méthodes mixtes (Rao, 1998). Mais la plupart de ces invocations sont restées virtuelles. Des propositions se font jour depuis une dizaine d'années pour associer à l'estimation d'effets causals moyens un degré de confiance dérivé non d'hypothèses statistiques mais d'enquêtes qualitatives (Humphreys et al., 2015). On trouve des exercices analogues en médecine autour des moyens de personnaliser les thérapies en croisant enseignements de RCT et observation clinique (Donatini, 2017). Pour autant que l'on accepte l'assimilation de l'enquête monographique à un examen clinique, la démarche paraît particulièrement opportune face aux cas d'effets individuels hétérogènes : le qualitatif et l'accès direct à la causation y retrouve une portée exclusive. Toujours dans l'optique d'une application formelle de la logique bayésienne, des méthodes basées sur la simulation de probabilités sont développées dans un cadre de TBE (Befani et al., 2021)[50].

Les avancées de l'inférence causale à base d'apprentissage statistique évoqués précédemment suggèrent une voie peut-être plus naturelle et polyvalente. Le recouvrement diagrammatique entre DAG et DLI introduit supra met en évidence l'utilité d'expliciter des aprioris qualitatifs sur la structure causale des situations sociales auxquelles s'applique l'intervention à évaluer. La greffe sur cette structure d'un réseau bayésien (carte de probabilités conditionnelles) quantifiant des associations probabilistes, établit formellement le cadre d'une analyse en méthodes mixtes. Il s'agirait d'un approfondissement de la TBE comme dispositif d'intégration d'approches processuelle et probabiliste de la démonstration causale. Un espace de collaboration interdisciplinaire où évaluer des interventions complexes, répondre aux exigences d'une EPP méthodique sans être technocratique.

VI. LES CONTRIBUTIONS DU NUMÉRO

Les contributions de ce second opus consacré à l'évaluation se répartissent entre discussions méthodologiques générales, réflexions illustrées par l'évaluation d'expérimentations (inscrites dans le PIC ou la « Stratégie pauvreté ») et synthèse évaluative (à propos des dispositifs d'incitation à négocier).

VI.1. TROIS DISCUSSIONS SUR LES MÉTHODES D'ÉVALUATION DE POLITIQUES PUBLIQUES

La première contribution du numéro prend la forme d'un entretien avec Adam Baïz, actuellement coordinateur de l'EPP à la Cour des comptes, autour de la doctrine élaborée à France Stratégie concernant l'application des méthodes mixtes à l'EPP. Cette doctrine part de l'appréciation la plus courante des différences entre « cultures » (Goertz et al., 2006) qualitative et quantitative : type de matériau empirique (cas versus population), rapport à la théorie (inductif versus hypothético-déductif) et modalité de l'inférence causale (processuelle versus probabiliste-contrefactuelle). Huit perspectives d'articulation en sont tirées : perspective de confirmation, d'enrichissement, de complexification, de généralisation, de triangulation, de complémentarité, de développement. Il s'avère que chacune de ces perspectives trouve au

[50] La simulation de probabilités à partir d'un *agent-based model* déroge pourtant au concept de probabilité subjective censée formaliser le rôle que prête la TBE aux expertises académiques ou profanes préexistantes.

moins une illustration dans les exercices d'évaluation mixtes proposés dans le numéro, que ces exercices se réclament de l'évaluation contrefactuelle, de l'analyse quali-quantitative comparée (Ragin, 1987) ou de l'analyse de contribution (Mayne, 2012). Adam Baïz n'élude pourtant pas les difficultés pratiques de l'évaluation mixte. Au regard de la technicité des méthodes propres à chacune des cultures, l'idée n'est pas de renoncer à la spécialisation mais que les spécialistes de l'une aient une compréhension suffisante des principes de l'autre pour savoir où se jouent les complémentarités. Que l'enquête soit qualitative ou quantitative, il s'agit de l'orienter de sorte à produire une connaissance à la fois exclusive (inaccessible à l'autre approche) et supplétive (fertile pour l'autre approche) ce qui suppose une bonne « culture générale » de l'évaluation. Le commentaire de Solveig Grimault porte sur le guide France Stratégie (Baïz et Revillard, 2022) mais aussi sur plusieurs publications institutionnelles récentes fixant des attentes méthodologiques en matière d'EPP. L'autrice note une propension dans la commande publique d'EPP à reléguer l'enquête qualitative, malgré son potentiel autonome, à un rôle subalterne. En documentant des mécanismes et des influences contextuelles, le concept de causalité processuelle (causation) permet de mieux comprendre les résultats des politiques publiques. Plutôt qu'à hiérarchiser les formes d'enquête évaluative, les préconisations institutionnelles gagneraient à définir une complémentarité symétrique qui intègre plutôt qu'elle n'oppose les éclairages quali- et quantitatifs.

L'article d'Yves de Curraize et Francesco Sergi explore l'évolution des préconisations en matière d'application des méthodes mixtes à l'évaluation des politiques de l'emploi menées en France. L'étude mobilise pour cela un corpus de littérature grise (guides méthodologiques, études et rapports d'évaluation) produite entre 2008 et 2023 ; les auteurs examinent la manière dont méthodes qualitatives et quantitatives sont définies, articulées et mises en pratique. Ils constatent que ces méthodes sont mises en œuvre par des acteurs différents, cabinets de conseil pour le qualitatif, services ministériels ou universitaires pour le quantitatif, et qu'elles sont plus souvent juxtaposées qu'articulées. Pour autant, leur couplage est recommandé sur l'ensemble de la période couverte. Il arrive pourtant que des variables utilisées pour l'enquête quantitative soient le résultat d'enquête qualitatives préalables. Sur la période récente, l'étude relève une pratique de la synthèse des résultats qualitatifs et quantitatifs essentiellement dans le cadre de rapports d'évaluation. Il apparaît qu'une réelle application des méthodes mixtes à l'évaluation réclame une synchronisation des étapes propres à chaque méthode qui est rarement obtenue.

VI.2. QUATRE RETOURS D'EXPÉRIENCE ALIMENTANT UNE RÉFLEXION MÉTHODOLOGIQUE

Deux articles relatent des évaluations menées dans le cadre du Plan d'investissement dans les compétences (PIC).

L'article d'Agathe Devaux-Spatarkis, Pauline Joly et Thomas Bouget illustre les enjeux d'une évaluation pragmatique. L'étude croise TBE et réflexivité des parties prenantes dans le cadre de l'expérimentation de dispositifs d'accès à la formation (« Illettrisme et Illectronisme » et « Mobilisation vers la formation ») ne permettant pas une évaluation d'impact contrefactuel. L'évaluation doit alors privilégier un objectif d'utilité. L'expérimentation porte notamment sur de nouvelles méthodes de repérage, de mobilisation et d'accompagnement des publics éloignés de l'emploi ou de la formation. L'évaluation part de la reconstruction d'une théorie synthétique représentant la logique d'intervention des dispositifs, puis explore les mécanismes à l'œuvre par enquête qualitative selon une perspective dite réaliste. Loin de schémas coût-bénéfice, les dispositifs considérés engagent des mécanismes psychosociaux de mobilisation (par sollicitation, orientation de proximité, par les pairs…) et de pédagogie (« du détour », de projet, des petits pas, du pied à l'étrier...). L'évaluation s'efforce de générer des connaissances utiles pour les parties prenantes, en identifiant des pratiques innovantes et en proposant des

recommandations pour améliorer les dispositifs existants ; elle constitue un outil de légitimation et de capitalisation d'expériences.

Toujours à propos de dispositifs du PIC, Carole Beaugendre, Elise Crovella, Jeoffrey Magnier et Isabelle Recotillet proposent une évaluation en méthode mixte dont ils tirent une réflexion sur l'articulation des cultures quali- et quantitative. Il s'agit de s'interroger sur les modalités pratiques d'une hybridation des approches dans le cas d'interventions complexes. La réflexion consiste en l'analyse des sources de divergences entre les conclusions de deux exercices parallèles : une mesure d'impact contrefactuel et une évaluation empruntant à l'enquête qualitative. L'analyse quantitative conclut à un impact positif du dispositif sur l'accès à la formation mais à un impact négatif sur l'accès à l'emploi, notamment pour les jeunes. L'analyse qualitative documente un impact positif en termes d'accès à la formation et à l'emploi ; elle met en lumière des mécanismes psychosociaux (gains en autonomie et en motivation) et des facteurs contextuels. Trois aspects concourent aux divergences constatées : les causations caractérisées par l'enquête qualitative mobilisent des influences contextuelles que l'analyse quantitative neutralise ; l'information quantitative, plus pauvre, ne permet pas de capter certains effets positifs ; les calendriers d'évaluation n'ayant pas été synchronisés, informations qualitatives et quantitatives n'ont pu être suffisamment intégrées. L'article conclut à l'importance d'une intégration pensée ex ante des enquêtes évaluatives quali- et quantitatives.

À nouveau dans une perspective d'évaluation en méthode mixte, l'article de Nicolas Farvaque, Renaud Garrigues, Elise Picon et Carole Beaugendre porte sur SEVE Emploi, un programme visant à améliorer les pratiques d'accompagnement des structures de l'insertion par l'activité économique (IAE). Il s'agit, dans une logique de médiation active, d'inciter les structures IAE à multiplier les mises en situations professionnelles au sein d'entreprises classiques, selon une logique de « sas » vers l'emploi durable. Le dispositif passe par un changement des pratiques des acteurs de l'insertion ; il prend la forme d'une formation-action comprenant neuf journées réparties sur un an, avec des actions concrètes à réaliser entre chaque séance. Le protocole d'évaluation combine, dans un cadre TBE unificateur, une approche réaliste et une méthode mixte : enquêtes monographiques, enquêtes par questionnaire (en ligne et par téléphone), analyse statistique comparative. L'étude documente les contributions respectives des différentes composantes de l'intervention aux résultats obtenus ; l'enjeu est notamment de caractériser des effets intermédiaires et finaux ainsi que leurs relations. Pour les premiers, les structures formées à SEVE montrent des traces de changement (refonte des fiches de poste, nouvelles postures professionnelles, collaborations nouvelles centrées sur la relation-entreprises) ; elles utilisent plus fréquemment les périodes d'immersion. Concernant les effets finaux, les salariés en IAE des structures formées à SEVE se déclarent mieux préparées à la recherche d'emploi et plus confiants ; les taux de sortie en emploi durable sont meilleurs. Les auteurs documentent aussi des spécificités contextuelles pouvant avoir contribué aux résultats.

L'article d'Annie Jolivet explore les effets et les limites d'un dispositif d'incitation à négocier en faveur de l'emploi des salariés âgés en France, en se basant sur des enquêtes de terrain réalisées entre 2010 et 2018. Créé en 2008 et appliqué jusqu'en 2017, le dispositif devait augmenter le taux d'emploi des 55-64 ans et diminuer la proportion de « seniors » ni en emploi ni à la retraite (dans un contexte de relèvement des âges d'ouverture des droits à la retraite à taux plein). L'incitation combinait loi et négociation au niveau des entreprises et des branches, prescrivant la conclusion d'un accord ou d'un plan d'action sous peine de sanction financière. L'article propose une réflexion critique sur la construction de pratiques d'entreprises impulsées par cette incitation et évalue sa portée et ses limites. L'évaluation montre un fort taux de conclusion d'accords ou de plans d'action, mais une faible pertinence des indicateurs et objectifs chiffrés, souvent impensés et peu discutés. Les entreprises ont adopté une approche minimale

et conformiste et les actions négociées ont été peu mises en œuvre. L'enquête permet cependant d'identifier quelques « bonnes pratiques »[51].

Pour conclure le numéro, les comptes rendus de lecture de Laure Bazzoli, Madlyne Samak, Sophie Dessein et Fabien Brugière portent (respectivement) sur quatre ouvrages décrivant l'état du travail et de l'emploi en France après des décennies de réforme néolibérale : Un compromis salarial en crise (Signoretto et Giraud, 2023), Essentiel.les et invisibles ? Classes populaires au travail en temps de pandémie (Gardes, 2022), L'hécatombe invisible (Lépine, 2023) et L'addiction au travail (Loriol, 2023). En discutant l'ouvrage dirigé par Yvon Karel, Le syndicalisme est politique (Yon, 2023), Riyad Manseri examine une voie (étroite) de re-politisation de la vie sociale par le travail. Au total, l'ensemble des ouvrages discutés disent l'urgence d'une intervention publique moins idéologique, moins indifférente à ses effets concrets et surtout… plus démocratique.

RÉFÉRENCES BIBLIOGRAPHIQUES

Aeberhardt R., Chiodi V., Crépon B., Gaini M., Vicard A., 2014, Revenu contractualisé d'autonomie – rapport d'évaluation. Fonds d'Expérimentation de la Jeunesse.

Aventur F., Galliot Y., Glover D., Rabner M.-J., 2016, L'impact de la démarche de prospection auprès des entreprises – une évaluation randomisée, Pôle emploi, Etudes et Recherches n°6.

Angrist J.D., Pischke J.S., 2009, Mostly harmless econometrics: An empiricist's companion, Princeton University Press.

Baguelin O., Duclos L., 2023, « Les politiques de l'emploi au régime de l'expérimentation : la preuve de concept », Socio-économie du travail, n°14, p. 17-54.

Baïz A., Nakhla M., 2018, « Pour une approche algorithmique de la nature protéiforme et fractale des instruments de l'action collective », Politiques et management public, 35(3-4), p. 153-172.

Baïz A., Revillard A., 2022, Comment articuler les méthodes qualitatives et quantitatives pour évaluer l'impact des politiques publiques, Un guide à l'usage des décideurs et des praticiens, France Stratégie.

Beach D., 2020, Multi-method research in the social sciences: A review of recent frameworks, a way forward, Government & Opposition, 55(1), p. 163-182.

Befani B., Elsenbroich C., Badham J., 2021, Diagnostic evaluation with simulated probabilities, Evaluation, 27(1), p. 102-115.

Behaghel L., Crépon B., Gurgand M., 2009, Évaluation d'impact de l'accompagnement des demandeurs d'emploi par les opérateurs privés de placement et le programme Cap vers l'entreprise, Rapport final.

Behaghel L., Crépon B., Gurgand M., Kamionka T., Lequien L., Rathelot R., Zamora P., 2013, L'accompagnement personnalisé des demandeurs d'emploi, Revue française d'économie, 131(1), p. 123-158.

Bezes P., Palier B., Surel Y., 2018, Le process tracing: du discours de la méthode aux usages pratiques, Revue française de science politique, 68(6), p. 961-965.

[51] Dont la notion est cependant discutée par l'autrice.

Blanchard T., 2018, Causalité (A), In: M. Kristanek (dir.), L'encyclopédie philosophique [en ligne] disponible <https://encyclo-philo.fr/causalite-a> [Consultation juin 2024].

Bamberger M., 2012, Introduction to mixed methods in impact evaluation, Impact evaluation notes, 3(3), p. 1-38.

Callon M., Latour B., 1997, « 'Tu ne calculeras pas !' ou comment symétriser le don et le capital » in Caillé A., Le capitalisme aujourd'hui, Mauss n°9, La découverte.

Chen H.T., Rossi P.H., 1987, The theory-driven approach to validity, Evaluation, program planning, 10(1), p. 95-103.

Cibois V., 2023, « Les expérimentations en droit de la formation professionnelle au bénéfice d'un changement de culture normative », Socio-économie du travail, n°14, p. 169-200.

Cinelli C., Forney A., Pearl J., 2024, A crash course in good, bad controls, Sociological Methods & Research, 53(3), p. 1071-1104.

Collier D., 2011, Understanding process tracing, PS: political science & politics, 44(4), p. 823-830.

Concato J., Shah N., Horowitz R.I., 2000, "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs", The New England Journal of Medecine, vol. 342, n°25, p. 1887-1892.

Couturier Y., Carrier S., 2003, « Pratiques fondées sur les données probantes en travail social : un débat émergent », *Nouvelles pratiques sociales*, vol. 16, n°2, p. 68-79.

Dawid A.P., 2000, Causal inference without counterfactuals, Journal of the American statistical Association, 95(450), p. 407-424.

Dahan-Gaida L., 2023, L'art du diagramme, Vincennes, PUV.

Deaton A., Cartwright N., 2018, Understanding, misunderstanding randomized controlled trials, Social science & medicine, 210, p. 2-21.

Delahais T., Devaux-Spatarakis A., Revillard A., Ridde V., 2021, Évaluation. Fondements, controverses, perspectives.

Desrosières A., 1989, L'opposition entre deux formes d'enquête : monographie et statistique, Cahiers du Centre d'Études de l'Emploi, 33, p. 1-9.

Desquinabo N., 2021, L'évaluation dans les politiques complexes. Les cas de la lutte contre l'habitat indigne et du traitement des copropriétés en difficulté, Revue française d'administration publique, (1), p. 115-129.

Devaux-Spatarakis A., 2014, L'évaluation « basée sur la théorie », entre rigueur scientifique et contexte politique, Politiques et management public, 31(1), p. 51-68.

Donatini B., 2017, La méthode Bayésienne pour aider à évaluer l'efficacité des thérapies personnalisées, Hegel, 7(2), p. 113-129.

Duclos L., 2017, Les conditions de mobilisation de la formation et de l'expérience pour l'emploi, Education permanente, n°213, p. 121-132.

Duclos L., 2024, « L'évaluation par l'expérimentation », Document d'études DGEFP-IDHES, ministère du Travail, décembre.

Erhel C., 2020, Les politiques de l'emploi, Que sais-je.

Fagot-Largeault A., 1992, Quelques implications de la recherche étiologique, Sciences sociales et santé, vol. 10, n°3, p. 33-45.

Gaini M., 2023, « Évaluer les politiques de l'emploi (et de santé) », Socio-économie du travail, n°14, p. 55-66.

Gagnon M., 1975, « Une analyse sémantique du concept de causalité est-elle possible ? », Philosophiques, 2(2), p. 187–205.

Gardes C., 2022, Essentiel.les et invisibles ? Classes populaires au travail en temps de pandémie, Paris, Éditions du Croquant.

Gautié J., 2023, « L'évaluation des politiques de l'emploi : l'économiste, le sociologue et l'expert », Socio-économie du travail, n°14, p. 67-78.

Geiger D., Verma T., Pearl J., 1990, Identifying independence in Bayesian networks, Networks, 20(5), p. 507-534.

Gertler P.J., Martinez S., Premand P., Rawlings L. B., Vermeersch C. M., 2011, L'évaluation d'impact en pratique, The World Bank.

Ginzburg C., 1980, Signes, traces, pistes : Racines d'un paradigme de l'indice, Le débat, (6), p. 3-44.

Glaser B., Strauss A., 1967, The Discovery of Grounded Theory, Aldine Publishing Company, Hawthorne, NY.

Glennan S., 2008, Mechanisms, In: The Routledge companion to philosophy of science, p. 404-412.

Goertz G., Mahoney J., 2006, A tale of two cultures: Contrasting quantitative, qualitative research, Political analysis, 14(3), p. 227-249.

Guérin M., 2021, *La troisième main : des techniques matérielles aux technologies intellectuelles*, Arles, Actes Sud.

Harding D.J., Seefeldt K.S., 2013, Mixed methods, causal analysis, In: Handbook of causal analysis for social research, p. 91-110.

Heckman J.J., 2010, Building bridges between structural, program evaluation approaches to evaluating policy, Journal of Economic literature, 48(2), p. 356-398.

Heckman J., Pinto R., 2024, Econometric causality: The central role of thought experiments, Journal of Econometrics, 105719.

Humphreys M., Jacobs A. M., 2015, Mixing methods: A Bayesian approach, American Political Science Review, 109(4), p. 653-673.

Huneman P., 2020, Pourquoi ? Une question pour découvrir le monde, Autrement.

Johnson R. B., Russo F., Schoonenboom J., 2019, Causation in mixed methods research: The meeting of philosophy, science, practice, Journal of Mixed Methods Research, 13(2), p. 143-162.

Jany-Catrice F., Fretel A., Gardin L., 2023, « De quoi l'inflation d'évaluations dans l'expérimentation 'Territoires zéro chômeur' est-elle le nom ? », Socio-économie du travail, n°14, p. 79-112.

Kistler M., 1998, Reducing causality to transmission, Erkenntnis, 48(1), p. 1-25.

Labrousse A., Zamora P., 2013, Expérimentations de terrain et politiques publiques du travail et de l'emploi, Apports récents et mises en perspective Introduction, Travail et emploi, (135), p. 5-13.

Lépine M., 2023, L'hécatombe invisible - Enquête sur les accidents du travail, Le Seuil, 2023.

Lewis D., 1973, Causation, The journal of philosophy, 70(17), p. 556-567.

Loriol M., 2023, L'addiction au travail. De la pathologie individuelle à la gestion collective de l'engagement, éditions Le Manuscrit, Paris.

Mejia J.M.L., Rey L., 2022, La modélisation des interventions en évaluation in L'évaluation en contexte de développements : enjeux, approches et pratiques, p. 195.

Mackie J.L., 1965, Causes, conditions, American philosophical quarterly, 2(4), p. 245-264.

Mahoney J., 2008, Toward a unified theory of causality, Comparative Political Studies, 41(4-5), p. 412-436.

Mahoney, J., 2012, « The Logic of Process Tracing Tests in the Social Sciences », *Sociological Methods & Research*, 41(4), p. 570-597.

Marceau R., 2022, Le processus d'évaluation, in Rey et al. (dir.), L'évaluation en contexte de développement, p. 215-230.

Marceau R., Sylvain F., 2022, La terminologie de l'évaluation, in Rey et al. (dir.), L'évaluation en contexte de développement, p. 37-56.

Mayne J., 2012, Contribution analysis: Coming of age?, Evaluation, 18(3), p. 270-280.

Mayne J., 2017, Theory of change analysis: Building robust theories of change, Canadian Journal of Program Evaluation, 32(2), p. 155-173.

Mejia R., Rey L., 2022, La modélisation des interventions en évaluation, in Rey et al. (dir.), L'évaluation en contexte de développement, p. 195-214.

Osiander C., 2021, Lessons from mixed-method evaluations—An example from labor market research, Research Evaluation, 30(1), p. 90-101.

Palier B., Trampusch, C., 2018, « Comment retracer les mécanismes causaux ? Les différents usages du process tracing », *Revue française de science politique*, vol. 68(6), p. 967-990.

Passeron J.-C., Revel J. (dir.), 2020, Penser par cas, Éditions de l'École des hautes études en sciences sociales.

Pearl J., 1985, Bayesian networks: A model of self-activated memory for evidential reasoning, in Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA, p. 15-17.

Pearl J., 2000, Causality: models, reasoning, inference, Cambridge, UK: Cambridge University Press, 19(2), p. 3.

Pearl J., Mackenzie D., The Book of Why, Allen Lane.

Powell S., 2018, The Book of Why: The New Science of Cause and Effect by J. Pearl and D. Mackenzie, 2018, book review, Journal of MultiDisciplinary Evaluation, 14(31), p. 47-54.

Prairat E., 2014, Valuation et évaluation dans la pensée de Dewey, Le Télémaque, 46(2), p. 175 (167-176).

Prowse M., Camfield L., 2013, Improving the quality of development assistance: What role for qualitative methods in randomized experiments?, Progress in Development Studies, 13(1), p. 51-61.

Ragin C.C., Becker H.S., 1992, What is a case? Exploring the foundations of social inquiry, Cambridge University Press.

Ragin C.C., 2007, Comparative methods, The Sage handbook of social science methodology, p. 67-80.

Ragin C.C., 2018, Casing, in Routledge Handbook of Interdisciplinary Research Methods, p. 104-107.

Raimondo E., 2023, « Tracage de processus » in Revillard A., éd., *Méthodes et approches en évaluation des politiques publiques*, Québec, Éditions Science et bien commun, p.193-203 (citation : p. 198).

Rao V., 1998, "Wife-Abuse, Its Causes and Its Impact on Intra-Household Resource Allocation in Rural Karnataka: A 'Participatory' Econometric Analysis." In Maithreyi Krishnaraj, Ratna Sudarshan, and Abusaleh Shariff, eds., Gender, Population, and Development. Oxford, U.K.: Oxford University Press.

Rao V., Woolcock M., 2003, Integrating qualitative, quantitative approaches in program evaluation, in The impact of economic policies on poverty, income distribution: Evaluation techniques, tools, p. 165-190.

Ravallion M., 2020, Should the randomistas (continue to) rule?, National Bureau of Economic Research No. w27554.

Reichenbach H., 1956, The Direction of Time, Berkeley, Los Angeles, University of California Press.

Retsin C., 2023, « Évaluation de l'expérimentation TZCLD selon une analyse processuelle », Socio-économie du travail, n°14, p. 145-168.

Rey L., Quesnel J.S., Sauvain V. (dir), 2022, L'évaluation en contexte de développement – Enjeux, approches et pratiques, École nationale d'administration publique, JFD Éditions.

Rohlfing I., Zuber C.I., 2021, Check your truth conditions! Clarifying the relationship between theories of causation, social science methods for causal inference, Sociological Methods & Research, 50(4), p. 1623-1659.

Rorty R., 1993, Contingence, ironie et solidarité, Paris : A. Colin.

Salmon W.C., 1994, Causality without counterfactuals, Philosophy of Science, 61(2), p. 297-312.

Scriven M., 2008, A summative evaluation of RCT methodology: & an alternative approach to causal research, Journal of multidisciplinary evaluation, 5(9), p. 11-24.

Signoretto C., Giraud B. (dir.), 2023, Un compromis salarial en crise. Que reste-t-il à négocier dans les entreprises ? Baptiste Giraud, Camille Signoretto, Editions du Croquant, Collection Dynamiques socio-économiques.

Simonnet V., 2014, Évaluation des politiques actives du marché du travail. Introduction, Travail et emploi, 139, p. 5-14.

Smith G.C.S., Pell J.P., 2003, "Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials", British Medical Journal, 327, p. 1459-1461.

Smith H.L., 2013, La causalité en sociologie et démographie. Retour sur le principe de l'action humaine, Philadelphia, PA : Population Studies Center, University of Pennsylvania, PSC Working Paper Series, PSC 13-11.

Stern E., Stame N., Mayne J., Forss K., Davies R., Befani B., 2012, Broadening the range of designs, methods for impact evaluations.

Tantot A., 2023, « Territoire zéro chômeur de longue durée – l'évaluation conflictuelle d'une expérimentation singulière », Socio-économie du travail, n°14, p. 113-144.

Thévenot L., 1983, L'économie du codage social, Critiques de l'économie politique, 23-24, p. 188-222.

Todd P.E., Wolpin K.I., 2020, The best of both worlds: Combining RCTs with structural modelling, Journal of economic literature.

Tsing A.L., 2012, « On nonscalability: The Living World Is Not Amenable to Precision-Nested Scales », *Common Knowledge*, vol. 18, issue 3, p. 505-524.

Verma T., Pearl J., 1990, Causal networks: Semantics, expressiveness, in Machine intelligence, pattern recognition, Vol. 9, p. 69-76.

White H., 2008, Of probits, participation: The use of mixed methods in quantitative impact evaluation, IDS bulletin, 39(1), p. 98.

Weber M., 2019, Economy, society: A new translation, Harvard University Press.

Weiss C.H., 1997, Theory-based evaluation: past, present, future, New directions for evaluation, 76, p. 41-55.

Yon K., 2023, Le syndicalisme et politique. Questions stratégiques pour un renouveau syndical, La Dispute.

Zamora P., 2011, La méthode d'évaluation aléatoire : apports et limites, Tracés – Revue de Sciences humaines, #11, p. 175-186.

Zask J., 2008, « Situation ou contexte ? Une lecture de Dewey », Revue internationale de philosophie, n°245, p. 313-328.

Zusman Y., 2021, *L'espace aléatoire*, Paris, PUF.