

Une critique du conformisme méthodologique en psychologie générale

On peut aussi chercher des relations générales restrictives

Stéphane Vautier

Université de Toulouse¹

¹ Université de Toulouse-Le Mirail. Pavillon de la Recherche, Octogone. 5 allées Antonio Machado, 31058 Toulouse cedex 9. Courriel : vautier@univ-tlse2.fr

Résumé

Une hypothèse comme « certaines conditions expérimentales influencent l'estime de soi » débouche typiquement sur le test statistique de l'égalité de scores moyens. Ce type de démarche repose sur deux illusions : (1) on sait mesurer une grandeur psychologique via le scorage des réponses à un questionnaire, (2) on produit des connaissances générales en étudiant des personnes typiques. Une méthodologie radicalement différente est disponible, fondée sur la recherche de relations générales et restrictives. Elle ne nécessite ni procédé numérolgique de quantification, ni test statistique de l'hypothèse nulle. Il serait donc faux de considérer que la démarche habituelle est la seule possible pour un chercheur qui prétend contribuer à l'avancement des connaissances en psychologie générale.

Mots-clés : épistémologie, mesurage, méthodologie, relation restrictive et générale

Abstract

A typical research project consists in starting from a general hypothesis like “self-esteem depends on some experimental conditions” and ending with the null hypothesis significant test of average scores. This kind of research conceals two illusions: (1) a test score measures a psychological, quantitative, attribute, (2) the study of a typical person produces general knowledge. A radically different methodology is available, which is based on the search of general and restrictive relations. It requires neither index-numerology nor null hypothesis significance testing. Consequently, it would be wrong to claim that the usual methodology is the sole one for a researcher who wants to advance knowledge in general

psychology.

Key-words: epistemology, measurement, methodology, restrictive and general relation

1. Introduction

Pour autant qu'une psychologie scientifique soit concevable, c'est, par définition, une psychologie générale, c'est-à-dire une discipline qui cherche des principes théoriques valables pour tout système psychologique (par exemple, tout être humain). C'est aussi une psychologie expérimentale, c'est-à-dire une psychologie qui se sert des faits d'observation pour tester ses principes théoriques. Une telle définition ne spécifie ni les principes ni les faits de la psychologie. Cette tâche de spécification ne se trouve pas à la portée d'un chercheur débutant, c'est pourquoi un enseignement de méthodologie est nécessaire à sa formation.

Mais cet enseignement peut s'avérer contre-productif si l'apprentissage d'un certain nombre de savoir-faire pour le recueil et l'analyse des données est effectué sans une formation à l'analyse épistémologique. L'analyse épistémologique consiste notamment à identifier les conditions de possibilité de la connaissance scientifique (e.g., Soler, 2000). Or, nous avons l'impression que la formation méthodologique des étudiants de psychologie souffre d'un clivage entre l'épistémologie de la discipline et une forme de statisticisme (Lamiell, 2013) auquel elle semble confier son salut scientifique. Les savoir-faire en analyse des données (pour résumer, l'ANOVA, la régression – linéaire, parfois logistique –, l'analyse factorielle, la classification – non supervisée – et quelques rudiments de

modélisation par variables latentes) sont souvent considérés par les étudiants – et les enseignants qui dirigent leurs travaux – comme une fin et non un moyen.

Lorsqu’il s’attelle à son travail de recherche, le chercheur débutant se trouve face au problème suivant : comment, à partir de prémisses théoriques fondées sur des intuitions (e.g., l’estime de soi dépend des comparaisons sociales, Redersdorff & Martinot, 2003), déboucher sur des résultats statistiques conformes à ceux qui sont habituellement rapportés dans la littérature spécialisée ? La réponse typique comprend au moins trois étapes (Vautier, 2011b) : (1) opérationnalisation des concepts théoriques, (2) recueil des données, (3) mise en évidence d’un résultat statistique conforme (par exemple, une différence significative, un pourcentage de variance résiduelle conforme à ce qui a été publié, une structure factorielle unidimensionnelle, etc.).

Ramener la méthodologie de recherche en psychologie à ce type de démarche optimise la capacité du chercheur à boucler un rapport de recherche dans le temps qui lui est imparti en même temps qu’à démontrer sa capacité à faire un exercice de style. Mais, lorsqu’une telle pratique s’institutionnalise, elle devient un conformisme socio-technique (Danziger, 1985; Michell, 2003). L’impératif de conformité, s’il n’est pas soigneusement justifié au sein de la discipline, permet que le travail de recherche soit mené en dépit du fait que les conditions de possibilité d’une réponse scientifique aux questions initiales ne soient pas établies.

Mais pour qu’un tel diagnostic de défaillance méthodologique puisse être posé, encore faut-il que les questions initiales qui sont supposées motiver la

méthodologie du projet de recherche soient explicites. L'exemple de l'idée d'une relation de dépendance entre l'estime de soi et la comparaison sociale est utile pour expliciter ces questions. La locution « estime de soi » signifie une grandeur. La question de la mesurabilité de cette grandeur se pose alors tout de suite. Mais cette question est immédiatement remplacée par la question pratique (et non pas théorique) : comment opérationnaliser l'estime de soi ? Et cette question conduit alors, étant donné l'usage en cours dans la discipline, à la recherche d'un questionnaire approprié pour « mesurer » ladite grandeur psychologique.

La première partie de cet article montre que la solution psychotechnique typique, qui consiste à assigner un score à un vecteur de réponses qualitatives, ne répond pas à la question de la mesurabilité de la grandeur invoquée. Pour évaluer la validité d'un tel verdict d'inefficacité, il faut disposer de critères d'évaluation. Le problème logique qui se pose consiste à spécifier puis à tester un principe de liaison entre le domaine théorique de la grandeur et le domaine empirique des réponses observables. La conséquence immédiate de cette analyse sera que si on ne dispose pas d'un tel principe, ou si ce principe est falsifié par les faits d'observation, alors on ne sait pas mesurer l'estime de soi et il est prématuré d'essayer de tester que l'idée que l'estime de soi, comme grandeur qu'on ne sait pas mesurer, puisse dépendre de quoi que ce soit. L'analyse épistémologique rappelle que les objectifs de recherche sont tributaires de l'état des connaissances, en ce sens qu'il existe une hiérarchie logique des problèmes. Si le problème B suppose que le problème A soit résolu, et si le problème A n'est pas résolu, alors la contribution espérée doit porter sur le problème A.

La réponse typique en trois étapes que nous évoquions plus haut conduit le chercheur débutant à se satisfaire d'une pseudo-solution pour construire le domaine de valeurs de sa variable dépendante, ce qui lui permet de se centrer sur la conception d'un plan expérimental (ou quasi-expérimental) requis pour la mise en œuvre de l'analyse statistique des données. Ici, l'impératif méthodologique, que Lamiell (2013) dénonce par le terme de *statisticisme*, consiste à tester l'hypothèse nulle de l'égalité des scores moyens d'estime de soi conditionnellement à telle ou telle situation expérimentale via une expérience d'échantillonnage néo-galtonienne (Danziger, 1987; 1990), c'est-à-dire une expérience qui agrège des individus dans des échantillons. La seconde partie de cet article montre que l'approche expérimentale telle qu'elle est formatée par l'impératif néo-galtonien implique la perte de la généralité du principe même qu'il s'agit de tester. Si l'estime de soi dépend de la comparaison sociale par exemple, ce principe de dépendance est par définition général, c'est-à-dire qu'il est valable pour toute personne susceptible de s'impliquer dans une comparaison sociale. Or le support empirique de l'estime de soi dans une comparaison de scores moyens néo-galtoniens n'est pas une personne quelconque, qui verrait l'intensité de l'estime de soi changer en fonction des conditions psychosociales dans lesquelles elle se trouve, mais une personne moyenne, dont l'ontologie, i.e., l'existence même, résulte essentiellement d'une expérience de pensée (Vautier, 2011b). Si bien que même si l'hypothèse « alternative » est acceptée comme conséquence logique du rejet de l'hypothèse « nulle », la proposition qui en résulte ne décrit personne en particulier.

La troisième partie de cet article propose un positionnement général de la problématique de recherche qui, s'il constitue une tentative d'émancipation vis-à-vis du conformisme socio-technique en vigueur, ne consiste pas en une apologie de l'anarchie méthodologique, puisqu'il identifie les contraintes méthodologiques qui découlent de notre manque de connaissances générales en matière de mesurage psychologique d'une part, et de détermination des phénomènes psychologiques observables d'autre part. La méthodologie de la recherche en psychologie doit (aussi) servir un objectif commun de recherche de relations, que nous appellerons restrictives et générales, entre des phénomènes psychologiques objectivables. Ainsi, la tâche du chercheur débutant se trouve définie à l'avance par la structure logique des connaissances existantes, et sa contribution singulière réside essentiellement dans sa capacité à développer une approche descriptive qui vise la mise en évidence de phénomènes relationnels spécifiques.

2. La correspondance entre le monde théorique de la grandeur et le monde empirique des observables

Dans cette partie, nous prendrons l'exemple de la grandeur « estime de soi » telle qu'elle est typiquement opérationnalisée par Redersdorff et Martinot (2003) à l'aide d'un questionnaire auto-évaluatif. Le questionnaire comprend sept items dont les réponses sont cotées sur une échelle ordinaire composée de sept degrés (de « pas du tout » à « tout à fait »). Il s'agira d'explicitier (1) les implications ordinales du concept de grandeur, (2) la structure logique de la variable dépendante et (3) le problème de la mise en correspondance de la

grandeur et des observables. Ensuite, nous montrerons que (4) le processus d'obtention du score d'estime de soi non seulement ne répond pas à la question de savoir comment il est possible que ce score mesure l'estime de soi en tant que grandeur théorique, mais qu'il résulte d'une confusion entre la signification ordinale des réponses à un item quelconque du questionnaire et la signification additive qu'il faut conférer à ces réponses pour obtenir un score « composite ». Enfin, nous montrerons que (5) le recours à la théorie classique des tests (e.g., Lord & Novick, 1968) pour interpréter ce score comme mesure d'une grandeur latente implique non seulement un décret de mesurabilité de la grandeur par les réponses aux items, mais encore l'acceptation de l'invalidité de toute interprétation fondée sur la différence de deux scores étant donnée l'absence de restriction qui affecte « l'erreur de mesure ».

2.1. La structure ordinale connectée d'une grandeur

Par définition, une grandeur est un attribut qui varie de trois manières possibles : soit il augmente, soit il diminue, soit il ne varie pas. Dans le langage de la théorie de l'ordre (e.g., Michell, 1990, appendice 1; Roberts, 1985, chapitre 1), la grandeur, en tant qu'ensemble de valeurs possibles, est un ordre simple strict (mais la réciproque est fautive : un ordre simple strict n'est pas nécessairement une grandeur). Un ordre simple strict est un ensemble doté d'une relation binaire antisymétrique, transitive et connectée. Cette dernière propriété, la connexion, nous intéresse de manière critique : quelles que soient deux valeurs distinctes a et b de la grandeur, soit a est plus petit que b soit b est plus petit que a . C'est cette

propriété qui définit la grandeur comme dimension unique, parce que toutes les valeurs qu'elle peut prendre s'ordonnent sur une seule ligne.

Remarquons que les scores d'estime de soi, quelle que soit la façon dont on les obtient, forment aussi un ordre simple strict, d'où leur attrait pour mimer des mesures de la grandeur. Nous utilisons le verbe « mimer » plutôt qu'un autre pour signifier que l'analogie ordinale entre les scores et la grandeur est nécessaire mais pas suffisante pour établir et tester le principe d'une correspondance entre la grandeur et le domaine des scores possibles. Un principe de liaison possible est celui d'une fonction par palier, qui à tout point de la grandeur associe un score, de telle sorte que les scores correspondent à des segments ordonnés. L'étendue de ces segments est cependant inconnue. Si elle était connue, le mesurage permis par les scores serait non pas ordinal mais quantitatif. Le problème qui se pose est qu'on ne sait pas tester la validité d'un tel principe de liaison entre la grandeur et les scores. Autrement dit, ce principe de liaison constitue une convention linguistique qui n'a aucune portée empirique. Bien entendu, les conventions sont nécessaires à l'activité scientifique (Granger, 1987; Lakatos, 1978). Ici cependant, ce qui est nécessaire n'est pas une convention mais une théorie testable spécifiant comment la grandeur est reliée à des événements empiriques observables avec le questionnaire.

2.2. La structure ordinale non connectée de l'ensemble des réponses au questionnaire

Il faut d'emblée distinguer deux sortes d'ensembles d'événements empiriques. Le premier est l'ensemble des réponses logiquement possibles au

questionnaire d'estime de soi. Une réponse au questionnaire signifie qu'on considère la réponse au premier item, la réponse au deuxième item, etc. jusqu'à la réponse au septième item. Ainsi, une réponse au questionnaire n'est pas la donnée d'une valeur, mais de sept valeurs ordonnées, c'est-à-dire d'un 7-uplet (ou d'un vecteur de sept réponses).

On peut facilement calculer le cardinal de cet ensemble. Comme l'échelle de réponse à tout item contient sept degrés (de « pas du tout » à « tout à fait »), le nombre de 7-uplets possibles est fini. Pour le connaître, il suffit d'envisager comment on peut en déterminer la liste complète. Le premier item autorise sept réponses. À chacune de ces réponses, le second item ajoute sept réponses possibles, ce qui donne un total de 7^2 couples de réponses. À chacun de ces couples de réponses, le troisième ajoute sept réponses possibles, soit un total de 7^3 triplets de réponses, et ainsi de suite jusqu'aux 7^7 7-uplets possibles.

Le second type d'ensemble de réponses possibles est l'ensemble des réponses qui en fait sont observables, étant données les contraintes psychologiques qui, par hypothèse, déterminent leur production. Fondamentalement, pour qu'un 7-uplet soit utilisable comme une indication ordinale de l'intensité de l'estime de soi, il faut supposer que sa production effective dépend uniquement de la valeur de l'estime de soi. Si on admet que la réponse au questionnaire dépend d'autre chose de manière non négligeable, alors on reconnaît d'emblée que la réponse ne peut pas servir de mesure ordinale de l'estime de soi, puisqu'une variation observée peut par hypothèse dépendre, de manière non négligeable, d'autre chose que de la variation théorique d'estime de

soi. De ce point de vue, le concept d'erreur de mesure, tel qu'il est promulgué en psychométrie, invalide d'entrée de jeu que les observations effectuées puissent être interprétées comme des mesures d'une grandeur unique. Nous reviendrons sur l'erreur de mesure en psychométrie dans le paragraphe 2.5. L'ensemble des réponses non pas logiquement mais psychologiquement possibles est un sous-ensemble de l'ensemble des réponses logiquement possibles (dans le cas des scores, cette distinction n'est pas nécessaire à l'argument).

2.3. Le problème de la correspondance

La question qui se pose ensuite est de savoir s'il est possible d'attribuer à chaque 7-uplet de l'ensemble des réponses psychologiquement possibles un intervalle de valeurs théoriques (ou un segment), conformément au principe de liaison d'une fonction par palier. Une analyse approfondie du problème dépasse le cadre de cet article et nous renvoyons le lecteur à notre cours de mesurage². Sauf dans le cas improbable où les réponses observées obéissent à une échelle de Guttman (1944), il n'existe pas de principe pour ordonner totalement (ou connecter) les 7-uplets possibles.

On peut saisir le problème en réfléchissant à ce qui se passe lorsqu'on compare deux 7-uplets différents. Le principe de base de l'échelle de réponse à tout item est qu'il existe une fonction par palier reliant la grandeur aux réponses. Ainsi, la septième réponse indique un niveau d'estime plus grand que le niveau qu'indique la sixième réponse, etc. jusqu'à la première réponse. Si, en considérant l'évolution des réponses item par item, on observe qu'un item indique une

² Il suffit de demander le document par courriel à l'auteur du présent article.

augmentation et qu'un autre item indique une diminution, alors il n'est pas possible que les deux items mesurent (ordinalement) la même grandeur, puisque si c'était le cas, comme la grandeur varie soit en augmentant, soit en diminuant, mais pas les deux à la fois, on ne pourrait pas observer que les deux types de variation se produisent en même temps. Par conséquent, dès qu'on reconnaît que les réponses possibles au questionnaire n'excluent pas la possibilité de changements dont les sens sont opposés, on falsifie l'idée qu'on saurait mesurer ordinalement la grandeur avec les 7-uplets psychologiquement possibles. La grandeur peut toujours exister dans le réseau nomologique du chercheur (Cronbach & Meehl, 1955), mais son accès empirique n'est pas établi.

La coutume dans la discipline n'est pas de conduire ce type d'étude préalable, mais de valider des tests et des questionnaires psychologiques sur des bases théoriques spécifiques, dites psychométriques (e.g., DeVellis, 2003; Laveault & Grégoire, 2002; Penta, Arnould, & Decruynaere, 2005), dont Michell (1990; 1997; 1999; 2000; 2006; 2008) a sévèrement, et avec constance, critiqué la valeur scientifique (voir aussi Borsboom, Cramer, Kievit, Scholten, & Franic, 2009; Vautier, Veldhuis, Lacot, & Matton, 2012).

2.4. Le score composite d'estime de soi est une pseudo-mesure de la grandeur

Le score d'estime de soi est calculé de la manière suivante. On numérote les degrés de réponse sur l'échelle de réponse de tout item de 1 à 7. Ainsi, tout 7-uplet de réponses est un vecteur de la forme (1, 2, 3, 1, 1, 4, 2) par exemple. Le score composite (i.e., le score d'estime de soi) s'obtient en sommant les nombres qui se trouvent dans le 7-uplet. Ainsi par exemple, on attribue le score 14 au

vecteur (1, 2, 3, 1, 1, 4, 2). Dans leur article, Redersdorff et Martinot (2003) divisent le score par 7, mais cette transformation n'implique rien de substantiel par rapport à la signification des scores obtenus par la somme des composantes.

Une telle pratique cristallise une confusion logique rédhibitoire du point de vue de la cohérence de l'utilisation des nombres. Avant d'analyser cette confusion, nous voudrions appuyer notre propos sur un argument de la philosophie des sciences. Au début du XXe siècle, Windelband a exprimé de manière convaincante la nécessité, dans le travail scientifique, de la cohérence logique en même temps que la compatibilité de cette discipline de cohérence avec la créativité théorique du chercheur.

C'est précisément dans la mathématique que la relation de la connaissance à ses objets, formés par cette connaissance même, possède une transparence extrêmement instructive. Car, si les configurations numériques et les figures spatiales particulières que le chercheur constitue en objets dépendent de son arbitraire, la pensée connaissante, dès que l'objet est construit, est pourtant complètement liée à lui et soumise à la légalité immanente de l'objet. La production du problème et sa solution ne peuvent rien viser d'autre que le développement des rapports de grandeur créés par la construction synthétique même. La pensée mathématique, aussi libre qu'elle soit dans la production de ses objets, fait précisément en cela l'expérience de la *contrainte de l'objectivité* [*Gegenständlichkeit*], qui se trouve contenue dans les configurations qu'elle construit et qui oppose sa puissance souveraine à

toute affirmation arbitraire. Nous voulons en outre qualifier cette relation, qui apparaît nécessairement du point de la psychologie subjective comme un miracle, de *logique de l'objet*.

Il en va toutefois de même pour les *sciences empiriques*. Si les habitudes de la pensée préscientifique, relevant du réalisme naïf, le dissimulent, la réflexion logique doit du coup d'autant plus rigoureusement mettre en évidence le fait que nulle connaissance ne peut se rapporter à des objets qu'elle n'a pas elle-même conceptuellement déterminés (Windelband, 2002, p. 222).

Le chercheur est libre de supposer une grandeur qu'il appelle « estime de soi », de la même manière qu'il est libre d'imaginer une technique d'observation fondée sur un questionnaire d'auto-évaluation. Mais il est tenu de respecter la logique de ces concepts. Or il se trouve que la logique d'une grandeur et la logique des phénomènes qu'on peut décrire avec un questionnaire d'estime de soi sont différentes. Dans le cas de la grandeur, on a affaire à un ordre simple strict (i.e., la relation d'ordre est connectée) tandis que dans le cas des réponses psychologiquement possibles, on a affaire à un ordre non connecté. C'est cette différence que le scorage psychotechnique escamote, en dépit de vaines protestations (Essex & Smythe, 1999; Johnson, 1935; 1936; 1943), en transmettant aux chercheurs du futur le legs d'une hypothétique justification à produire quant à la légitimité de l'opération.

En effet, pour que l'addition mobilisée par le scorage des 7-uplets observés ait un sens théorique (et non pas seulement métaphorique), il faut que les

étiquettes numériques entrant dans la composition des 7-uplets réfèrent à des quantités. Or ce n'est pas le cas, puisqu'elles désignent des rangs sur l'échelle de réponse de tout item. Les composantes des 7-uplets indiquent la position de la réponse dans une série ordonnée, et, par définition, un ordre relatif n'est pas une quantité. On peut buter sur ce défaut de sens en remplaçant l'étiquetage numérique par un étiquetage alphabétique. Ainsi, le 7-uplet (1, 2, 3, 1, 1, 4, 2) peut aussi bien être symbolisé par le 7-uplet (a, b, c, a, a, d, b) et on voit bien qu'il n'y a pas de sens à sommer ces lettres. Le même exercice conduit à la même conclusion si on utilise la notation (1^{ère}, 2^{ème}, 3^{ème}, 1^{ère}, 1^{ère}, 4^{ème}, 2^{ème}).

Par conséquent, les scores ne résultent pas d'un processus de mesurage bien défini et corroboré, mais d'un jeu de langage qui permet, par un subtil glissement de sens, d'assigner des nombres à des événements dont la structure ordinale non connectée n'est pas compatible avec la logique ordinale de la grandeur théorique visée.

2.5. L'interprétation du score par la théorie classique des tests

La théorie classique des tests (e.g., Lord & Novick, 1968) ignore le problème de la correspondance entre la grandeur et le score. Elle décompose le score comme la somme d'un score vrai et d'une erreur de mesure. Le score vrai de Paul qui a répondu hier à 14 heures au questionnaire est la moyenne de tous les scores que l'on obtiendrait si on était capable de remonter dans le temps, de répliquer l'expérience (faire remplir le questionnaire à Paul dans les mêmes conditions) indéfiniment, et, une fois qu'on aurait fini ces répliques, de calculer la moyenne des scores observés (Vautier et al., 2012). Étant donné le score

observé, l'erreur de mesure est la différence entre ce score et le score vrai. Par conséquent, il est impossible de connaître l'erreur de mesure puisqu'il est impossible de connaître le score vrai.

Arrivé à ce point de notre analyse, il est possible que le lecteur pense au fait que la théorie classique des tests permet d'approcher le problème de l'erreur de mesure grâce à ce qu'on appelle la fidélité d'un test. Avant de discuter ce point pour rejeter cette solution comme inappropriée, nous le prions d'avoir la patience de tirer deux conséquences logiques de l'axiomatique ainsi introduite.

Tout d'abord, le score vrai comme concept mathématique – le score espéré – est parfaitement légitime. Mais il est psychologiquement illégitime parce que rien ne permet d'identifier l'estime de soi en tant que grandeur théorique au domaine des scores vrais associés à l'échelle des scores composites : le score d'estime de soi est une convention numérique fondée sur la confusion logique entre la signification ordinale des réponses aux items, dont la sémantique a bien sûr trait à celle de l'estime de soi, et la signification proprement additive qui leur est ensuite conférée pour le calcul. Cette illégitimité psycho-métrique, au sens strict, tient à ce que le score ne résulte pas de la mise en relation préalable du score vrai avec l'estime de soi. On ne sait si une augmentation de l'estime de soi implique une augmentation du score vrai, et réciproquement, on ignore si l'augmentation du score vrai implique l'augmentation de l'estime de soi.

Si on identifie le score vrai, comme grandeur artefactuelle, à l'estime de soi, comme grandeur théorique (et, espère-t-on, réelle), alors on crée un problème catastrophique pour la théorie de l'estime de soi, à savoir qu'il existe autant de

types d'estime de soi que de questionnaires d'estime de soi. Cette déclinaison infinie génère la nécessité d'établir que toutes ces opérationnalisations (qui, au sens strict, n'en sont pas parce qu'elles mobilisent des opérations imaginaires) peuvent être mises en correspondance. Mais comme on ne connaît pas le score vrai que Paul aura obtenu en répondant à un questionnaire quelconque, on ne connaît pas les scores que Paul aura obtenus en répondant à quelques questionnaires d'estime de soi. C'est pourtant une condition nécessaire pour tester que le vecteur des scores vrai de Paul se comporte de manière homogène lorsque les conditions dans lesquelles Paul répond à ces questionnaires changent.

Une deuxième conséquence de l'axiomatique du score vrai concerne l'invalidité d'un raisonnement comparatif. Par exemple, Paul dans telle condition a un score d'estime de soi plus faible que dans telle autre condition expérimentale. Si on identifie score vrai et valeur de l'estime de soi, en admettant sans preuve empirique qu'il existe une fonction croissante entre ces deux domaines, peut-on en déduire que son estime de soi a augmenté de la première à la seconde condition ? Non parce que l'erreur de mesure qui est attachée à chaque score observé est inconnue. Étant donnés les scores observés, il est possible que le score vrai n'ait pas changé, comme il est possible qu'il ait diminué au lieu d'augmenter, comme il est aussi possible qu'il ait augmenté.

La théorie classique des tests, que l'on pourrait aussi appeler théorie de la fidélité, est efficace à partir du moment où on s'intéresse non plus à une personne quelconque, mais à une population de personnes (e.g., Gaudron & Vautier, 2007; Vautier, 2011a; Vautier, Gaudron, & Jmel, 2004; Vautier & Jmel, 2003; Vautier,

Steyer, & Boomsma, 2008). Cette population, dont nous préciserons la structure logique dans la prochaine partie de l'article, est le domaine de définition d'une variable. Les valeurs de cette variable sont des scores attribués aux unités statistiques de la population. La fidélité d'une variable (et non pas d'un score ni d'un test comme nous l'avons énoncé plus haut faute de pouvoir tout discuter en même temps, voir Thompson & Vachaa-Haase, 2000) est alors définie comme le rapport de la variance des scores vrais sur la variance des scores observables. Dans le cas limite où ce rapport vaut 1, la variance d'erreur est nulle, ce qui implique que toute erreur de mesure est nulle puisque, par définition du score vrai, la moyenne de l'erreur de mesure est nulle. Dès que la fidélité d'une variable est inférieure à 1, il est logiquement impossible d'encadrer de manière utile l'erreur de mesure associée à un score particulier. Lorsque nous disons « de manière utile », nous voulons signifier qu'on peut exclure certaines valeurs, grandes, de l'erreur de mesure. Par exemple, lorsqu'un pédiatre mesure la taille d'un nouveau né, il n'a pas à considérer que l'erreur de mesure puisse dépasser disons un centimètre. Et si une telle erreur survenait, c'est la compétence du médecin qui serait incriminée et non pas la fidélité de la variable des mesures effectuées sur une population de nouveaux nés, parce qu'on ne tolère pas que l'erreur de mesure soit illimitée au regard des valeurs que le procédé de mesurage rend possibles.

Pour conclure cette première partie, nous voyons que la psychotechnique, i.e., le bricolage numérique des réponses aux items du test, confortée par l'élaboration psychométrique de la théorie classique des tests (pour le cas de la modélisation de la réponse à l'item, voir Vautier et al., 2012), ne répond pas à la

question de la possibilité du mesurage effectif de l'estime de soi. Elle conduit le chercheur à aller trop vite en besogne. Une telle impatience a un prix : il doit dorénavant contribuer à une science non plus des personnes (i.e., des systèmes psychologiques), mais des variables définies sur des populations statistiques de systèmes psychologiques. D'où le pouvoir irrésistiblement « enchanteur » des « sirènes » de la modélisation statistique, qui se nourrissent effectivement de populations et de variables (Fisher, 1925, introduction).

3. Du système psychologique quelconque au système psychologique typique

Dans cette partie, nous n'entrerons pas dans le détail de la problématique psychosociale et cognitive concernant le caractère causal de l'activité psychologique de comparaison sociale pour l'estime de soi. Notre objectif est de permettre au lecteur de comprendre dans le détail comment l'impératif néogaltonien du test de l'hypothèse nulle, qui stipule que le score moyen d'estime de soi ne dépend pas des conditions expérimentales, implique la perte de la généralité de l'hypothèse psychologique qui justifie tacitement la méthodologie de la recherche. L'hypothèse psychologique est que l'estime de soi de toute personne dépend du résultat, appréciateur ou dépréciateur, d'un certain jugement évaluatif qu'elle porte sur elle-même et « les autres », groupe dont elle fait partie ou pas, à tel sujet (par exemple, la performance dans tel domaine de tâches).

Pour tester l'égalité de deux moyennes (peu importe qu'elles soient indépendantes ou « appariées ») à l'aide du t de Student, on a besoin de considérer deux lois normales $N(\sigma_1, \mu_1)$ et $N(\sigma_2, \mu_2)$. La question qui se pose au chercheur

soucieux de se conformer à l'impératif néo-galtonien est de savoir, étant données les estimations (s_1, m_1) et (s_2, m_2) permises par deux échantillons d'observations, si on peut considérer que l'égalité $\mu_1 = \mu_2$ est fausse. Notre propos ne consiste pas à critiquer les fondements logiques de l'inférence statistique ni les interprétations invalides des résultats calculés par les logiciels d'analyse de données (à ce sujet, voir par exemple Cohen, 1994; Gliner, Leech, & Morgan, 2002; Gonzalez, 1994; Guenther, 2002; Harcum, 1990; Jones & Sommerlund, 2007; Kalinowski, Fidler, & Cumming, 2008; Kluger & Tikochinsky, 2001; Krueger, 2001; Lecoutre & Poitevineau, 2000; Marewski & Olsson, 2009; Meehl, 1997; Morgan, 2003; Mulaik, Raju, & Harshman, 1997; Nickerson, 2000). Il s'agit plutôt de montrer que le rejet de l'hypothèse nulle telle qu'elle est définie dans l'approche néo-galtonienne constitue un « hors sujet » au regard de l'hypothèse, générale au sens strict, qu'une variable psychologique quantitative dépende d'une variable indépendante donnée.

Supposons qu'on ait rejeté l'égalité $\mu_1 = \mu_2$ à juste titre, c'est-à-dire qu'il soit vrai que les lois normales (ou tout au moins les lois de probabilité) en question aient des moyennes distinctes. Le chercheur peut-il affirmer qu'il a testé son hypothèse psychologique ? Non, parce qu'il n'a fait aucun test approprié. Les objets sur lesquels il a acquis une connaissance sont des lois normales qui ne décrivent aucun système psychologique concret. Les lois normales, pour recevoir une interprétation psychologique, doivent être conçues comme le résultat d'une expérience de pensée analogue à celle qui est nécessaire à la définition du score vrai (voir 2.5) : si on met une personne quelconque dans telle situation, si on est

capable de répliquer l'expérience, peu importe la personne observée, un nombre de fois infini, si on admet que la loi normale est un modèle acceptable de la distribution des scores observés, alors on obtiendra un score moyen qui diffèrera du score moyen qu'on obtiendrait si on changeait, de manière spécifique, la condition dans laquelle le score est observé. Ainsi, le score observé est logiquement déconnecté de la personne qui l'a produit.

La population statistique sur laquelle est fondée la variable des scores d'estime de soi dans telle ou telle condition est définie par des unités d'observation qui sont des triplets (personne, temps, condition expérimentale). Par exemple, le triplet (Paul, hier à 14 heures, condition 1) appartient à la population statistique des triplets (personne, temps, condition 1) parce qu'hier à 14 heures, Paul se trouvait dans la condition 1. Cette population désigne un ensemble d'états du monde (passés et futurs) du point de vue du score d'estime de soi. Cet ensemble d'états du monde est représenté par une distribution de probabilités théorique (discrète et finie puisque les scores sont définis dans une échelle de score discrète et finie), dont le chercheur ne peut avoir que des aperçus, via des échantillons d'observations. La moyenne (ou, plus convenablement, la médiane) de cette distribution décrit un triplet typique qui est non pas une abstraction au sens strict, c'est-à-dire une propriété qui serait commune à tous les triplets de la population, mais une virtualité caractéristique – un paramètre statistique.

Ce paramètre est sans pertinence empirique pour le psychologue dès lors que celui-ci définit son objet comme un être humain (un système psychologique) particulier descriptible dans le temps. Les personnes particulières sont nécessaires

à la définition des populations requises par le test de l'égalité de deux moyennes néo-galtoniennes. Mais une personne particulière, disons Paul, placée dans la condition 1, définit une autre population statistique, à savoir l'ensemble des triplets (Paul, temps, condition 1), où le temps désigne l'ensemble des « moments » au cours desquels Paul a été ou sera placé dans la condition 1. L'inclusion de cette dernière population dans la première possède une asymétrie qui empêche que la population formée par les triplets (personne, temps, condition 1) décrive adéquatement la population formée par les triplets (Paul, temps, condition 1) du point de vue du score d'estime de soi.

Il ne nous appartient pas de justifier l'intérêt du test de l'égalité de deux moyennes tel qu'il est prescrit par la méthodologie néo-galtonienne, puisque nous défendons l'idée que cette approche est méthodologiquement « hors sujet » pour une psychologie générale. L'erreur fondamentale que recèle le test de l'hypothèse statistique nulle lorsqu'il porte sur des groupes de personnes, est qu'il conduit le chercheur à perdre de vue les implications théoriques de son hypothèse psychologique initiale. S'il raisonne en termes de tendance caractéristique d'une personne quelconque, alors le chercheur n'a d'autre choix que de reformuler son hypothèse de la façon suivante : quelle que soit une personne, elle aura tendance à produire un score d'estime de soi plus faible dans telle condition que dans telle autre condition. Pour qu'il puisse évaluer une telle hypothèse, le chercheur doit collecter des observations assez nombreuses de l'estime de soi de cette personne dans les deux conditions, pour examiner la différence des deux moyennes observées. Par conséquent, le chercheur devra revenir au problème du mesurage

de l'estime de soi, non pas sous l'angle de la mesurabilité ordinaire cette fois, mais sous l'angle de sa capacité à obtenir des mesures indépendantes, c'est-à-dire telles qu'une *n*ème mesure ne dépende pas des mesures précédentes mais seulement de l'estime de soi lors du *n*ème mesurage. Or la technique des auto-évaluations paraît peu compatible avec l'exigence technique de l'indépendance expérimentale, puisque la formation des jugements nécessaires à la production des réponses aux items laisse vraisemblablement des traces dans la mémoire du répondant et qu'il est difficile d'exclure que ces traces ne jouent aucun rôle causal dans la production des réponses lors de la *n*ème passation du questionnaire. On assiste ici, en quelque sorte, au « retour du refoulé épistémologique » quant aux conditions nécessaires à une connaissance objective des phénomènes mentaux. Il est pratiquement difficile d'obtenir les deux distributions de scores d'estime de soi qui sont nécessaires pour tester l'idée que chez une personne donnée, les tendances auxquelles on s'intéresse sont différentes. On remarquera que cette difficulté permet de comprendre en quoi les agrégats de l'approche néo-galtonienne sont attrayants, puisque la notion de réplication expérimentale ne pose pas de problème : les réponses de Paul hier à 14 heures lorsqu'il était dans la condition 1 sont indépendantes des réponses de Julie aujourd'hui à 10 heures lorsqu'elle était dans la condition 1.

En supposant que le chercheur surmonte ces difficultés pratiques et rejette l'hypothèse nulle selon laquelle, chez Paul, les deux tendances sont identiques, il ne peut généraliser ce résultat à toute personne autre que Paul de manière (logiquement) valide. C'est pourquoi il doit reconnaître que l'état de ses

connaissances (et de celles de la communauté des chercheurs concernés par son thème de recherche), nécessite de progresser cumulativement, système psychologique par système psychologique. Cette nécessité méthodologique s'applique également lorsqu'il s'agit de tester que les réponses d'une personne particulière possèdent des propriétés d'ordre simple (cf. 2.3).

On peut récapituler le résultat de notre analyse de la manière suivante : le rejet de l'hypothèse nulle, hypothèse qui stipule que le score moyen d'estime de soi ne dépend pas de la condition expérimentale dans laquelle on place n'importe quelle personne, conduit le chercheur à s'illusionner de deux manières. Premièrement, il croit à tort qu'il sait mesurer l'estime de soi (s'il ne le croit pas, on ne comprend pas pour quelles raisons scientifiques il effectue le test de l'hypothèse nulle de l'égalité des deux moyennes). Le scorage des réponses au questionnaire d'estime de soi produit un résultat qui imite le résultat d'un processus de mesurage qui reste inconnu ou qui n'est pas testé. Deuxièmement, une différence statistiquement significative de deux moyennes néo-galtoniennes permet de mettre en évidence un phénomène statistique dont le support empirique est indéterminé. Ce phénomène résulte du fait qu'on invoque des populations statistiques qui ne caractérisent les tendances de personne en particulier. Par conséquent, le rapport d'une différence statistiquement significative n'a pas de pertinence pour l'hypothèse psychologique qui a motivé le travail de recherche.

4. La recherche de relations générales et restrictives

Le manque de pertinence de la méthodologie néo-galtonienne (i.e., fondée sur l'agrégation de personnes dans des groupes expérimentaux ou quasi-expérimentaux) pour la recherche en psychologie générale ne doit pas surprendre étant donné le faible niveau d'élaboration théorique de l'hypothèse psychologique. Au moment de conclure, nous aurons à nuancer une telle affirmation en envisageant la possibilité que, perdant son hégémonie, cette méthodologie est néanmoins utile pour répondre à certaines questions de « débroussaillage ». Pour l'instant, nous considérons que le chercheur subit l'hégémonie néo-galtonienne.

Tout d'abord, l'hypothèse psychologique porte sur une grandeur théorique à laquelle on n'a pas accès. Le chercheur a seulement accès aux réponses apportées aux sept items du questionnaire d'estime de soi. S'il souhaite proposer une hypothèse de recherche en utilisant ces réponses pour construire sa variable dépendante, alors il doit réaliser que ses conjectures portent sur la manière dont les personnes répondent au questionnaire, puisque c'est ce qu'il observe. En particulier, l'ensemble des observables est composé de 7-uplets ordonnés de manière incomplète (il suffit de montrer qu'il existe deux 7-uplets observés qui indiquent des changements de sens opposés). Ensuite, la notion de détermination des réponses par les conditions expérimentales demande à être détaillée.

Nous avons vu plus haut (cf. 2.2) qu'il paraît invraisemblable que les réponses dépendent uniquement d'une grandeur psychologique qu'on a envie d'appeler l'estime de soi. En effet, on s'attend à ce que les réponses aux items

dépendent d'un processus psychologique complexe qui est, pour ainsi dire, une boîte noire (voir Tourangeau, Rips, & Rasinski, 2000). Il est possible que le caractère dépréciatif ou appréciatif des cognitions auto-évaluatives influence le choix des réponses aux items, mais le détail de ce processus n'est pas conceptualisé. La remarque de Windelband que nous avons reproduite plus haut s'applique ici pleinement : on ne peut connaître que ce qui a été préalablement conceptualisé. À supposer qu'on dispose d'un modèle cognitif assez détaillé pour qu'on puisse concevoir comment les cognitions auto-évaluatives font « pencher la balance », le problème de la caractérisation empirique du processus de choix, qui se substitue alors à la grandeur « estime de soi », demeure entier. Tant qu'il n'est pas résolu, le modèle cognitif demeure une conjecture non testable.

Face à un tel diagnostic épistémique, le chercheur est forcé de considérer que son hypothèse psychologique est au minimum prématurée. On ignore à quelles lois psychologiques les réponses au questionnaire pourraient obéir. Cette ignorance est reconnue par les psychométriciens, dont le point de départ consiste à postuler qu'une réponse à un item est le résultat d'une expérience aléatoire (non pas empirique, mais imaginaire), ce qui entraîne que les modèles psychométriques ne sont valables que pour des applications actuarielles (Lord, 1980, p. 3; Rasch, 1960, p. 11), ce que Vautier et al. (2012) appellent l'ingénierie sociale. Dès lors qu'on distingue la psychologie générale de l'ingénierie sociale, la pertinence des méthodes de cette ingénierie pour la psychologie générale demande par principe à être évaluée.

La réflexion épistémologique est utile pour recouvrer le sens de ce qui caractérise les connaissances scientifiques et nous nous appuyerons cette fois sur un épistémologue contemporain : « [...] les sciences ne tirent pas leur part de certitude d'un accès aux (hypothétiques) propriétés intrinsèques des choses, mais d'une maîtrise des conditions d'établissement de relations entre des phénomènes eux-mêmes relatifs aux formes de connaissances » (Bitbol, 2010, p. 13). Nous préciserons dans un premier temps ce que nous comprenons par « relations entre des phénomènes » et dans un deuxième temps nous en tirerons les conséquences méthodologiques pour une psychologie authentiquement générale.

4.1. La recherche de relations générales et restrictives

Avant de développer ce que signifie « une maîtrise des conditions d'établissement de relations entre des phénomènes », il faut clarifier ce qu'on entend par « relations entre des phénomènes ». En effet, les phénomènes peuvent désigner des réalités hypothétiques, comme des variations de la grandeur « estime de soi », auquel cas il est nécessaire de détailler comment on fait pour pouvoir dire quelque chose de valable à propos de ce qui reste une hypothèse. Ou bien les phénomènes désignent ce qui entre dans la catégorie du manifeste (ou encore de l'empirique), comme par exemple un 7-uplet de réponses. Du point de vue du mesurage, une grandeur théorique (comme la longueur d'un objet) est un concept qui correspond à un réseau de concepts logico-empiriques. Par exemple, si j'étire un élastique, sa longueur augmente. Autrement dit, nous considérons arbitrairement que les propriétés relationnelles que possède un objet concret dans un contexte englobant d'autres objets (l'élastique, ses points de fixation, etc.)

constituent une condition de possibilité pour qu'il y ait un sens à résumer ces propriétés à l'aide d'un langage abstrait (le langage de la longueur). Tant qu'on n'a pas réussi à établir une telle correspondance, on ne dispose que du langage des phénomènes manifestes pour décrire les objets étudiés – le langage métaphorique ne répondant pas aux besoins descriptifs. Par conséquent, nous interprétons les phénomènes dans la citation de Bitbol comme les phénomènes qui peuvent se manifester.

Il s'agit alors de clarifier la logique descriptive des phénomènes, ce qui revient à définir les modalités d'une variable, et à définir la nature logique de cet ensemble de modalités en le munissant de relations logiques. Il suffit ici de considérer que les modalités descriptives sont mutuellement exclusives. Par exemple, supposons qu'on sache décrire un système psychologique d'un certain point de vue qui est l'ensemble des possibilités de manifestation (ou modalités) $B = \{b_1, b_2, b_3\}$. Un phénomène est alors un événement au sens de la théorie des probabilités (e.g., Falmagne, 2003). Par exemple, tel objet placé dans telle condition se manifeste selon la modalité b_1 . On voit tout de suite apparaître une seconde variable descriptive dont les modalités sont les conditions « initiales » possibles. Appelons cet ensemble $A = \{a_1, a_2\}$ en supposant qu'on se limite à deux conditions seulement.

Nous pouvons alors définir une relation empirique binaire entre les deux types de phénomènes. Cette relation relie les modalités de A à celles de B. C'est une relation empirique dès lors qu'on considère des êtres empiriques dont la manifestation peut être décrite (connue) dans le référentiel

$$A \times B = \{(a_1, b_1), (a_1, b_2), \dots, (a_2, b_3)\},$$

qui est l'ensemble de tous les états qu'on peut concevoir en considérant conjointement les deux points de vue A et B. La connaissance de la relation empirique est donc connaissance de la relation qui se manifeste lorsqu'on s'intéresse à un être empirique particulier. Connaître cet être dans le référentiel de description $A \times B$ est équivalent à connaître le graphe de la relation qui le caractérise, c'est-à-dire l'ensemble des couples de $A \times B$ qui sont empiriquement possibles pour cet être particulier.

Ici, il est crucial de bien distinguer ce qui relève de la connaissance empirique et ce qui relève de la connaissance théorique de la relation. La connaissance empirique de la relation est l'ensemble des observations qui ont été réalisées (et mémorisées dans la communauté). Par exemple, un millier d'observations de Paul ont montré qu'il s'est trouvé dans les états $G_{\text{Paul}} = \{(a_1, b_1), (a_1, b_2), (a_2, b_1), (a_2, b_2), (a_2, b_3)\}$. La connaissance théorique de la relation est une conjecture. Par exemple, on suppose que le graphe théorique de Paul est le même que son graphe empirique, c'est-à-dire qu'il est psychologiquement impossible que Paul se trouve dans l'état (a_1, b_3) . La difficulté fondamentale est alors l'invention d'une explication de ce fait théorique (cf. Hempel, 2004, pp. 22-23), si, comme c'est le cas dans l'immense majorité des études psychologiques croyons-nous, on ne dispose pas d'une théorie préexistante de laquelle déduire ce graphe. Une théorie psychologique du choix du 7-uplet selon que le répondant a produit un jugement comparatif appréciateur ou dépréciateur sur lui-même n'existe pas. La théorie psychologique sur laquelle le chercheur s'appuie a pour

variable dépendante l'estime de soi et, comme on ne sait pas observer l'estime de soi, c'est une théorie non-empirique.

Cette relation théorique est restrictive parce que son graphe est un sous-ensemble strict du référentiel de la relation. Une relation est non restrictive si son graphe est égal à son référentiel ($A \times B$). Son caractère restrictif lui confère une valeur prédictive, puisqu'on peut prédire que si Paul se trouve dans l'état a_1 , alors il ne peut pas se trouver simultanément dans l'état b_3 . Cette prédiction n'est pas ponctuelle parce qu'il existe deux états possibles, b_1 et b_2 , conditionnellement à a_1 . Elle est générale parce que son domaine de généralité est un ensemble, celui de tous les « moments » dans lesquels Paul s'est trouvé et se trouvera dans l'état a_1 . Mais elle n'est pas générale dès lors qu'on s'intéresse à Julie par exemple, parce qu'on ne connaît pas la relation qui se manifeste dans le référentiel de description lorsque c'est Julie qui est observée.

En pratique, on ne dispose pas d'un millier d'observations de Paul dans le référentiel, mais tout au plus de quelques dizaines d'observations. Le chercheur peut alors introduire une supposition supplémentaire. Ce qui vaut pour Paul vaut pour quiconque. Dans la tradition néo-galtonienne de l'étude de personnes typiques, il peut alors s'intéresser à la relation qui émerge de l'observation de quelques dizaines de personnes dans le référentiel $A \times B$. Supposons que la même relation émerge de ces observations. Alors, il peut, par induction, généraliser ce résultat et proposer l'hypothèse, générale, que quelle que soit une personne, son comportement du point de vue du référentiel obéit à cette relation. Les connaissances empiriques produites par la méthodologie néo-galtonienne portent

sur des personnes typiques et non pas sur des personnes réelles parce qu'elles reposent sur des relations non-restrictives, qui n'ont aucune valeur prédictive puisqu'elles n'excluent rien. Au contraire, l'hypothèse d'une relation générale restrictive est falsifiable par l'observation d'un événement qui n'appartient pas à son graphe.

On voit ici que le recours au test de l'hypothèse nulle telle qu'elle est prescrite par le statisticisme est inutile. L'hypothèse théorique, qui est descriptive et non pas explicative (voir Canguilhem, 1958), est définie positivement au lieu d'être la négation, peu spécifique, d'une hypothèse nulle ponctuelle dans l'approche néo-galtonienne (voir aussi Meehl, 1967; 1990). Et son test consiste à observer ce qui se passe quand quiconque se trouve dans l'état a_1 , auquel cas l'état b_3 est un falsificateur, ou bien dans l'état b_3 , auquel cas l'état a_1 est un falsificateur.

4.2. Conséquences méthodologiques de la redéfinition de la question de recherche

Maintenant que nous avons explicité ce qu'est une relation restrictive et générale entre des phénomènes, nous pouvons comprendre la clé de la proposition de Bitbol (2010), qui est centrée sur la « maîtrise des conditions d'établissement de relations entre des phénomènes eux-mêmes relatifs aux formes de connaissances ». Les phénomènes sont relatifs aux formes de connaissances en ce sens que des connaissances préalables sont nécessaires pour qu'une description soit possible, c'est-à-dire pour qu'un référentiel de type $A \times B$ soit utilisable – dans le langage méthodologique de la psychologie, le référentiel est exactement $M(VI) \times M(VD)$, où $M(VI)$ est l'ensemble des modalités de la variable

indépendante et $M(VD)$ est l'ensemble des modalités de la variable dépendante. Maîtriser les conditions d'établissement de relations entre des phénomènes signifie alors, dans la perspective qui est la nôtre, trouver au moins un référentiel tel que, lorsqu'on s'intéresse à une collection de systèmes psychologiques (le singleton {Paul}, ou l'ensemble des adultes francophones par exemple), la relation descriptive qui lui est associée soit restrictive et générale. Nous avons appelé ailleurs une telle relation un fait empirique général non tautologique (Vautier, 2011b; 2013).

La tâche du chercheur, qu'il soit novice ou expert, consiste alors à examiner l'état des connaissances dans le domaine qui est le sien, en répondant à la question suivante : quels sont les référentiels de description qui ont déjà été utilisés ? S'il en trouve au moins un, la question suivante est alors : quelles sont les relations qui ont été décrites dans la littérature à partir de ce référentiel ? Le conformisme socio-technique que nous critiquons a pour conséquence que le chercheur ne pourra probablement pas répondre à la question, parce ce que ce qui est habituellement rapporté n'est pas le graphe de relations descriptives, mais quelque chose comme la significativité de la valeur d'une « variable test » associée à un test d'hypothèse nulle (voir aussi Krause, 2011). Nous voulons ici souligner le caractère contre-productif du conformisme socio-technique lorsqu'il formate l'activité intellectuelle des chercheurs en psychologie, puisqu'il les enjoint de ne pas s'intéresser à ce qui constitue le cœur même de la connaissance empirique générale. Si le chercheur ne trouve pas de référentiel de description, alors sa contribution possible est : en proposer un.

Arrivés à ce point de l'analyse, il nous semble que nous venons d'identifier une situation de crise, d'autant plus aiguë que le chercheur a une situation institutionnelle précaire, puisqu'il se trouve face à un dilemme. Soit il ignore l'impératif socio-technique de la communauté dans laquelle il cherche à s'inscrire et cherche un fait général, auquel cas, et s'il le trouve, il s'expose à un risque majeur de rejet de son rapport de recherche, qu'il a bien entendu soumis à une revue orthodoxe sélective. Comme la politique éditoriale de ladite revue est déterminée par « l'impact factor » (Ségalat, 2009), la probabilité que le manuscrit soumis à évaluation, s'il était publié, soit cité par de futurs articles, eux-mêmes soumis au même impératif bibliométrique, est nécessairement un critère d'évaluation majeur. Par conséquent, la résistance induite par le caractère insolite de la méthodologie mise en œuvre joue contre le manuscrit qui sera classé comme une bizarrerie. Soit il se soumet à l'impératif méthodologique en vigueur et ignore donc ce qui constitue à nos yeux la finalité d'une étude empirique standard. Il peut aussi essayer de convaincre sa communauté qu'un changement d'approche méthodologique est non seulement possible mais fécond, mais cette stratégie est pour le moins risquée.

Ce qui nous amène à examiner plus en détail les conséquences méthodologiques de l'impératif scientifique que nous défendons. Quelles sont les conditions nécessaires à la découverte d'une relation restrictive et générale lorsque la variable dépendante a par exemple $7^7 = 823543$ modalités ? Pour qu'une relation empirique puisse être restrictive, il faut qu'il existe au moins un état possible non observé. Mais pour qu'on puisse avancer qu'il s'agit d'un fait

théorique, c'est-à-dire que ce « trou » dans le référentiel résulte d'une contrainte réelle, il faut que la possibilité logique de ce « trou » soit très faible (Popper, 1973; 1992). Dans le cas où un état élémentaire quelconque est un 7-uplet parmi 7^7 , la possibilité d'un état quelconque qui est la réunion de p états élémentaires vaut $p/7^7$. Par conséquent la possibilité de sa négation vaut $1 - p/7^7$. Cette négation constitue ce qui est recherché. La possibilité de cette négation étant donné qu'on dispose de n observations indépendantes est donc $(1 - p/7^7)^n$. Cette possibilité est celle d'une relation restrictive générale spécifiant l'impossibilité psychologique (et non pas logique) de cet état. Si n est petit comparativement à 7^7 , la possibilité logique de la relation est grande, et l'édification de cette relation comme fait théorique sera peu convaincante. Autrement dit, il faut que le nombre d'observations soit largement plus important que le nombre d'états élémentaires possibles dans le référentiel de description pour que l'observation d'une relation restrictive puisse retenir l'intérêt du chercheur. Par conséquent, la taille de son référentiel de description doit être adaptée à sa capacité d'exploration, c'est-à-dire à sa capacité d'accumuler des observations indépendantes.

Cette contrainte épistémique est redoublée dès lors que le chercheur s'intéresse à un phénomène de changement d'états, ce qui est le cas dans notre exemple : si on passe d'un jugement dépréciatif à un jugement appréciatif, le vecteur des réponses évolue d'une certaine manière, et si on passe d'un jugement appréciatif à un jugement dépréciatif, le vecteur de réponses évolue d'une autre manière. Le problème consiste alors à spécifier ces « manières » dans un espace d'événements constitués de $7^7 \times 7^7$ 14-uplets possibles.

On peut aborder le problème en considérant que l'échelle de réponse à tout item permet d'observer plusieurs types de changements : -4 degrés (de 5 à 1), -3 degrés (de 5 à 2 ou de 4 à 1), -2 degrés (de 5 à 3, etc.), -1 degré, 0 degré, $+1$ degré, $+2$ degrés, $+3$ degrés ou $+4$ degrés. On peut alors définir un vecteur de changement en comptant le nombre de fois qu'on observe tel type de changement. Par exemple le vecteur de changement (0, 0, 0, 0, 3, 2, 1, 1, 0) indique que tous les changements observés sont nuls ou positifs. Un résumé encore plus concis du phénomène de transition d'états est fourni par le vecteur (0, 3, 4) qui compte le nombre de changements respectivement négatifs, nuls et positifs. Une hypothèse générale est alors que quelle que soit une personne testée dans la condition (dépréciation \rightarrow appréciation), son vecteur (x, y, z) est tel que $x < z$. Pour des raisons de place, nous épargnerons au lecteur le travail qui consiste à expliciter ce nouveau référentiel de description, à spécifier le graphe de la relation théorique et à évaluer le degré de sa testabilité, c'est-à-dire sa possibilité (logique).

Une telle approche permettrait vraisemblablement de repérer des contre-exemples. Ces contre-exemples, pourvu qu'ils ne soient pas trop nombreux (voir Vautier, 2011b), constituent une information de première importance pour la recherche, bien qu'elle soit systématiquement ignorée par la méthodologie néogaltonienne du test de la nullité de la différence des scores d'estime de soi moyens. Contre-exemple par contre-exemple, il s'agit d'essayer de comprendre comment il se fait que les jugements apportés par des personnes particulières ne se conforment pas à l'attente théorique, de manière à en affiner la formulation si cela est possible. C'est ici que la notion de maîtrise des conditions d'établissement

de la relation attendue prend pleinement son sens : quelles sont les conditions psychologiques suffisantes pour que le phénomène relationnel puisse se manifester (dans la même veine, voir aussi Krause, 2010) ? Répondre à une telle question nécessite que la recherche puisse, de manière orthodoxe, se centrer sur l'analyse de quelques cas individuels au nom d'une recherche de généralité (et non pas d'une singularité subjective dont la trivialité ne justifie aucune recherche scientifique). En raison même de la généralité des propositions qu'elle vise, la recherche en psychologie générale doit accorder une place centrale à la méthodologie clinique, dans la mesure où ce terme désigne la rencontre avec une personne particulière.

5. Conclusion

Cet article est une tentative pour attirer l'attention de la communauté des chercheurs en psychologie générale sur le fait que notre responsabilité scientifique n'est pas pleinement assumée dès lors que les normes socio-techniques qui contraignent la manière de faire de la recherche cessent d'être critiquables. La réflexion critique que nous avons développée repose sur l'idée que la méthodologie de la recherche est davantage qu'un effort de docilité vis-à-vis de certaines prescriptions – quantifier la variable dépendante conformément à la tradition psychométrique, échantillonner des observations et procéder à un test d'inférence statistique conformément à une norme que Danziger (1987; 1990) appelle néo-galtonienne. La méthodologie de la recherche contient aussi une réflexion critique qui répond à une recherche de pertinence : étant donné l'objectif

que s'assigne la psychologie générale, c'est-à-dire la découverte de principes psychologiques généraux, c'est-à-dire valables pour tout système psychologique et falsifiables (i.e., tels que l'observation d'un cas particulier peut en droit falsifier la « théorie »), les moyens mis en œuvre pour constituer une démarche de recherche concrète sont-ils pertinents ?

L'exemple du test de l'égalité des scores moyens d'estime de soi obtenus sur deux groupes de personnes est illustratif d'une docilité méthodologique qui conduit le chercheur à manquer de pertinence. Tout d'abord, le scorage de l'estime de soi à partir de vecteurs de réponses à un questionnaire d'auto-évaluation n'accomplit le mesurage d'une grandeur théorique pour quiconque répond au questionnaire que si on veut bien admettre qu'un certain nombre d'analyses psychométriques, i.e., effectuées à l'échelle d'un agrégat, fonde la précision métrologique d'un procédé numérolgique qui opérationnalise une métaphore. Mais aucun mesurage n'est accompli, si on entend par mesurage le fait qu'on dispose d'une théorie corroborée de la manière dont les vecteurs dépendent, de manière unique, de ladite grandeur. Par conséquent, l'impératif de l'opérationnalisation des grandeurs psychologiques fonctionne comme un obstacle (Bachelard, 1983) au repérage des problèmes à résoudre. Il est possible que le mesurage des grandeurs psychologiques ne trouve pas de solution même dans un avenir lointain. Il est donc vain, du point de vue de la contribution à la connaissance scientifique, de s'atteler à l'étude des déterminants d'une grandeur psychologique tant qu'on ne sait pas la mesurer.

Ensuite, la mise en place d'un plan expérimental (ou quasi-expérimental) fondé sur des agrégats de systèmes psychologiques conduit le chercheur à définir son objet d'étude comme un système psychologique typique. Il se sert des systèmes psychologiques particuliers pour édifier une abstraction. Certes, l'abstraction est consubstantielle à la connaissance scientifique mais le trait définitoire d'une abstraction scientifique est qu'elle caractérise tout objet concret (elle est une classe d'équivalence), ce qui n'est pas nécessairement le cas du système psychologique typique. Une hypothèse psychologique est générale si elle se laisse formuler par un « quel que soit ». Quelle que soit une personne, on affirme quelque chose qui est supposé vrai à son propos. L'énoncé d'une différence de moyennes n'implique aucune généralité. Krause (2011) a récemment discuté ce point en détail dans le contexte des applications psychothérapeutiques. Une différence de moyennes peut résulter du fait que le domaine des valeurs possibles dans telle condition ne recouvre pas celui des valeurs possibles dans telle autre condition, auquel cas il existe une relation générale restrictive ou encore un fait général, mais la réciproque est fautive. Le schéma « scorage de l'estime de soi, échantillonnage néo-galtonien et test de l'hypothèse nulle de l'égalité des moyennes » conduit le chercheur, lorsqu'il rejette cette hypothèse, à affirmer quelque chose qui n'est valable pour personne en particulier dès que les faits empiriques ne sont pas des faits généraux.

La réflexion critique ne doit pas conduire au décret autoritaire de non pertinence inconditionnelle de ce schéma. Nous visons seulement à saper les fondements de son caractère hégémonique. Un chercheur qui veut vérifier qu'une

condition est un « facteur causal » par rapport à une autre condition peut bien entendu, en guise de préliminaire, attester qu'il existe un phénomène de masse qu'il mettra en évidence avec une variable dépendante de convenance. Mais cela constitue le début de la quête de connaissances générales et non pas une fin.

Nous espérons que la notion de relation générale et restrictive puisse à bon droit entrer dans le cahier des charges d'une formation en méthodologie. Les concepts nécessaires pour la maîtrise des conséquences de cette réflexion méthodologique sont les mêmes que les concepts nécessaires au maniement éclairé des probabilités (voir aussi l'avant-propos de Noël, 2013). La notion de graphe restreint est exactement équivalente à la notion probabiliste d'un événement composé (Rényi, 1966, chapitre 1) dont la probabilité est nulle. On comprend que si la formation méthodologique en psychologie consiste à privilégier le savoir faire au détriment du savoir pourquoi faire, l'utilité sociale au détriment de l'intelligibilité désintéressée, bref le conformisme socio-technique au détriment de l'audace et de la rigueur intellectuelle, les chances de découvrir des phénomènes empiriques généraux en psychologie sont faibles. Il ne tient qu'à nous, enseignants-chercheurs en charge des enseignements de méthodologie et de la direction des travaux de recherche, d'assumer nos responsabilités pédagogiques et scientifiques vis-à-vis de la jeune génération.

Références

Bachelard, G. (1983). *La formation de l'esprit scientifique* (12e ed.). Paris: Vrin.

- Bitbol, M. (2010). *De l'intérieur du monde: pour une philosophie et une science des relations*. Paris: Flammarion.
- Borsboom, D., Cramer, A., Kievit, R. A., Scholten, Z., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135-170). Charlotte, NC: Information Age Publishing.
- Canguilhem, G. (1958). Qu'est-ce que la psychologie ? *Revue de Métaphysique et de Morale*, 1, 12-25.
- Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- Cronbach, L. J., & Meehl, P. H. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Danziger, K. (1985). The methodological imperative in psychology. *Philosophy of the Social Sciences*, 15, 1-13.
- Danziger, K. (1987). Statistical method and the historical development of research practice in American psychology. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution, Vol. 2: Ideas in the sciences* (pp. 35-47). Cambridge: MIT Press.
- Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. New York: Cambridge University Press.
- DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks: Sage.

- Essex, C., & Smythe, W. E. (1999). Between numbers and notions: A critique of psychological measurement. *Theory & Psychology, 9*, 739-767.
- Falmagne, J. C. (2003). *Lectures in elementary probability theory and stochastic processes*. Boston: McGraw Hill.
- Fisher, R. A. (1925). *Statistical methods for research workers*. London: Olivier & Boyd.
- Gaudron, J. P., & Vautier, S. (2007). Estimating true short-term consistency in vocational interests: A longitudinal SEM approach. *Journal of Vocational Behavior, 70*, 221-232.
- Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *Journal of Experimental Education, 71*, 83-92.
- Gonzalez, R. (1994). The statistics ritual in psychological research. *Psychological Science, 5*, 321, 325-328.
- Granger, G.-G. (1987). Conventions, normes, axiomes dans la connaissance des faits humains. *Dialectica, 41*, 39-67.
- Guenther, R. K. (2002). How probable is the null hypothesis? *American Psychologist, 67*-68.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.
- Harcum, E. R. (1990). Deficiency of education concerning the methodological issues in accepting null hypotheses. *Contemporary Educational Psychology, 15*, 199-211.

- Hempel, C. (2004). *Eléments d'épistémologie* (2e ed.). Paris: Armand Colin.
- Johnson, H. M. (1935). Some neglected principles in aptitude testing. *American Journal of Psychology*, *47*, 159-165.
- Johnson, H. M. (1936). Pseudo-mathematics in the mental and social sciences. *American Journal of Psychology*, *48*, 342-351.
- Johnson, H. M. (1943). Index-numerology and measures of impairment. *American Journal of Psychology*, *56*, 551-558.
- Jones, A., & Sommerlund, B. (2007). A critical discussion of null hypothesis significance testing and statistical power analysis within psychological research. *Nordic Psychology*, *59*, 223-230.
- Kalinowski, P., Fidler, F., & Cumming, G. (2008). Overcoming the inverse probability fallacy: A comparison of two teaching interventions. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4*, 152-158.
- Kluger, A. N., & Tikochinsky, J. (2001). The error of accepting the "theoretical" null hypothesis: The rise, fall, and resurrection of commonsense hypotheses in psychology. *Psychological Bulletin*, *127*, 408-423.
- Krause, M. S. (2010). Trying to discover sufficient condition causes. *Methodology*, *6*, 59-70.
- Krause, M. S. (2011). Statistical significance testing and clinical trials. *Psychotherapy*, *48*, 217-222.
- Krueger, J. (2001). Null hypothesis significance testing. *American Psychologist*, *56*, 16-26.

- Lakatos, I. (1978). Falsification and the methodology of scientific research programmes. In J. Worrall & G. Currie (Eds.), *Imre Lakatos: Philosophical papers. Vol. 1: The methodology of scientific research programmes* (pp. 8-101). Cambridge: Cambridge University Press.
- Lamiell, J. T. (2013). Statisticism in personality psychologists' use of trait constructs: What is it? How was it contracted? Is there a cure? *New Ideas in Psychology, 31*, 65-71.
- Laveault, D., & Grégoire, J. (2002). *Introduction aux théories des tests en psychologie et en sciences de l'éducation*. Paris: De Boeck Université.
- Lecoutre, B., & Poitevineau, J. (2000). Aller au-delà des tests de signification traditionnels : vers de nouvelles normes de publication. *L'Année Psychologique, 100*, 683-713.
- Lord, F. M. (1980). *Applications of item response theory to practical problems*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA.: Addison-Wesley.
- Marewski, J. N., & Olsson, H. (2009). Beyond the null ritual: Formal modeling of psychological processes. *Zeitschrift für Psychologie/Journal of Psychology, 217*, 49-60.
- Meehl, P. H. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

- Meehl, P. H. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108-141.
- Meehl, P. H. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393-425). Mahwah, NJ: Erlbaum.
- Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355-383.
- Michell, J. (1999). *Measurement in psychology*. Cambridge: Cambridge University Press.
- Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology, 10*, 639-667.
- Michell, J. (2003). The quantitative imperative: Positivism, naïve realism, and the place of qualitative methods in psychology. *Theory & Psychology, 13*, 5-31.
- Michell, J. (2006). Psychophysics, intensive magnitudes, and the psychometricians' fallacy. *Studies in History and Philosophy of Biological and Biomedical Sciences, 17*, 414-432.

- Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspectives*, 6, 7-24.
- Morgan, P. L. (2003). Null hypothesis significance testing: Philosophical and practical considerations of a statistical controversy. *Exceptionality*, 11, 209-221.
- Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65-115). Mahwah, NJ: Lawrence Erlbaum Associates.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241-301.
- Noël, Y. (2013). *Psychologie statistique avec R*. Paris: Springer.
- Penta, M., Arnould, C., & Decruynaere, C. (2005). *Développer et interpréter une échelle de mesure : Applications du modèle de Rasch*. Sprimont: Mardaga.
- Popper, K. R. (1973). *La logique de la découverte scientifique*. Paris: Payot.
- Popper, K. R. (1992). *Realism and the aim of science*. London: Routledge.
- Rasch, G. (1960). *Probabilistic models for some intelligence tests and attainment tests*. Copenhagen: Denmark's Paedagogiske Institut.
- Redersdorff, S., & Martinot, D. (2003). Impact des comparaisons ascendantes et descendantes sur l'estime de soi : importance de l'identité mise en jeu. *L'Année Psychologique*, 104, 411-444.
- Rényi, A. (1966). *Calcul des probabilités*. Paris: Dunod.

- Roberts, F. S. (1985). *Measurement theory, with applications to decision-making, utility, and the social sciences*. Reading, MA: Addison-Wesley.
- Ségalat, L. (2009). *La science à bout de souffle ?* Paris: Seuil.
- Soler, L. (2000). *Introduction à l'épistémologie*. Paris: Ellipes.
- Thompson, B., & Vachaa-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195.
- Tourangeau, R., Rips, L. J., & Rasinski, A. (2000). *The psychology of survey response*. New York: Cambridge University Press.
- Vautier, S. (2011a). Measuring change with multiple visual analogue scales: Application to tense arousal. *European Journal of Psychological Assessment, 27*, 111-120.
- Vautier, S. (2011b). The operationalisation of general hypotheses versus the discovery of empirical laws in Psychology. *Philosophia Scientiae, 15*, 105-122.
- Vautier, S. (2013). How to state general qualitative facts in psychology? *Quality and Quantity, 47*, 49-56.
- Vautier, S., Gaudron, J. P., & Jmel, S. (2004). Modèles factoriels linéaires pour l'analyse de la fidélité des variables composites. *Revue d'Epidémiologie et de Santé Publique, 52*, 441-453.
- Vautier, S., & Jmel, S. (2003). Transient error or specificity? An alternative to the staggered equivalent split-half procedure. *Psychological Methods, 8*, 225-238.

- Vautier, S., Steyer, R., & Boomsma, A. (2008). The true-change model with individual method effects: Reliability issues. *British Journal of Mathematical and Statistical Psychology*, *61*, 369-399.
- Vautier, S., Veldhuis, M., Lacot, E., & Matton, N. (2012). The ambiguous utility of psychometrics for the interpretative founding of socially relevant avatars. *Theory & Psychology*, *22*, 810-822.
- Windelband, W. (2002). *Qu'est-ce que la philosophie? et autres textes*. Paris: Vrin.