SEPT PROBLÈMES MÉTHODOLOGIQUES (ET LEURS SOLUTIONS)

Introduction

Au travers de ce document, nous allons considérer 7 problèmes méthodologiques courants, qui ont tous la capacité de sérieusement compromettre la qualité d'une recherche. La plupart de ces problèmes sont communs à tous types de plan de recherche, mais nous nous focaliserons ici plus particulièrement sur les plans expérimentaux, car ce sont eux qui offrent le plus de possibilités de contrôle sur bon nombre de ces problèmes. Nous commencerons donc par quelques rappels généraux sur les types de plans de recherche, puis nous entrerons dans le vif du sujet et aborderons en détail les 7 problèmes, ainsi que certaines solutions permettant de les éviter. Enfin nous conclurons en résumant quelques points clés et en proposant quelques recommandations synthétiques.

Les types de plan de recherche

Commençons par passer en revue les « grandes familles » de types de plan de recherche (plans corrélationnels, plans expérimentaux ; approches transversales, approches longitudinales). Nous nous attarderons ensuite un peu plus en détail sur les plans expérimentaux, en particulier pour clarifier la différence entre *les plans à groupes indépendants* et les *plans à mesures répétées*, ainsi que pour expliciter leurs avantages et inconvénients respectifs.

Plans corrélationnels et plans expérimentaux

Une première distinction importante à comprendre est celle qui sépare plan de recherche corrélationnel et plan de recherche expérimental.

Les *plans corrélationnels* sont une grande famille de plans de recherche qui permettent d'étudier, par l'*observation*, toutes sortes de phénomènes naturels, notamment ceux qui ne peuvent pas (ou difficilement) être manipulés par l'expérimentateur. Les plans corrélationnels consistent typiquement à mesurer le degré de dépendance – ou de covariation ou de corrélation – entre deux ou plusieurs variables, qui peuvent être de toutes sortes (sexe, lieu d'habitation, statut marital, centres d'intérêts, salaire, bien-être subjectif, intelligence, extraversion, etc.). Dans ce genre de plan, le recueil de données se fait généralement par questionnaire ou entretien téléphonique, mais pas seulement. Il est tout à fait possible (et même relativement fréquent) de réaliser des études corrélationnelles en laboratoire, par exemple en faisant passer des batteries de tests cognitifs et en analysant ensuite les corrélations entre les différentes aptitudes que ces tests mesurent.

Les *plans expérimentaux* quant à eux consistent essentiellement en la *manipulation* d'un petit nombre de variables. Dans un plan expérimental, les variables ne sont pas simplement observées. Les participants à une expérience sont *assignés* à des conditions expérimentales. La personne qui conçoit l'expérience *décide* qui sera exposé à telle ou telle situation expérimentale et donc quelle sera sa valeur sur la variable indépendante (VI) d'intérêt. Ensuite, on mesure l'effet de cette manipulation sur la variable dépendante (VD) – un temps de réaction, une opinion, une réaction émotionnelle, un comportement d'approche ou d'évitement, etc.

Les plans expérimentaux sont caractérisés par une très forte validité interne. En effet, en dehors de la manipulation de la VI, tous les paramètres de la situation expérimentale sont soigneusement contrôlés. Toutefois, le prix à payer pour ce degré de contrôle élevé est une validité externe souvent faible ; les situations expérimentales sont souvent assez artificielles et ne reflètent pas toujours bien la complexité de la vie réelle. (Nous reviendrons plus loin sur les questions liées à la validité, les variables parasites et les méthodes de contrôle.) Les plans expérimentaux sont typiquement réalisés en laboratoire, mais pas seulement. Il est en effet tout à fait possible de concevoir une étude expérimentale qui se passe dans la rue, par exemple si on compare le temps où des passants regardent différentes affiches, dont le contenu est manipulé par les expérimentateurs.

Approches transversales et approches longitudinales

Une autre distinction importante est celle existant entre les plans ou approches *transversales* et les approches *longitudinales* (ou à mesures répétées).

La caractéristique centrale d'une étude *transversale* est le fait qu'elle repose sur *un seul temps de mesure* – chaque participant à l'étude ne participe qu'une seule fois. Ce genre d'étude est idéal pour mesurer un grand nombre de variables à relativement moindre coût, notamment à l'aide de questionnaires ou d'entretiens téléphoniques. Du fait de son coût logistique et financier relativement modéré, c'est également le genre d'étude qui est le plus souvent mené auprès de grands échantillons, de plusieurs centaines voire plusieurs milliers de personnes.

Les études *longitudinales*, elles, sont caractérisées par le fait qu'elles incorporent *plusieurs temps de mesure*. Les termes « longitudinal » et « à mesures répétées » sont donc relativement synonymes. Toutefois, en pratique, « longitudinal » implique souvent un long suivi dans le temps, par exemple en mesurant les mêmes variables chaque année pendant plusieurs années, permettant ainsi de mesurer et modéliser du changement. Par contraste, « mesures répétées » implique plutôt un plan expérimental et une durée plus courte, parfois de quelques heures ou quelques minutes seulement.

Vue d'ensemble des plans de recherche

En résumé, les deux distinctions évoquées jusqu'ici donnent lieu à quatre cas de figure typiques représentés dans le Tableau 1 ci-dessous. Dans la section qui suit, nous allons considérer un peu plus en détail les caractéristiques des deux types de plans expérimentaux représentés dans la dernière ligne de ce tableau.

(Les plans quasi-expérimentaux ne sont pas représentés ici ; ils se situent en fait simplement « entre » les plans corrélationnels et les plans expérimentaux. En d'autres termes, dans un plan quasi-expérimental, certaines variables sont manipulées alors que d'autres sont simplement observées. Les plans quasi-expérimentaux n'ont donc aucun attribut propre ; ils constituent simplement un intermédiaire entre l'expérimental et le corrélationnel.)

Tableau 1. Quatre grands types de plan de recherche

	Un seul temps de mesure	**Plusieurs temps de mesure**
On observe les VIs	Plan corrélationnel transversal	Plan corrélationnel longitudinal
On manipule les VIs	Plan expérimental à groupes indépendants	Plan expérimental à mesures répétées

Détails sur les plans expérimentaux à groupes indépendants

Commençons par les plans expérimentaux à groupes indépendants, qui sont aussi parfois appelés « design inter-sujets ». Dans ce type de plan, chaque personne ne passe qu'une seule condition expérimentale. Il y a donc autant de groupes de participants que de conditions expérimentales. S'il y a une condition A et une condition B, nous aurons deux groupes distincts de participants ; s'il y a trois conditions, nous aurons trois groupes, et ainsi de suite.

Un avantage de ce type de plan est que les participants ne sont pas trop sollicités – ils n'ont pas à passer au travers de plusieurs conditions expérimentales. Cela empêche toute sorte d'effets problématiques liés à la fatigue ou à l'apprentissage. D'une manière générale, il n'y a aucun effet d'ordre, qui est une faiblesse caractéristique des plans à mesures répétées.

L'inconvénient en revanche est qu'il y a un risque de non-équivalence des groupes. Par exemple, s'il y a plus d'hommes que de femmes dans un groupe que dans l'autre, plus de participants fatigués dans un groupe que dans l'autre, plus de participants stressés dans un groupe que dans l'autre, etc. La solution typique pour éviter la non-équivalence des groupes est d'*assigner aléatoirement* les participants aux différentes conditions expérimentales. En pratique, il s'agit simplement de *tirer au sort* qui va dans quel

groupe. Une autre approche, qui peut être utilisée en complément à la première, est de contrôler certaines variables d'une façon systématique. On peut par exemple, en plus d'une assignation aléatoire, s'assurer qu'il y a autant d'hommes que de femmes dans chaque groupe.

Détails sur les plans expérimentaux à mesures répétées

Considérons maintenant les plans expérimentaux à mesures répétées, parfois aussi appelés « design intra-sujet ». Dans ce type de plan, il y a un seul ensemble de personnes pour toute l'expérience. En d'autres termes, chaque participant va être amené à passer dans toutes les conditions expérimentales les unes après les autres.

L'avantage principal est qu'il n'y a aucun risque de non-équivalence des groupes, puisque ce sont toujours les mêmes personnes au travers de toutes les conditions. On dit parfois que « le sujet est son propre contrôle ». Cette situation va être avantageuse du point de vue de l'analyse de données car on va pouvoir modéliser ces « effets sujet ». Concrètement, si une personne est très motivée, cette forte motivation sera retrouvée dans les deux conditions expérimentales – et non pas dans une seule comme dans un design à groupes indépendants.

Un inconvénient, en revanche, est qu'à force de répétitions, les participants puissent détecter ce qui est attendu d'eux et biaiser les résultats, volontairement ou non. Il y a aussi simplement des problèmes liés à la fatigue ou à la lassitude. D'une manière générale, tous ces problèmes renvoient à la notion d'*effet d'ordre*. La solution typique pour pallier aux problèmes des effets d'ordre est le *contre-balancement*. Cette technique consiste à diviser l'ensemble des participants en deux sous-groupes. S'il y a deux conditions expérimentales, le premier sous-groupe passera d'abord la condition A puis la B, alors que le second sous-groupe fera l'inverse. On aura donc, par exemple, autant de participants fatigués dans la condition A que dans la condition B.

Le contre-balancement a toutefois des limites. Premièrement, il peut devenir très compliqué à mettre en place s'il y a beaucoup de conditions expérimentales. À titre d'exemple, trois conditions donnent déjà lieu à six ordres différents ABC, ACB, BAC, BCA, CAB et CBA. Cela peut vite devenir un cauchemar logistique ! Par ailleurs, certaines tâches ou certaines situations expérimentales se prêtent mal au contre-balancement et, d'une manière générale, aux plans à mesures répétées. Par exemple, si l'on conçoit une expérience sur les effets de la caféine, il faudra attendre plusieurs heures entre les deux parties de l'expérience ; on ne peut clairement pas mettre un participant dans une condition « sans caféine » alors qu'il vient d'ingérer une heure auparavant une dose de caféine pour la condition « avec caféine ». Problème logistique ici également.

Les 7 problèmes en question

Après ces rappels sur les différents types de plans, nous pouvons enfin aborder nos 7 problèmes méthodologiques, en commençant par deux problèmes très généraux, qui concernent indifféremment tous les types de plans.

Problème 1. Décalage entre la théorie et l'opérationnalisation

Le premier problème porte sur le décalage qu'il peut y avoir entre un concept ou une variable *théorique* et une variable *opérationnalisée*.

Théorique et opérationnalisée: définition et implications

Quand on parle de variables théoriques, il s'agit souvent de « grands » concepts, de notions complexes et abstraites comme l'intelligence, la personnalité ou la créativité. Ces grands concepts, leurs définitions et leurs relations avec d'autres notions complexes suscitent en général des débats théoriques souvent très passionnés.

Les variables opérationnalisées, elles, renvoient à des manifestations concrètes – parfois simplistes, parfois sophistiquées – mais en général beaucoup moins ambiguës. L'avantage est qu'avec les variables opérationnalisées, on sait de quoi on

parle exactement, ce qui est loin d'être évident avec les variables théoriques. Cependant, pour une variable théorique donnée, il n'en existe pas une opérationnalisation, mais une quasi infinité.

De nombreux malentendus viennent donc du fait que les variables théoriques peuvent être opérationnalisées de bien des façons différentes et que, lors de discussions théoriques, on a vite fait d'oublier de préciser de quoi on parle concrètement – de quel type d'opérationnalisation. Par exemple, là où deux personnes parlent de créativité, l'une pense peut-être « plaisir que les enfants éprouvent à s'exprimer en dessinant » alors que l'autre pense aux « pays qui cumulent le plus de prix Nobel ».

La question des liens entre créativité, intelligence et personnalité a effectivement fait couler beaucoup d'encre et donné lieu à des milliers d'articles scientifiques (littéralement). Toutefois, après plus d'un siècle de recherches (littéralement aussi), certaines questions théoriques, en particulier celles qui sont très générales, restent encore très débattues. Il n'est même pas rare de trouver des conclusions contradictoires, même en ne considérant que des études globalement sérieuses et bien faites.

Un problème incontournable… mais qui n'excuse pas tout !

En pratique, il est impossible de s'affranchir des problèmes qui peuvent émerger quand on passe du théorique à l'opérationnalisé (et inversement) : toute conclusion sérieuse à un niveau théorique doit inévitablement se baser sur des variables opérationnalisées. S'il n'y a pas d'opérationnalisation, il n'y a pas de mesure et il n'y a rien de concret ; les choses sont alors pires encore, c'est la porte ouverte à toutes les spéculations les plus sauvages, à toutes sortes d'interminables débats d'opinions purs et simples.

Néanmoins, même dans les recherches relativement sérieuses, qui ne renoncent pas à toute forme de mesure, on constate parfois qu'une opérationnalisation particulièrement pauvre empêche toute remontée crédible au niveau théorique. Par exemple, il serait complétement abusif de parler d'intelligence à un niveau théorique si l'opérationnalisation consiste en une bête tâche de calcul mental de cinq minutes. Hélas, ce genre d'études existe bel et bien, même si elles sont en général publiées dans des revues de seconde zone – ce qui n'empêche pas que certains résultats qui en sont issus soient parfois relayés par la presse.

Au-delà de ces cas un peu « limites », il faut admettre que le hiatus entre le théorique et l'opérationnalisé est, jusqu'à un certain point, incompressible. En effet, toute opérationnalisation implique nécessairement des choix, des contraintes et des limites. Il est par exemple absolument impossible de concevoir une étude sur la créativité qui incorpore ne serait-ce que 10% de toutes les opérationnalisations proposées pour mesurer cet insaisissable concept. Quand on décide de faire une étude sur la créativité, on va donc inévitablement devoir faire des choix d'opérationnalisation qui vont considérablement réduire l'envergure du concept théorique – par définition, une variable opérationnalisée est toujours plus limitée, beaucoup moins grandiose que sa grande sœur théorique.

Conclusion sur le décalage entre théorie et opérationnalisation

En définitive, il faut donc toujours faire un travail d'équilibriste entre la théorie et l'opérationnalisation en évitant trois écueils principaux : (1) assimiler des mesures simplistes à des concepts complexes ; en effet, des variables théoriques complexes requièrent le plus souvent de multiples opérationnalisations complémentaires ; (2) manquer de transparence (le plus souvent par négligence) sur le type d'opérationnalisation qui est en arrière-plan de toute discussion théorique ; (3) renoncer à toute forme de mesure et d'opérationnalisation sous prétexte que « c'est immesurable », ce qui revient purement et simplement à renoncer à toute approche scientifique.

Problème 2. Mesure de mauvaise qualité

Le deuxième problème, assez étroitement lié au premier, concerne également tous les types de plans de recherche. Il s'agit de l'utilisation de mesures de mauvaise qualité ou, en termes plus

techniques, de mesures aux médiocres propriétés psychométriques. Il s'agit en quelque sorte d'une version aggravée du premier problème. Là où le premier problème concernait plutôt les difficultés à connecter le théorique et l'opérationnel, nous allons maintenant considérer les problèmes qui font qu'une opérationnalisation est tout simplement mauvaise et donc foncièrement inutilisable. La psychométrie retient en général trois caractéristiques clés pour déterminer si une mesure est de bonne qualité : la validité, la fidélité et la sensibilité. Regardons de plus près ces trois notions.

Validité

La validité peut porter soit sur une *recherche* entière, soit sur une *mesure* en particulier.

Dans le premier cas, on parle souvent de validité interne et de validité externe. La *validité interne* renvoie à la « solidité interne » de l'étude ; elle est haute si le degré de contrôle d'une étude est haut, en particulier en termes de contrôle des variables parasites. Comme nous l'avons évoqué plus haut, la validité interne est souvent le point fort des études expérimentales. La *validité externe*, quant à elle, renvoie à la généralisabilité des résultats (on parle parfois aussi de validité écologique). En d'autres termes, est-ce que ce qui a été observé dans une étude est généralisable à la « vraie vie », c'est-à-dire à des situations plus vastes, plus complexes, moins aseptisées que celle qui a été étudiée ? Si l'on trouve que la caféine permet de recopier plus rapidement une liste de symboles, peut-on conclure que « la caféine améliore la productivité » ? Si l'on trouve que recevoir 10 CHF permet d'améliorer sensiblement l'humeur, peut-on conclure que « l'argent fait le bonheur » ? Ce point est étroitement lié au précédent problème concernant l'opérationnalisation. Il est également lié aux questions de généralisabilité que nous aborderons plus loin (problème 7).

Passons maintenant aux formes de validité qui concernent une mesure en particulier. Il y en a essentiellement deux types : la validité convergente et la validité discriminante.

La *validité convergente* (appelée aussi parfois validité de construit) renvoie au fait que « l'outil mesure bien ce qu'il doit mesurer ». Si cela a l'air simple, ce n'est pourtant pas évident de le démontrer. En pratique, il y a deux approches pour évaluer la validité convergente d'une mesure. La première consiste simplement à comparer l'outil en question avec un autre qui est censé mesurer la même chose ; les deux outils devraient être assez fortement corrélés, sinon c'est qu'ils mesurent deux choses différentes. La seconde approche consiste à évaluer la pertinence pratique de l'outil en question. Dans le cas d'une tâche d'intelligence, si sa validité est bonne, on s'attend à ce qu'elle ait une certaine valeur pratique. Admettons que l'on ait besoin de recruter des personnes très intelligentes pour exécuter des tâches très complexes. Si l'on sélectionne ces personnes à l'aide du test d'intelligence en question, on s'attend à ce que ces personnes s'en sortent ensuite plutôt bien dans lesdites tâches complexes (de ce fait, on parle aussi parfois de *validité prédictive*). Sinon, le test n'est pas valide.

La *validité discriminante*, qui est en quelque sorte le pendant de la validité convergente, renvoie au fait que « l'outil ne mesure pas ce qu'il ne doit pas mesurer ». Si l'on garde notre exemple du test d'intelligence, on veut non seulement qu'il mesure bien l'intelligence (validité convergente) mais aussi qu'il ne mesure pas des choses qui n'ont rien à voir, comme l'anxiété ou l'extraversion. Si un test d'intelligence est très bien réussi par des personnes très extraverties et systématiquement raté par des personnes très anxieuses, il s'agit clairement d'un mauvais test d'intelligence, avec une validité discriminante médiocre. Pour s'assurer qu'un outil de mesure a une validité discriminante satisfaisante, il faut donc tester qu'il ne corrèle pas avec des variables qui n'ont rien à voir, ainsi qu'il corrèle peu avec des variables qui sont similaires mais néanmoins distinctes (on souhaiterait par exemple que le test d'intelligence ne corrèle que modérément avec un test de créativité ; une corrélation trop forte signifierait que les deux tests sont indistincts).

Fidélité

Si la validité d'un test est nécessaire, elle n'est hélas pas suffisante. Une mesure de qualité doit aussi avoir une forte *fidélité*. La fidélité renvoie à la précision de la mesure. Le score issu du test est-il fiable, précis ? Est-ce que l'erreur de mesure est faible ? Un test fidèle aura une faible erreur de mesure et reflétera bien ce qu'on appelle parfois le « score vrai » d'une personne, c'est-à-dire le score « idéal », sans erreur de mesure. Il existe différents types de fidélité, mais tous reflètent cette idée fondamentale de faible erreur de mesure.

On parlera notamment de *fidélité inter-item* pour les questionnaires basés sur plusieurs questions, dont on fait ensuite la moyenne pour obtenir un score total. Si j'utilise un questionnaire de 10 questions pour mesurer l'extraversion, je m'attends à ce que ces dix questions (ces dix items) soient très fortement corrélées. Si ce n'est pas le cas, mon questionnaire mesure tout et n'importe quoi et, une fois que j'aurai fait la moyenne des 10 questions, ma variable finale d'extraversion ne sera rien de plus qu'une superbe compilation d'erreurs de mesure. Evidemment, ce n'est jamais noir ou blanc, et il est rare que la fidélité soit totalement nulle. Il existe donc des indices statistiques (typiquement l'*alpha de Cronbach*) qui permettent de synthétiser les corrélations inter-item et d'évaluer rapidement si la fidélité est excellente (proche de 1, qui représente le maximum théorique), plutôt bonne (autour de .70 ou .80) ou à la limite de l'acceptable (autour de .60).

Un autre type de fidélité est la *fidélité inter-juges*. Cette fidélité est particulièrement utile dans les recherches qui impliquent des évaluations subjectives. Imaginons par exemple que l'on demande à trois chercheurs d'évaluer, à l'aide d'une grille prévue à cet effet, le degré d'agressivité manifestée par des enfants lors d'un jeu. On s'attend à ce que les trois avis soient à peu près similaires, que les scores des trois juges soient fortement corrélés, comme devraient l'être nos 10 items du questionnaire d'extraversion. Si un enfant est jugé particulièrement agressif par un juge et pas du tout par un autre, on a un problème de fidélité inter-juges. Cette fidélité inter-juges repose fortement sur la qualité des outils d'observation – grille ou autre forme de protocole – qui doivent être aussi clairs et standardisés que possible afin de permettre aux évaluateurs d'être le plus précis possible. Cette notion de fidélité inter-juges est importante en psychologie car de nombreuses variables d'intérêt ne sont tout simplement pas appréhendables à l'aide de mesures objectives.

(Un dernier type de fidélité, un peu à part, est la *fidélité test-retest*. Cette fidélité-là renvoie à la stabilité d'une mesure au travers du temps. Si l'on mesure par exemple l'extraversion d'un échantillon d'adultes aujourd'hui, on s'attend à trouver des résultats similaires si on administre une seconde fois le questionnaire un mois plus tard. En d'autres termes, on s'attend donc à trouver une forte corrélation entre la première mesure et la deuxième. Cela ne concerne évidemment que les construits qui sont relativement stables ou, du moins, qui ne se développent que très lentement, comme la personnalité chez l'adulte.)

Avant d'aborder la sensibilité, soulignons le fait qu'il est tout à fait possible qu'un outil soit *valide mais peu fidèle*. Dans notre exemple de l'agressivité des enfants, il est possible que nos trois juges ne soient que très peu en accord (fidélité faible) mais que, pour autant, la moyenne de leurs trois évaluations permette néanmoins de prédire (modestement) d'autres manifestations d'agressivité de ces enfants dans un autre contexte (bonne validité). Evidemment, plus la fidélité est faible, plus le risque est grand que la validité soit faible elle aussi ; à l'extrême, une fidélité nulle est finalement équivalente à de la pure erreur de mesure, qui, évidemment, n'a aucune valeur prédictive – autant utiliser des chiffres tirés au sort ! À l'inverse, il est également possible qu'un outil soit *fidèle mais peu valide*. En gardant le même exemple, cela impliquerait que nos trois juges sont tout à fait d'accord entre eux (fidélité forte) mais que ces scores ne permettent pas du tout de prédire d'autres manifestations d'agressivité de la part des enfants (validité [convergente/prédictive] faible ou nulle).

Sensibilité

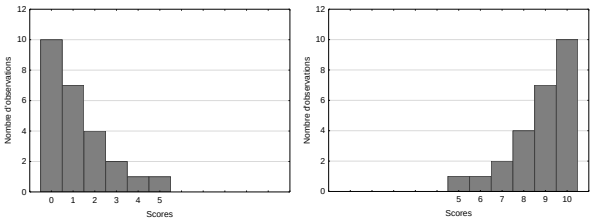
La sensibilité renvoie à la capacité d'un instrument à détecter les différences entre les

personnes que l'on observe. Pour prendre un exemple extrême, une question oui/non (« Aimez-vous faire du jogging, oui ou non ? ») a une sensibilité minime : elle ne permettra pas de distinguer le joggeur du dimanche du passionné de marathon, ni toutes les personnes entre les deux. Pour avoir une meilleure sensibilité, il faudrait poser plusieurs questions relatives à la fréquence et l'intensité des séances de jogging, le plaisir ressenti, etc. On voit donc à travers cet exemple simple, qu'à moins d'utiliser des mesures vraiment simplistes, la sensibilité est rarement un problème. De fait, la sensibilité est un peu « la troisième roue du carrosse » de la psychométrie ; on s'en préoccupe généralement beaucoup moins que des deux autres. Ceci est en partie justifié, car en pratique, il est beaucoup plus rare de se heurter à des problèmes de sensibilité qu'à des problèmes de validité ou de fidélité. Toutefois, la sensibilité renvoie également aux notions d'effet plancher et d'effet plafond, qui, eux, peuvent être problématiques.

L'*effet plancher* renvoie à une mesure ou un test qui est trop « difficile » pour une population donnée. Par exemple, si on administre à des enfants un test d'intelligence pour adultes ou si on utilise un questionnaire pour mesurer la dépression dans une population non clinique. Tous les scores seront alors très bas, donnant lieu à un histogramme très « tassé à gauche » (voir la Figure 1 ci-dessous). On constate une disproportion de petits scores et quasiment aucun score élevé, alors que l'outil de mesure le permet en principe. La distribution d'une mesure qui présente un effet plancher souffre d'asymétrie positive ; il n'y a pas assez de scores positifs (au-dessus de la moyenne).

L'*effet plafond* correspond au problème inverse : on a utilisé un outil de mesure trop « facile » pour la population d'intérêt – un test d'intelligence pour enfants administré à des adultes, un questionnaire de mauvaise humeur conçu pour la population générale administré à une population clinique souffrant de troubles de l'humeur, etc. Dans ce cas, tous les scores seront très hauts, donnant lieu à un histogramme très « tassé à droite ». On constate une disproportion de scores élevés, le score maximum de l'outil est atteint par un très grand nombre de personnes, alors qu'il y a très peu de faibles scores, même si, à nouveau, l'outil de mesure permettrait d'en observer en principe.

Figure 1. Représentation graphique de l'effet plancher (à gauche) et de l'effet plafond (à droite), dans les deux cas avec un outil de mesure fictif dont les scores possibles pourraient aller de 0 à 10



Ces deux problèmes d'effet plancher et plafond sont à éviter autant que possible car ces distributions non normales peuvent poser problème lors de l'analyse de données. Plus fondamentalement, les effets plancher et plafond reflètent le fait qu'une partie des personnes de l'échantillon a été mal mesurée : dans le cas de l'effet plancher, il y a beaucoup de scores bas qui sont potentiellement surestimés (ceux qui sont tout à gauche), alors que dans le cas de l'effet plafond il y beaucoup de scores haut qui sont potentiellement sous-estimés (ceux qui sont tout à droite).

Conclusion sur la qualité de la mesure

En conclusion, on admettra volontiers que le travail psychométrique qui permet de développer des outils de qualité n'est pas simple. Développer une nouvelle mesure, c'est une activité de recherche à part entière. À partir de là, lorsque l'on conçoit une recherche, deux solutions s'offrent à nous : soit on développe un nouvel outil et on est alors prêt à fournir le travail psychométrique qui permet de démontrer sa qualité ; soit on utilise des outils connus, dont la qualité a déjà été démontrée par d'autres.

Par extension, lorsqu'on lit de la littérature scientifique, il est important de rester très critique en ce qui concerne les instruments de mesure. Les outils utilisés dans une recherche sont-ils (re)connus ? Sont-ils issus d'un travail de validation sérieux ? Si les outils de mesure sont nouveaux, peut-on admettre qu'un travail de validation suffisamment sérieux a été fait ? Les questions de validité et de fidélité, en particulier, sont-elles abordées en détail par les auteurs ?

Quels éléments sont proposés par ces derniers pour démontrer les bonnes qualités psychométriques de leurs outils ?

Problème 3. Mauvais contrôle des variables parasites

Le troisième problème concerne le vaste monde des variables parasites, variables confondues et autres « troisièmes variables ». Contrairement aux deux premiers problèmes, les incarnations de ce troisième problème sont sensiblement différentes en fonction des types de plan de recherche. Typiquement, les notions (quasi interchangeables) de variable parasite et de variable confondue renvoient à une faiblesse majeure dans un plan expérimental, alors que la notion de « troisième variable » renvoie plutôt à une réalité quasiment inévitable des designs corrélationnels. Détails ci-dessous.

Variables parasites (ou confondues)

Une variable parasite (ou variable confondue) est une variable – ou, plus généralement, un phénomène – qui va interférer profondément et irrémédiablement avec l'interprétation des résultats. On utilise le terme (très négatif) de variable parasite essentiellement pour les plans expérimentaux car, par définition, ce type de plan est censé offrir un niveau de contrôle optimal et donc éviter les variables parasites. Plus concrètement, une variable parasite est *une variable dont les valeurs sont confondues avec celles de la variable indépendante*. Si quelqu'un a une valeur « A » sur la VI, il aura une valeur « A » sur la variable parasite ; si quelqu'un a « B » sur la VI, il aura « B » sur la variable parasite, etc.

Imaginons par exemple que l'on conçoive une expérience permettant de tester une nouvelle méthode pour arrêter de fumer. Au lieu d'attribuer aléatoirement les participants soit au « groupe traitement », soit au « groupe contrôle », on les laisse choisir librement. Il est alors possible que l'on se retrouve dans la situation suivante : les personnes très motivées à arrêter de fumer vont choisir le groupe traitement (car elles veulent maximiser leur chance d'effectivement réussir à arrêter de fumer) alors que celles qui sont peu motivées vont choisir le groupe contrôle (car elles s'imaginent que ça sera moins pénible, qu'elles auront une excuse si elles n'arrivent pas à arrêter). Une fois l'étude terminée, si l'on observe que les personnes du groupe traitement ont réussi à arrêter de fumer mais pas les autres, il est impossible de savoir si c'est grâce au traitement ou à la motivation de départ des participants.

Par conséquent, les variables parasites constituent une menace majeure à la validité interne. Les conclusions de l'étude sont au mieux rendues très incertaines et au pire complètement anéanties : on ne sait pas si les effets observés sont dus à la VI ou à la variable confondue.

La « troisième variable »

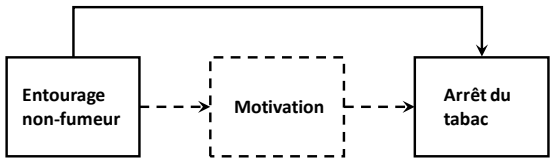
La notion de troisième variable est similaire à celle de variable parasite, en moins dramatique. Elle concerne plutôt les plans de recherche corrélationnels et ne constitue pas une insuffisance méthodologique majeure comme celle évoquée ci-dessus. En effet, dans un plan corrélationnel, aucune VI n'est manipulée, il est donc techniquement impossible de contrôler l'effet de toutes les autres variables qui peuvent avoir une influence sur la variable dépendante.

Par contraste avec l'exemple ci-dessus, imaginons que l'on réalise une grande enquête par questionnaire permettant de savoir quels sont les facteurs qui déterminent si une personne va réussir à arrêter de fumer ou pas. Dans une telle étude, on ne va rien manipuler ; il n'y a ni traitement ni groupe contrôle ; on va simplement mesurer (observer) toutes sortes de facteurs que l'on suppose importants dans l'arrêt de la consommation de cigarettes. Admettons qu'on décide de mesurer des variables liées à la motivation (forte ou faible volonté d'arrêter de fumer) et quelques variables supplémentaires, disons des variables liées à la personnalité (propension à l'anxiété) et des variables liées à l'environnement (stress dans la vie privée et stress dans la vie professionnelle). Imaginons que l'on réalise notre étude et que tout se passe pour le mieux ; l'étude est bien conçue, bien implémentée, avec de bons outils de mesure, etc., et n'a donc pas de limitation méthodologique particulière. Après analyse des données, on trouve que la motivation est le

facteur déterminant et que l'effet du stress est négligeable. L'étude est publiée dans *Nature* ; les chercheurs et chercheuses font fortune.

Un jour, une nouvelle étude est lancée. Il s'agit de la même étude que la précédente, mais avec, en plus, la prise en compte de variables liées à l'entourage, en particulier la proportion de fumeurs dans l'entourage immédiat. Les résultats montrent alors que c'est ce facteur qui est déterminant. Mieux encore, la proportion de fumeurs dans l'entourage explique à la fois le fait de parvenir à arrêter de fumer *et* la forte motivation : les fumeurs qui fréquentent peu de fumeurs sont les plus motivés à arrêter, et en étant très motivés à arrêter, ils y parviennent effectivement.

Figure 2. Le lien entre X et Y (motivation et réussite de l'arrêt de consommation du tabac) expliqué par une troisième variable (entourage majoritairement non-fumeur)



Il s'agit là d'un cas typique de « troisième variable ». Une étude trouve un lien entre X et Y ; une autre étude parvient à démontrer que ce lien est en fait expliqué par une troisième variable Z (cf. Figure 2). Il n'y a pas d'erreur méthodologique intrinsèque à la première étude. On peut certes reprocher aux chercheurs d'avoir manqué de présence d'esprit en oubliant d'inclure des variables liées à l'entourage, mais ce n'est pas en soi une erreur méthodologique. L'erreur se situe plutôt au moment de la conclusion, si les chercheurs oublient que X et Y ne racontent pas toute l'histoire et que Z a peut-être aussi son mot à dire.

Les variables contrôles

D'une manière générale, les variables contrôles sont celles qui vont permettre de renforcer la solidité et la précision des conclusions que l'on peut tirer concernant le lien entre deux autres variables. Avec tout ce que nous avons vu jusqu'ici, il devient assez simple d'expliquer ce que font concrètement les variables contrôles.

Dans un plan expérimental, une variable contrôle est une sorte de « troisième variable » que l'on va mesurer car on sait qu'elle est potentiellement pertinente. Ce type de démarche représente déjà un certain raffinement de l'approche expérimentale, car en principe, il suffit d'assigner au hasard les participants aux conditions expérimentales, et le tour est joué. Toutefois, en pratique, il arrive que l'on utilise des variables contrôles supplémentaires.

Imaginons que l'on veut tester l'effet d'un nouveau médicament pour guérir la dépression. Pour diverses raisons (budget, nombre de patients disponibles), il n'est pas réaliste de concevoir une étude randomisée avec deux groupes de 100 participants. On doit se débrouiller avec deux groupes de 15 personnes. Avec des effectifs si petits, il est possible que, malgré une assignation aléatoire dans les règles de l'art, on joue d'un peu de malchance et que les groupes ne soient pas tout à fait équivalents – on redoute par exemple d'avoir les deux patients les plus déprimés dans le même groupe. Pour se prémunir de ce problème, on peut décider de mesurer l'intensité de la dépression chez tous les patients avant de procéder à l'assignation aléatoire. Avec cette mesure en plus, on peut alors optimiser notre assignation aléatoire et répartir les participants comme suit : le plus déprimé va dans le groupe A, le second plus déprimé dans le groupe B, le troisième dans le groupe A, le quatrième dans le B, etc.

Dans un plan corrélationnel, la variable contrôle est simplement une variable que l'on mesure afin d'écarter des explications alternatives à l'origine d'un lien constaté entre deux autres variables. En d'autres termes, les variables contrôles sont simplement des « troisièmes variables » que l'on a effectivement mesurées et dont on va pouvoir tester l'effet éventuel. Dans notre exemple sur la cigarette, en admettant que c'est le lien entre motivation et arrêt de la consommation de cigarettes qui intéresse le plus les chercheurs, toutes les autres variables que nous avons vues peuvent être envisagées comme des variables de contrôle. Il existe aussi des variables de contrôle « standard » ou « incontournables » que l'on

mesure quasiment systématiquement, telles que l'âge et le sexe.

Conclusion sur les variables parasites et contrôles

Pour conclure, ce que nous retiendrons ici c'est que la notion de variable parasite (ou confondue) est étroitement associée aux plans de recherche expérimentaux, en l'occurrence plans expérimentaux de piètre qualité. S'il y a variable parasite, c'est qu'il y a faiblesse méthodologique. Les variables parasites sont en général évitées avec des procédures telles que l'assignation aléatoire ou le contre-balancement. La notion de variable confondue est également en lien avec la notion d'erreur systématique, traitée ci-après.

La notion de troisième variable, quant à elle, n'implique pas automatiquement une faiblesse méthodologique. C'est plutôt une variable que l'on a oublié de mesurer (ou pas pu mesurer) et qui pourrait être pertinente pour expliquer le lien entre deux autres variables. Si l'on identifie clairement une troisième variable potentielle et qu'on la mesure effectivement, elle prend alors le statut de variable contrôle. Cette variable n'est plus seulement hypothétique ; on l'a identifiée et mesurée, ce qui va permettre de tester son effet et de la mettre en concurrence avec d'autres prédicteurs (typiquement dans un modèle de régression linéaire multiple).

Notons enfin que la différence entre une variable contrôle et une variable indépendante est essentiellement fonction du point focal de telle ou telle recherche. En effet, les variables contrôles des uns sont les variables indépendantes des autres : tel chercheur contrôle l'âge systémiquement dans toutes ses recherches, mais sans s'y intéresser en particulier, alors que telle chercheuse, qui travaille sur le vieillissement cognitif, s'intéresse à l'âge de façon bien plus approfondie – et il n'est pas une seule de ses études où l'âge n'est la VI star.

Problème 4. Mauvaise gestion de l'erreur

Ce quatrième problème qui concerne « l'erreur » (vaste chapitre !) en cache en fait deux bien distincts : le problème de l'erreur systématique et le problème de l'erreur aléatoire. Les différents types de plan sont concernés par différentes manifestations de ces deux formes d'erreur. Découvrons cela progressivement.

Erreur systématique

L'erreur systématique tout d'abord. Pour faire court : c'est le mal incarné. L'erreur systématique est étroitement liée à (pour ne pas dire à la source de) deux autres problèmes majeurs : les biais de mesure et les variables confondues.

Les *biais de mesure* représentent une surévaluation ou une sous-évaluation systématique de ce que l'on cherche à mesurer. On a souvent des problèmes de biais de mesure lorsqu'on pose des questions sur des sujets dits « sensibles ». Ainsi, les violences domestiques, la consommation d'alcool et l'adhésion à des idées subversives vont avoir tendance à être systématiquement sous-évaluées. À l'inverse, la gentillesse, la générosité et tous les comportements vertueux vont être systématiquement surévalués. Evidemment, dans un cas comme dans l'autre, ce n'est pas souhaitable. Pour ce genre de sujets sensibles, il faut donc soigneusement concevoir les outils de mesure pour limiter ces problèmes, par exemple en ne formulant pas les questions d'une façon trop directe ou trop moralisatrice.

Il arrive également que l'erreur systématique soit une source de *variable confondue*. Par exemple, dans le cas d'un plan expérimental à deux groupes indépendants, nous avons un problème si tous les participants du groupe A réalisent l'expérience le matin à 10h alors que tous ceux du groupe B la passent à 14h. Les seconds seront en effet probablement dans des conditions d'éveil moins favorables et auront de ce fait des scores moins bons. Cette variable "moment de la journée" mène donc à une erreur systématique (sous-estimation de tous les scores issus de la tranche "14h") et est totalement confondue avec la variable "condition expérimentale", puisque ce sont tous les participants du groupe B et aucun du groupe A qui ont participé à 14h. Le même genre de malédiction peut frapper les plans à mesures répétées, typiquement s'il n'y a pas de contre-balancement et que tous les participants passent la condition A avant la condition B, ou l'inverse.

Dans un tel cas de figure, en admettant qu'on constate ensuite une différence entre les deux conditions, il est dès lors impossible de déterminer si (1) il y a une différence due aux conditions expérimentales ; (2) il y a une différence due à l'heure de la journée ou (3) une différence due aux conditions *et* à l'heure. Et s'il n'y a aucune différence, il est impossible de déterminer si (1) les conditions A et B sont équivalentes ou si (2) des différences liées aux conditions et celles liées à l'heure s'annulent mutuellement. Ce genre de situation, qui constitue une insuffisance méthodologique majeure (mais pas toujours aussi évidente que dans notre exemple) est absolument impossible à rattraper lors de l'analyse de données.

L'erreur systématique peut également frapper fort – et faire très mal – dans des recherches qui impliquent des évaluateurs subjectifs, dans des procédures telles que celle que nous avons évoquée dans la section sur la fidélité. Reprenons notre exemple dans lequel des chercheurs doivent évaluer le degré d'agressivité de différents enfants lorsqu'ils jouent. Imaginons plus spécifiquement que l'on se trouve dans le cadre d'une expérience où l'on a cherché à savoir si le visionnage de dessins animés violents avait une influence sur l'agressivité dans le jeu. Dans une condition « violence », des enfants ont regardé des dessins animés violents ; dans une condition « non-violence », d'autres enfants ont regardé des dessins animés non violents. Les enfants ont joué après avoir regardé le dessin animé d'une sorte ou d'une autre ; ces moments de jeu ont été enregistrés sur vidéo, et nous en arrivons au moment où des évaluateurs doivent juger l'agressivité des enfants.

Dans un tel contexte, un comportement ambigu (prendre un jouet, bousculer légèrement, jouer à la guerre, etc.) sera très vraisemblablement jugé plus agressif si l'évaluateur sait que l'enfant en question était dans la condition « dessin animé violent ». Sans le vouloir, si les évaluateurs savent dans quelle condition était quel enfant, ils seront plus prompts à évaluer un comportement ambigu comme agressif ; à l'inverse, s'ils savent que l'enfant était dans la condition « dessin animé non violent », ils seront probablement plus indulgents et se diront que « ce n'était peut-être pas fait exprès » ou que « ce n'est pas si agressif que ça ». Et là, c'est le drame : erreur systématique. En effet, si l'on trouve ensuite une différence entre les deux groupes, impossible de savoir si c'est vraiment à cause du dessin animé ou si c'est à cause du biais d'observation des évaluateurs. Le seul moyen d'éviter cela consiste à rendre les évaluateurs *aveugles* aux conditions expérimentales : les personnes en charge de faire ces évaluations de comportement ne doivent pas savoir dans quelle condition expérimentale étaient les enfants qu'ils sont en train d'évaluer.

Idéalement, on recherchera même à implémenter un *double-aveugle* : (1) les participants ne devraient pas savoir dans quelle condition expérimentale ils se trouvent, ni ce que font les autres groupes, sans quoi ils risquent de modifier leur comportement (se conformer aux attentes des expérimentateurs, chercher à compenser une situation qu'ils jugent défavorable, etc.) et (2) dans la mesure du possible, la personne qui fait passer l'expérience (celle qui accueille les participants, fait passer les tests ou fait différents types d'observations) ne devrait pas non plus savoir dans quelles conditions se trouvent les participants, sinon il y a toujours un risque que cette personne se comporte (même sans le vouloir) d'une façon systématiquement différente avec les participants de chaque condition. Naturellement, ce n'est pas toujours possible d'implémenter un double-aveugle.

Erreur aléatoire

L'erreur aléatoire, elle, par définition, « va un peu dans tous les sens », si l'on peut dire. Elle va parfois mener à des surévaluations, d'autres fois à des sous-évaluations. Il s'agit d'une erreur beaucoup plus inoffensive (sauf si elle atteint des proportions invraisemblables) et qui est même « naturelle », inévitable, lorsqu'on mesure n'importe quel phénomène. À tel point que l'erreur de mesure (sous-entendue aléatoire) a sa place dans la théorie classique des tests, qui stipule que *Score observé = score vrai + erreur de mesure*.

Ceci implique clairement que le « score vrai » est inaccessible, qu'on ne peut s'en approcher qu'à l'aide d'un score observé, qui contient inévitablement de l'erreur – parce que

l'instrument de mesure utilisé est imparfait, parce que la personne qui passe le test est particulièrement en forme ou particulièrement fatiguée ce jour-là, parce que les consignes sont interprétées d'une façon légèrement différente d'un participant à l'autre, parce que les conditions d'éclairage et de chauffage sont légèrement changeantes d'un jour à l'autre, etc.

Toutefois, ce type d'erreur aléatoire ne fait que diminuer la précision de la mesure, mais ne cause aucun biais particulier. En reprenant l'exemple évoqué plus haut sur la confusion entre « condition A et passation à 10h » et « condition B et passation à 14h », on peut éviter le problème d'erreur systématique simplement en assignant aléatoirement les participants du groupe A et du groupe B à l'une ou l'autre des deux tranches horaires. On aura donc toujours de l'erreur liée au moment de la journée, mais elle sera cette fois-ci répartie aléatoirement entre les deux groupes (parfois des participants du groupe A seront favorisés en passant le matin, parfois ce seront des participants du groupe B qui seront favorisés). Dans l'ensemble, on a donc de la surestimation et de la sous-estimation dans les deux groupes, mais rien de systématique. Donc a priori pas de problème.

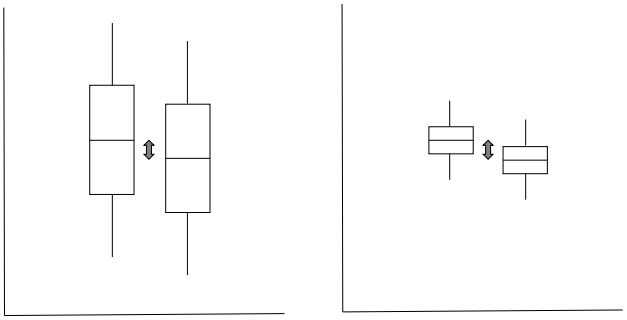
Conclusion sur le problème de la gestion de l'erreur

Comme nous l'avons vu, l'erreur *systématique* n'est guère pardonnable. Une étude qui est contaminée par de l'erreur systématique risque fort d'être purement et simplement ininterprétable.

L'erreur *aléatoire* est en revanche tout à fait tolérable, voire normale. Néanmoins, en pratique, en particulier dans le cadre d'une expérience, on essaye autant que possible de minimiser l'erreur aléatoire en uniformisant les conditions de passation.

En effet, plus il y a d'erreur (aléatoire) dans nos observations, plus les effets que nous cherchons à mettre en évidence seront difficiles à détecter. Dans une expérience, la variabilité des conditions de passation va mener à une augmentation pure et simple de la variance aux travers des conditions expérimentales. Or, plus la variance entre les conditions expérimentales est grande, plus il sera difficile de détecter une différence de moyennes (Figure 3).

Figure 3. Une différence de moyennes donnée (petite flèche grise) est plus difficile à détecter si la variabilité est forte (boxplots de gauche) que si elle est faible (boxplots de droite)



C'est pour le même genre de raison que l'on cherche à avoir des questionnaires avec une fidélité élevée et de forts accords inter-juges. D'une manière générale, plus il y a d'erreur ou « de bruit » dans les données, plus il est difficile de détecter des effets significatifs. On a donc tout intérêt, autant que possible, à tâcher de minimiser l'erreur aléatoire.

Problème 5. Mauvais choix de groupe contrôle

Notre cinquième problème concerne uniquement les plans expérimentaux. En effet, la notion de groupe contrôle n'a pas vraiment lieu d'être dans un plan corrélationnel. (Dans un tel contexte, on parle parfois de groupe de référence ou de groupe de comparaison, mais c'est une notion qui est relativement éloignée de celle de groupe contrôle dans le sens qui va nous occuper ici ; nous ne nous y attarderons donc pas.)

Le groupe contrôle est une des pierres angulaires des designs expérimentaux. En effet, c'est essentiellement au groupe contrôle que nous allons comparer tous types de manipulations expérimentales. Le groupe contrôle constitue donc le plus souvent un point de référence fondamental. Pour cette raison, il est crucial de bien réfléchir à ce que fait et ne fait pas le groupe contrôle. En effet, selon la façon dont est définie le groupe contrôle, il est facile de faire « apparaître » ou « disparaître » des différences…

Groupe contrôle trop différent du groupe expérimental

Le premier « extrême », celui qui va avoir tendance à faire « apparaître » des résultats, est le cas du groupe contrôle qui ne fait « rien ». Prenons l'exemple des psychothérapies, qui est assez parlant ici. Imaginions que j'ai inventé une potion magique contre la dépression (non, je ne dévoilerai pas la recette). Je décide de démontrer à l'aide d'une expérience fort rigoureuse que ma potion magique permet d'améliorer l'état des patients souffrant de dépression. Mon protocole est simple : je reçois les participants du groupe « potion magique » à mon bureau, situé dans un prestigieux hôpital universitaire (on peut rêver), ceci à raison d'une fois par semaine pendant 10 semaines. À chaque fois, je discute 15 minutes avec ces patients, leur demande comment ils vont, etc., puis je leur fais boire le remède miracle. Le groupe contrôle quant à lui, ne fait rien. J'ai assigné aléatoirement les patients à l'un ou l'autre groupe, j'utilise des outils validés pour évaluer la dépression, etc. ; mon design expérimental est a priori irréprochable. L'étude se termine, j'ouvre les données dans mon logiciel préféré, je les analyse et je découvre que la dépression dans le groupe « potion magique » a diminué de 10%, contre 0% dans le groupe contrôle. La différence est significative, les postulats de l'analyse sont respectés, tout va pour le mieux.

Un problème ? Peut-être… Une manière de le voir est de le considérer sous l'angle de la variable confondue. En effet, plusieurs choses diffèrent systématiquement entre mon groupe contrôle et mon groupe expérimental : non seulement le groupe contrôle n'a pas reçu la potion magique, mais, en plus, il n'a pas eu le privilège de venir dans un somptueux bureau, ni l'opportunité de parler de ses problèmes chaque semaine. La différence de résultats que j'observe est « forcée » par cette méthodologie médiocre. Si cela paraît évident ici, il est bien des cas où ce n'est pas aussi spectaculaire – et plus pernicieux.

Autre exemple : je me lance maintenant dans une étude portant sur l'effet positif de la méditation sur l'humeur. Cette fois-ci, je suis bien décidé à mieux soigner ma méthodologie. Les participants du groupe expérimental viendront à 10 séances de méditation de 30 minutes, données par un instructeur certifié. Je prévois de faire venir les participants du groupe contrôle à la même fréquence, pour la même durée, au même endroit, sauf qu'au lieu de de suivre les instructions pour méditer, ils seront simplement invités à s'asseoir et à attendre pendant 30 minutes. Apparemment, c'est déjà un peu mieux que l'autre expérience. Mais est-ce suffisant ? Un problème est que dans un cas, il y a une personne qui guide la séance et dans l'autre pas – indépendamment des instructions elles-mêmes.

Groupe contrôle trop similaire au groupe expérimental

À mesure que je vais tenter d'augmenter mon honnêteté scientifique, si l'on peut dire, en supprimant variable parasite après variable parasite, je risque de me rapprocher d'un autre extrême, qui va avoir tendance à faire « disparaître » les différences (ou en tout cas à les estomper sérieusement), car le groupe contrôle va devenir de plus en plus similaire au groupe expérimental.

Je pourrais par exemple décider d'ajouter un instructeur à mon groupe contrôle. De cette façon, il y aura un instructeur dans chaque groupe, et si je trouve une différence entre les deux groupes, on ne pourra pas me dire que c'est parce que dans un cas « on s'occupe des participants » alors que dans l'autre « ils sont livrés à eux-mêmes » et que les différences que j'observe sont peut-être dues à cela, et pas à la méditation en elle-même.

Néanmoins, si j'ajoute un instructeur dans le groupe contrôle, il va bien falloir qu'il dise quelque chose ; il ne va pas juste rester planter là – ça n'arrangerait rien au problème des variables confondues. Je décide donc que dans la condition contrôle, l'instructeur dira simplement quelque chose comme « détendez-vous, respirez calmement » une fois toutes les cinq minutes. Mais est-ce suffisant ? Si l'autre instructeur donne des instructions quasiment sans discontinuer, est-ce vraiment comparable ? Si je trouve une différence entre mes deux groupes, est-ce que ça ne pourrait pas simplement être dû au fait que dans le groupe « méditation », les participants sont bercés par la voix de

l'instructeur, alors que dans l'autre groupe, ce dernier parle dix fois moins ? Que faire alors ? Faire répéter à l'instructeur du groupe contrôle « détendez-vous, respirez calmement » toutes les trente secondes ? Mais si, grâce à ces instructions, les participants du groupe contrôle se focalisent sur leur respiration tout en évitant de penser à d'autres choses, n'est-ce pas finalement de la méditation ? Si je ne trouve pas de différence, n'est-ce pas simplement parce que les participants des deux groupes sont entrés dans un état méditatif pendant ces séances ?

Conclusion sur le problème du choix du groupe contrôle

Choisir un bon groupe contrôle est souvent difficile car il faut parvenir à identifier la caractéristique clé d'un traitement. C'est cette caractéristique clé qu'on va laisser présente dans le groupe expérimental et qu'on va supprimer dans le groupe contrôle. Dans le cas d'une étude sur un médicament, c'est relativement facile : la procédure est rigoureusement la même dans les deux groupes, sauf qu'un groupe reçoit le médicament et l'autre reçoit un placebo. Mais dans des cas complexes comme la psychothérapie et la méditation, quelle est vraiment la caractéristique clé ? Quel est le placebo ? Ce n'est pas toujours évident à déterminer.

Une chose est sûre néanmoins, c'est que dans ce genre de cas complexes, s'il s'avère que le groupe contrôle ne fait « rien », comme dans l'exemple de la potion magique, l'étude en question ne vaut pas tripette. C'est sans aucun doute le cas le plus problématique, méthodologiquement parlant. L'autre extrême (groupes trop similaires) a aussi son lot de problèmes (on s'arrache vite les cheveux !), mais il reflète en tout cas un raffinement scientifique bien plus grand. Plus on s'approche de ce second cas de figure, mieux on arrive à identifier clairement ce qui fait que « ça marche ».

Arrivé à ce point, on se rend compte que la dichotomie groupe contrôle *versus* groupe expérimental est rapidement un peu limitée. Une solution consiste à avoir plusieurs groupes. Dans le cas de mon étude sur la dépression, une solution serait d'avoir un groupe qui ne fait « rien », un autre qui vient parler 15 minutes à quelqu'un de qualifié à l'hôpital, un troisième qui fait la même chose que le second mais qui boit un simple verre d'eau, et un quatrième qui fait la même chose que le troisième mais qui boit la « potion magique ». La comparaison de tous ces groupes permettrait alors de distinguer les effets de contexte, du « facteur humain » et de la potion.

Problème 6. Inférence causale abusive

Passons maintenant à notre sixième problème, qui concerne l'inférence causale abusive. On entend par inférence causale abusive le fait de conclure qu'une variable X *cause* une variable Y, alors que la méthodologie ne permet pas de tirer ce genre de conclusion. Ce problème est assez répandu et peut se manifester de différentes façons.

On peut tout d'abord souligner le fait que l'inférence causale vraiment solide n'est possible que dans les plans expérimentaux, car comme nous allons le voir, ce sont les seuls qui permettent d'atteindre un degré de contrôle suffisant sur toutes les autres « troisièmes variables » qui pourraient expliquer le lien entre X et Y. Hélas, en pratique, les choses ne sont pas si claires et il existe de nombreuses zones grises entre l'inférence causale solide et l'inférence causale abusive.

Confusions grossières

Voyons d'abord les cas de confusion grossière, où l'inférence causale est clairement abusive. Un premier cas concerne l'amalgame entre le type d'analyse et type de design. Bien souvent, lorsqu'une étude expérimentale est réalisée, les données sont ensuite analysées avec un modèle du type ANOVA, qui est le plus naturel pour comparer des groupes. Une première erreur (vraiment grossière, mais que je préfère expliciter) est le fait de se dire que dès que l'on a affaire à une ANOVA, on se trouve dans le contexte d'un plan expérimental et que donc, on peut tirer des conclusions sur un éventuel lien de causalité entre la VD et les VI. Ce n'est évidemment pas le cas. L'inférence causale est avant tout une question de méthodologie et le type d'analyse de données n'a que peu

d'importance ; ce n'est quasiment jamais l'analyse de données – en tout cas jamais l'analyse de données seule – qui permet de tirer des conclusions fiables sur le lien de causalité entre deux variables.

Dans la même veine, on ne peut pas, sur la seule base d'un modèle de régression (dans le cadre duquel on dit souvent que la variable X *prédit* la variable Y) conclure qu'il y a un lien causal entre X et Y. En effet, les modèles de régression sont bien souvent utilisés pour analyser des données issues de designs corrélationnels transversaux, or ce type de design ne permet jamais d'inférence causale solide car, comme nous l'avons vu plus haut, on ne peut jamais exclure avec certitude qu'une troisième variable explique le lien observé entre X et Y. Ceci est également valable, évidemment, pour une simple corrélation ; la corrélation ne démontre *jamais* la causalité. Le lien entre deux variables n'est en fait qu'une condition nécessaire à la démonstration d'une relation causale, mais elle n'est pas, à elle seule, une condition suffisante. Cela peut paraître évident, mais en pratique, il arrive souvent que, par négligence ou abus de langage, on conclue que X cause Y après avoir simplement constaté que X et Y sont significativement corrélées.

Des cas plus subtils

Considérons maintenant des cas un peu plus subtils, où l'inférence causale demeure abusive, mais d'une façon moins grossière. Un premier cas concerne les plans quasi-expérimentaux et toutes les variables non manipulables. Une variable non manipulable (âge, sexe, revenus, classe sociale, lieu d'habitation, etc.) appartient en fait toujours au monde des designs corrélationnels, si l'on peut dire, avec toutes les limitations que cela implique. Il arrive toutefois que, lorsque ces variables sont « embarquées » dans un plan quasi-expérimental, avec un degré de contrôle et de sophistication méthodologique généralement plus grand que dans une simple étude corrélationnelle transversale, on oublie que l'on ne peut fondamentalement pas tirer de conclusions solides sur le fait que ces variables soient causalement associées à la VD.

Imaginons par exemple que l'on mène une étude sur les systèmes d'éducation alternatifs. On suppose par exemple que les enfants qui sont dans une école privée (où telle ou telle pédagogie alternative est utilisée) savent mieux lire et plus tôt que les enfants dans le système scolaire classique. On fait venir ces enfants dans un centre de recherche, on leur fait passer des batteries de tests de lecture dans des conditions très standardisées. On trouve que ces enfants lisent effectivement mieux que ceux dans le système classique. Même si cela est tentant, on ne peut pas vraiment conclure que c'est vraiment la pédagogie alternative qui permet de mieux apprendre à lire. Il est possible que les parents des enfants qui sont dans cette école soient d'une manière générale plus préoccupés de l'éducation de leurs enfants et qu'ils fassent également avec eux toutes sortes d'activités à la maison qui favorisent directement ou indirectement l'apprentissage de la lecture. Dans un tel cas, il serait donc assez clairement abusif de conclure à un lien de causalité entre méthode et apprentissage.

En pratique toutefois, il faut admettre que certaines études de ce genre parviennent à prendre en compte un nombre considérable de variables contrôles, qui permettent d'exclure l'effet de la plupart des « troisièmes variables » potentielles. Imaginons que l'on réalise la même étude que celle ci-dessus, mais en contrôlant en plus le niveau d'intelligence général des parents et des enfants, toutes sortes de variables liées aux enseignants (années d'expérience, qualité de la relation qu'ils ont avec les enfants, style pédagogique), toutes sortes de variables liées à l'école (qualité des infrastructures, nombre d'espaces verts, temps passé à l'extérieur par les enfants), toutes sortes de variables liées à l'éducation que les parents donnent à leur enfant (style parental, degré de stimulation proposé à la maison) et encore toutes sortes de variables socio-économiques (classe sociale, degré d'éducation des parents, etc.). Au bout d'un moment, après plusieurs études de ce genre, si l'effet de la méthode pédagogique alternative propre à l'école spécialisée persiste – si cet effet reste significatif après qu'on ait pris en compte toutes les variables contrôles – on admet souvent que c'est bien X qui cause Y, donc ici la méthode qui cause les meilleures capacités de lecture. Strictement parlant, c'est abusif, mais c'est

quand même beaucoup plus crédible que si l'on tire la même conclusion à partir de la corrélation entre l'appartenance à telle ou telle école et les notes obtenues en lecture.

Ce genre d'études sophistiquées est d'autant plus solide si, en plus de toutes les variables contrôles que l'on vient d'évoquer, l'étude en question est *longitudinale*. Dans une telle étude, on aurait alors, en plus de tous les raffinements méthodologiques déjà décrits, mesuré les capacités des enfants plusieurs fois ; on aurait par exemple évalué leur capacité de prélecture avant l'entrée à l'école, puis leur capacité de lecture en 1ère, 2ème et 3ème années d'école primaire. Une telle approche renforce beaucoup la solidité des conclusions et peut permettre, notamment, de constater qu'avant l'entrée à l'école, il n'y a pas de différence entre les enfants et que les écarts apparaissent et se creusent en fonction de l'avancée dans le cursus scolaire. Ce genre d'études longitudinales de grande qualité est ce que l'on peut faire de mieux pour clarifier la relation causale entre deux variables qui ne sont pas manipulables expérimentalement. En effet, dans le cas de notre exemple ici, il serait impossible de concevoir une étude expérimentale puisque, pour des raisons éthiques, on ne peut pas aléatoirement assigner les enfants à telle ou telle école.

Conclusion sur l'inférence causale

En conclusion, soulignons bien le fait que *trois* conditions sont nécessaires à une inférence causale solide – et aucune n'est suffisante seule ! Il s'agit des conditions suivantes :

- *Il y a un lien significatif entre les deux variables* ; ce point peut être démontré avec n'importe quel plan de recherche et testé par une simple corrélation.
- *Il y a antériorité de la cause* ; ceci implique inévitablement un design longitudinal ou expérimental ; une étude corrélationnelle ne peut jamais démontrer ce point, puisqu'il n'y a qu'un temps de mesure et aucune manipulation de la VI.
- *On a exclu toutes les autres causes potentielles* ; strictement parlant cela implique un design expérimental, où la seule chose qui varie est la VI et où tous les autres paramètres de la situation sont maintenus constants. On peut toutefois s'approcher de l'exclusion des autres causes potentielles en utilisant un très grand nombre de variables contrôles soigneusement sélectionnées.

C'est donc bien seulement les plans expérimentaux qui permettent de réunir ces trois conditions. Néanmoins, les études longitudinales très sophistiquées peuvent s'en approcher d'assez près ; et l'on est parfois obligé de s'en contenter, car toutes les variables ne sont pas manipulables expérimentalement.

Problème 7. Généralisation abusive

Ce septième et dernier problème est probablement le plus controversé d'entre tous ; il s'agit de la généralisation abusive. Cette notion de généralisation abusive est étroitement liée à la question de la représentativité de l'échantillon ; il y a généralisation abusive lorsqu'on généralise un résultat à une certaine population à partir d'un échantillon qui n'est pas représentatif de cette population. Comme pour l'inférence causale abusive, il y a toute une étendue de cas plus ou moins sévères, plus ou moins graves. Ce problème concerne indifféremment tous les types de plans de recherche.

Avant d'aller plus dans les détails de ce problème, commençons par une clarification importante, afin de ne pas confondre cette question de la généralisation avec celle de l'inférence statistique. Quand on trouve, par exemple, une différence significative entre deux groupes, on conclut que cette même différence existe dans la population. Bien sûr ce n'est pas nécessairement *exactement* la même différence ; peut-être que la différence observée dans l'échantillon est de 2.5 et que la « vraie » différence dans la population est de 2.4 ou 2.6. Cette légère imprécision est liée à l'erreur d'échantillonnage et est tout à fait normale – on peut même l'estimer avec précision, en particulier en calculant des intervalles de confiance. Tout cela concerne l'inférence statistique, qui ne se préoccupe pas du tout de la question de la représentativité de l'échantillon. Bien au contraire, la question de la

représentativité de l'échantillon est même en principe une condition *préalable* à toute inférence statistique.

D'une façon générale, la psychologie est une discipline qui, bien souvent, est notoirement peu préoccupée par les questions d'échantillonnage et de représentativité (au contraire, par exemple de la sociologie). Si, en tant que psychologue, on connaît les techniques d'échantillonnage (aléatoire simple, par grappes, stratifié, etc.), la plupart des psychologues ne les utilisent en fait jamais dans leur recherche. Pourquoi ?

Une brève histoire de la psychologie

Historiquement, la psychologie générale et la psychologie expérimentale (les premières variantes de la psychologie dite scientifique, basée sur la mesure et les tests statistiques) se sont toujours intéressées à des mécanismes généraux, supposés similaires chez chaque être humain. Ces disciplines s'intéressaient en particulier à des questions liées à la mémoire, l'attention et la perception. Elles reposent sur le principe que, mises à part quelques variations aléatoires et quelques différences d'aptitude d'un individu à l'autre, les capacités cognitives de tout être humain sont organisées fondamentalement de la même façon.

Les chercheurs en psychologie cognitive expérimentale menaient leurs études essentiellement à l'aide de plans expérimentaux, souvent d'une exemplaire rigueur, et n'avaient le plus souvent que faire de variables liées au sexe, à l'âge, à la culture ou à l'ethnie d'origine – sauf, évidemment, dans des cas extrêmes (par exemple un nourrisson comparé à une personne âgée). On concevait un plan de recherche expérimental, on faisait varier tels ou tels paramètres dans telles et telles conditions expérimentales, on assignait aléatoirement les participants à l'une ou l'autre de ces conditions, et voilà ! On trouvait telle différence ici, tel effet de telle variable là, et on concluait (implicitement) que c'était pour tout le monde pareil. (Je parle ici au passé, mais cette manière de faire se porte aujourd'hui encore très bien !)

Les choses ont commencé à changer avec la psychologie différentielle. Au lieu de s'intéresser de façon quasi exclusive aux « effets moyens », supposés les mêmes pour tout être humain, la psychologie différentielle s'est focalisée plutôt sur la variabilité liée à des facteurs individuels, que les expérimentalistes considéraient simplement comme de l'erreur. La psychologie différentielle s'intéresse fondamentalement aux différences interindividuelles et donc aux effets d'âge, de sexe, de culture, etc. C'est aussi la psychologie différentielle qui est à l'origine des conceptions des tests d'intelligence (et, par extension, de toutes sortes de tests). Or, pour que ces tests soient utilisables, il est impératif qu'ils soient normés, et ces normes ne peuvent être construites qu'à partir d'un échantillon représentatif. Dans ces cas-là, qui sont beaucoup plus l'exception que la règle, la représentativité de l'échantillon était soigneusement considérée et les techniques classiques d'échantillonnage utilisées.

Pendant longtemps, les choses ont peu changé : pour les tests, on fait des efforts concernant l'échantillonnage ; dans les autres cas, on ignore le problème ou on ajoute une simple phrase à la conclusion de chaque article, qui dit que « l'échantillon utilisé dans cette étude n'est peut-être pas parfaitement représentatif de la population » ou que « les résultats devraient être répliqués avec un autre échantillon ».

La psychologie interculturelle et les « WEIRD »

Depuis peu toutefois, une petite révolution est en train de s'opérer. (Il est un peu abusif de dire « depuis peu » car, en fait, cela ne fait pas loin d'une cinquantaine d'années ; mais toutes proportions gardées, c'était quand même assez long à venir…) On assiste en effet depuis les années 1970 à l'émergence – et aussi à la visibilité grandissante depuis une ou deux décennies – de ce qu'on appelle la psychologie interculturelle. Cette sous-discipline de la psychologie questionne profondément la validité externe des études expérimentales. Constatant que la plupart des études en psychologie, pendant pas loin d'un siècle, ont porté sur des échantillons constitués d'Européens ou de Nord-Américains, la psychologie interculturelle s'interroge sur la généralisabilité de ces résultats à d'autres ethnies et à d'autres cultures. Ces doutes concernent aussi bien la cognition (classiquement étudiée par la psychologie

générale expérimentale) que la personnalité (plutôt étudiée par des approches issues de la psychologie différentielle) ou encore la psychopathologie.

Certains auteurs parlent même de *psychologie WEIRD* (qui veut dire « bizarre, étrange, curieux » en anglais). Cet acronyme WEIRD dénonce le fait que la plupart des études en psychologie sont menées sur des participants issus de sociétés *de l'Ouest* (**W**estern), *avec un haut niveau d'éducation* (**E**ducated), *industrialisées* (**I**ndustrialized), *riches* (**R**ich) et *démocratiques* (**D**emocratic). Les personnes qui adhèrent à cette approche critique – à juste titre sans doute, en tout cas jusqu'à un certain point – considèrent que les membres des sociétés WEIRD ne sont pas représentatifs du genre humain, voire même que ce sont les moins représentatifs. Pire encore, la majorité des études en psychologie sont en effet menées sur des étudiants de psychologie en début de cursus (donc un sous-ensemble des WEIRD !). Quoique parfois excessivement alarmiste, cette « mode du WEIRD » a au moins l'avantage de rendre ces questions explicites et d'ouvrir le débat sur ce qui est peut-être le plus grand tabou ou le plus grand non-dit de la recherche scientifique en psychologie.

Conclusion sur les problèmes de généralisation

Comme toujours, certains vont loin, dans un sens ou dans l'autre. Néanmoins, ne dramatisons rien. Nonobstant ce qu'en disent les plus fervents partisans du WEIRD, toute la recherche en psychologie du siècle dernier n'est pas bonne à jeter à la poubelle. Il s'agit plutôt d'un encouragement à tester si des résultats que l'on pensait être universels le sont effectivement. D'une manière générale, la psychologie interculturelle permet donc d'aborder frontalement ces questions, plutôt que de les éluder systématiquement.

Maintenant, d'un point de vue pragmatique, voici quelques précautions raisonnables et mesurées (me semble-t-il) sur ces questions de généralisation abusive. Première distinction : parle-t-on d'une *moyenne* sur une variable ou d'un *lien entre deux variables* ? À l'évidence, si je fais passer un test cognitif ou un questionnaire de personnalité à un échantillon d'étudiants en psychologie, il serait ridicule que je me serve de ces données pour définir les normes du test en question, constatant par exemple que la moyenne est à 10 et l'écart-type de 3, puis affirmant que ceci est valable pour tout le monde. Très certainement, les moyennes des autres populations seront très différentes.

À l'inverse, si je mène une étude sur les effets du stress chronique sur le bien-être (un lien entre deux variables, donc), il semble beaucoup plus probable que les résultats, dans les grandes lignes, s'appliqueront à n'importe quel sous-groupe humain. Quoiqu'il existe des différences importantes de réactivité au stress d'un individu à l'autre, et qu'un évènement stressant dans une culture n'est pas forcément stressant dans une autre, on peut sans doute raisonnablement admettre que l'effet du stress chronique sera globalement le même pour tout le monde (en particulier s'il est mesuré par des marqueurs biochimiques comme la surreprésentation de certaines hormones dans le sang) et sera inévitablement associé à de la tension psychologique et à une diminution du bien-être.

Ceci nous amène à une autre distinction qui est celle de *la thématique* de la recherche dont on cherche à généraliser les résultats. Si l'on s'intéresse aux facteurs qui améliorent la mémoire, il est fort possible qu'ils soient (relativement) universels. En revanche, si l'on s'intéresse au lien entre violence à la télévision et morts par arme à feu, la généralisation risque d'être compliquée, toutes les sociétés n'ayant pas le même rapport à la télévision (certaines ne l'ont même pas du tout) et toutes les sociétés n'ayant pas le même rapport à la violence et aux armes à feu. Il s'agit là d'exemples contrastés, mais il n'en demeure pas moins que cette question de la thématique et de la vraisemblance de son universalité est, en attendant les études qui la testent directement, le seul moyen d'éviter de tomber dans deux extrêmes aussi absurdes l'un que l'autre, à savoir « rien n'est généralisable car tout le monde est différent » et « c'est pour tout le monde pareil car nous sommes tous humains ».

Synthèse et recommandations

- Ne jamais assimiler une opérationnalisation simpliste à une variable théorique complexe. Ne pas oublier l'opérationnalisation qui est inévitablement derrière toute variable théorique ; chercher à savoir quelle est cette opérationnalisation et clarifier si elle est crédible.
- Utiliser des mesures de qualité – valides, fidèles et suffisamment sensibles. Rester critique vis-à-vis des mesures utilisées dans une recherche.
- Contrôler correctement les variables parasites, notamment grâce à un plan expérimental avec, si possible, une procédure en double-aveugle et un placebo. Si ce n'est pas possible, ne jamais oublier qu'une troisième variable peut expliquer la relation observée entre deux autres variables.
- Eviter absolument l'erreur systématique en utilisant
    - des outils de mesure qui n'induisent pas de biais de réponse,
    - l'assignation aléatoire dans les plans expérimentaux à groupes indépendants,
    - le contre-balancement dans les plans expérimentaux à mesures répétées,
    - la procédure en double-aveugle + un placebo si c'est possible.
- Minimiser l'erreur aléatoire, en utilisant des outils de mesure précis et des conditions de passation relativement uniformes.
- Choisir soigneusement le groupe contrôle; en prendre plusieurs si nécessaire; éviter le groupe contrôle « qui ne fait rien ».
- Ne jamais faire d'inférence causale à partir d'un plan corrélationnel transversal.
- L'inférence causale ne peut être faite qu'à partir d'un plan expérimental solide (ou, à la limite, à partir d'une étude longitudinale de grande qualité, avec de nombreuses variables contrôles).
- Toujours considérer la représentativité de l'échantillon, évaluer l'importance de cette dernière en fonction du type de conclusion tirée et de la thématique.

Conclusion

Nous nous sommes focalisés ici sur des aspects méthodologiques; il y a bien sûr aussi 1'000 façons de faire des erreurs lors de l'analyse de données…

Et on peut même encore commettre des erreurs *après* l'analyse de données, en interprétant incorrectement des résultats statistiques. Les exemples les plus criants sont l'amalgame entre la taille d'effet et la significativité (« La *p*-valeur est petite, donc l'effet est important ») ou encore le fait d'interpréter un résultat qui n'est pas significatif (« Ce n'est pas significatif, mais il y a clairement une tendance »)…

La liste de problèmes présentée ici est donc loin d'être exhaustive ; restez vigilants!