De l'influence du prototypage et du design expérimental sur l'efficacité d'un Nudge

Etinene BRESSOUD, Directeur Général Délégué, BVA Nudge Unit, eb@bvanudgeunit.com

Virginie MAILLE, Professeur, SKEMA Business School, virginie.maille@skema.edu

Ali DRHIMEUR, Consultant, Cognizant, ali.drhimeur@gmail.com

Pierrick RIVIERE, PhD, Nudge France, pierrickriviere@yahoo.com

Richard BORDENAVE – Directeur Général Afjoint, Innovation, Transformation, BVA, richard.bordenave@bva-group.com

Les auteurs remercient l'association NudgeFrance, qui a mis ses données à disposition, les étudiants et enseignants qui ont participé au NudgeChallenge COP21 organisé par NudgeFrance, et/ou qui ont testé les Nudges créés pour ce challenge, et le Ministère de la Transition écologique et solidaire qui a subventionné les expérimentations.
Nous remerciaons particulièrement Michel Hourdebaigt (Ministère de la Transition écologique et solidaire), David Nahon (AgroParisTech), Agnès Helme-Guizon (IAE Grenoble), Frédéric Gonthier (Sciences Po Grenoble), Sarah Mussol (Université de Montpellier), Christophe Benavent (Université Paris-Nanterre) et Jean-Baptiste Légal (Université Paris-Nanterre).

The influence of prototyping and experimental design on Nudge efficacy

Abstract: Nudging provides multiple subtle ways of encouraging individuals to adopt sustainable behaviors by helping them switch from intending to acting. Proving the efficacy of a nudge implies to run experimentation, while some scientific papers question the replication of previous studies particularly in social sciences. This article focuses on the impacts of the design and execution of a nudging experiment on the proof of its efficacy. 22 groups of students executed and tested nine nudging ideas to incentivize responsible behaviors in favor of the climate, with respect to the targets fixed by the COP21. An exploratory analysis shows that starting from the same video brief, the execution and the experimental design of a nudge vary mong the different conducted tests. Those differences in the experiment drive different conclusions on the efficacy of a nudge: the similarity of both sample composition and Key Performance Indicators for different nudge executions are the two main factors influencing replicability more than prototyping and experimentation duration. Results confirm the need for conducting in parallel both the generation and the execution of a nudge, and the need for adapting the test according to the nudge development step and the objective of its evaluation.

Keywords: Nudge, Sustainable development, Prototype, Design, Experiments

De l'influence du prototypage et du design expérimental sur l'efficacité d'un Nudge

Résumé : Le Nudge est une incitation douce, pour favoriser l'adoption de comportements durables de la part des individus, en les aidant à passer de l'intention à l'acte. Prouver l'efficacité d'un Nudge nécessite de l'expérimenter. Dans ce cadre, cet article s'intéresse à l'influence de l'exécution et du design d'expérimentation d'un Nudge sur la preuve de son efficacité et s'inscrit dans le débat autour de la réplicabilité des études en sciences sociales. 22 groupes d'étudiants ont exécuté et testé neuf idées de Nudges incitant à des comportements responsables en faveur du climat, conformément aux objectifs fixés par la COP21. Une analyse exploratoire montre que, à partir d'un même brief, l'exécution et le design d'expérimentation d'un Nudge peuvent varier et amener à conclure différemment quant à l'efficacité d'un Nudge : des échantillons et des indicateurs d'efficacité identiques sont des facteurs facilitant la réplicabilité, au-delà de la similarité des prototypes et de la durée de l'expérimentation. Ces résultats confirment l'intérêt d'avancer en parallèle l'idéation et l'exécution d'un Nudge, et d'adapter la méthodologie de test en fonction de la phase de développement du Nudge et de la finalité de son évaluation.

Mots clés : Nudge, Développement durable, Prototype, Design, Expérimentation

Introduction

Si 61 % des consommateurs dans le monde se déclarent très concernés par les enjeux environnementaux (Greendex, 2014), bien moins agissent réellement pour la planète. L'écart entre intentions et comportements réels est tout aussi important lorsqu'il s'agit, par exemple, de bouger plus ou d'arrêter de fumer, et voilà des décennies que les sciences comportementales s'intéressent aux règles qui régissent ce type de comportement. Ayant montré que la façon de présenter les différentes options peut influencer la décision, Thaler et Sunstein (2008) ont proposé d'aider les individus à passer de l'intention au comportement en leur soumettant des Nudges. Ces « coups de pouce » consistent à créer des « architectures de choix » qui augmentent l'intérêt de l'option la plus avantageuse pour les individus, sans contraindre leur choix.

De nombreuses publications traitent des Nudges. Il ne s'agit pas ici de démontrer à nouveau leur intérêt, ni d'expliquer comment les créer. Nous reviendrons plutôt sur la façon d'évaluer leur efficacité. Sunstein (2014) recommande pour cela la « preuve par le test », impliquant de passer de l'idée de Nudge à un prototype, puis de concevoir et mettre en œuvre un plan d'expérimentation. On part donc de l'idée du Nudge et d'un brief. Si l'on connaît exactement l'environnement physique et temporel auquel le Nudge est destiné, le brief peut être précis, et le prototype du Nudge et les conditions du test proches de la solution finalement mise en œuvre. Un tel test devrait avoir un pouvoir prédictif relativement élevé. Mais le Nudge peut être implémenté dans plusieurs environnements à la fois, par ailleurs pas toujours observables au moment du test (notamment quand ils n'existent pas encore), et il est alors difficile de le tester dans les conditions exactes de sa future mise en œuvre. On en testera plutôt le principe, laissant une marge de manœuvre à l'expérimentateur quant au prototype et au plan d'expérimentation à créer. A l'heure où la littérature académique pose la question de l'impartialité du chercheur (Mazar et Ariely, 2015) et de la réplicabilité des études (Baker, 2016), notamment en sciences sociales, quel crédit accorder aux conclusions de tels tests ? Quelles conclusions pourront en tirer les académiques quant à leur modèle ? Et quelles décisions pourront prendre les professionnels ?

Le présent travail vise à étudier l'influence de l'exécution et du design d'expérimentation d'un Nudge sur la preuve de son efficacité. Après avoir défini ce que sont les Nudges et comment tester leur efficacité, une analyse de la littérature portant sur les réplications permet d'identifier les éléments du test qui peuvent conduire à des conclusions divergentes. Ayant ensuite confié à 30 groupes d'étudiants le test de 15 idées de Nudge, une étude empirique exploratoire met en évidence combien un même brief peut déboucher sur des prototypes et plan d'expériences variés, et combien ces différences peuvent impacter le sens des conclusions. Compte tenu de son caractère exploratoire et de ses limites, ce travail ne permet pas à ce stade d'établir une liste exhaustive de règles à suivre pour tester un Nudge, mais quelques pistes sont déjà discutées.

Que sont les Nudges, comment les créer et tester leur efficacité ?

Les courants de pensée dominants en économie théorique supposent que l'individu décide de façon rationnelle, dans une logique *d'homo economicus* régie par les prix et l'utilité perçue. Ces facteurs constituent la base de la plupart des modèles économiques des entreprises et des politiques économiques et monétaires.

Cela dit, les avancées en sciences comportementales des dernières décennies ont montré les limites d'une telle rationalité, la complexité du processus de prise de décision, et l'influence des biais cognitifs. L'hypothèse d'une fonction d'utilité représentative des préférences suppose une consistance de choix rarement vérifiée en réalité. De même, imposant un effort de réflexion mathématique important, l'hypothèse de maximisation s'avère difficilement défendable. Dès

1957, Herbert Simon introduit la notion de « rationalité limitée ». Du fait de ses capacités cognitives et du coût de l'information à obtenir, l'individu est plus enclin à faire un choix non optimal et basé sur l'heuristique. En 2003, Kahneman distingue deux systèmes de réflexion, l'un, intuitif, associatif et rapide dans le traitement de l'information, l'autre, plus réfléchi et plus lent. Dans un environnement sous pression, l'individu tend à utiliser le premier, utilisant des raccourcis cognitifs qui le rendent plus vulnérable aux erreurs de jugement.

Dans ce cadre, l'environnement social et physique et la façon dont les différentes options sont disposées peut influencer la décision et conduire l'individu à faire un choix qui n'est pas dans son intérêt, en tout cas pas à long terme. Introduite par Thaler et Sunstein (2008), « l'architecture du choix » met en évidence l'influence que peut avoir la disposition des options disponibles sur la prise de décision. Le « Nudge » consiste alors à modifier cette architecture afin d'aider l'individu à opter pour l'option la plus avantageuse pour lui. Il ne s'agit pas de contraindre son choix, aucune option n'est éliminée, ni aucune incitation économique mise en avant (Hausman et Welch 2010). En 2014, Sunstein retient quatre critères d'une intervention Nudge : la liberté de choix, la transparence, l'efficacité, et enfin, la preuve par le test.

De nombreuses publications, dont Goepel, Rossini Rahme et Svenhall (2015), expliquent comment concevoir un Nudge. Mais tout de suite après la phase d'idéation du Nudge se pose la question de son efficacité. Le chercheur doit d'abord passer de l'idée de Nudge à son exécution, puis définir et mettre en œuvre un plan d'expérimentation. Si l'idée de Nudge est normalement matérialisée par un brief très précis, ce n'est pas toujours possible. Il arrive par exemple qu'un Nudge soit destiné à plusieurs environnements différents, imposant des adaptations du Nudge et de son test d'un environnement à l'autre. Il se peut aussi que le Nudge soit destiné à un environnement qui n'existe pas encore. C'est le cas, par exemple, lorsque l'OGIC envisage de « Nudger » les habitants d'un futur bâtiment (Launiau, Bakoula, et Théry, 2017) que l'expérimentateur devra imaginer, ou pour un événement futur comme Paris 2024. C'est ainsi qu'une même idée de Nudge pourra être exécutée et testée de différentes façons. En théorie, si l'expérimentateur est impartial et expert, tous les tests – ou réplications – devraient conduire à une même conclusion : le Nudge est efficace, ou non. Mais la réalité est tout autre... Quels facteurs peuvent causer ces divergences ? Et que penser de réplications qui ne débouchent pas sur des résultats convergents ? Cela remet-il en cause l'efficacité du Nudge testé ?

Les réplications en sciences sociales et les sources d'hétérogénéité

Par « réplication », le *Subcommittee on Replicable Science* (SBE) désigne la reproduction des résultats d'une étude, en collectant de nouvelles données au moyen des mêmes procédures que celles de l'étude initiale (Sayre et Riegelman, 2018). La « *replicability* » fait l'objet ces dernières années d'une littérature abondante en sciences sociales, source d'inspiration des Nudges. Au demeurant, elle souffre aussi de confusions. On la nomme parfois « *reproducibility* », à tort puisque la reproduction ne désigne normalement qu'une nouvelle analyse statistique des données originales de l'étude (SBE). De même, si la réplication utilise les « mêmes procédures » que l'expérience initiale, elle ne consiste pas en une simple « répétition » de l'expérience, laquelle vise à répéter à l'identique le protocole expérimental initial (Serra, 2012).

Qu'est donc exactement une réplication et quel intérêt a-t-elle ? A la différence de la reproduction, on collecte de nouvelles données. Et pour ce faire, à la différence de la répétition, le protocole initial subit quelques modifications mineures, l'idée étant de répliquer un phénomène (par exemple, une réaction favorable après exposition à un Nudge), pas une expérience. On utilise donc des protocoles légèrement différents (paramètres, instructions, incitations, caractéristiques des sujets…), qui déboucheront sans doute sur des données différentes, mais dans le but d'inférer l'existence du même phénomène. En cela, la réplication permet de tester la robustesse du résultat initial (Serra, 2012), qui s'avère d'autant plus robuste

que des résultats similaires sont obtenus à partir d'un grand nombre de réplications. Selon Simons (2014), il s'agit du "*best and possibly the only believable evidence for the reliability of an effect*". C'est sans doute la raison pour laquelle, jusque-là peu valorisées par la communauté académique, les réplications sont aujourd'hui encouragées par les revues, qui leur dédient notamment des rubriques ou numéros spéciaux. Recherchant dans PsycINFO les articles académiques dont le titre contient le terme « réplication » (ou un synonyme), Anderson et Maxwell (2016) observent ainsi une accélération significative de ce type de publication : 82 articles en 2003, 121 en 2008, 154 articles en 2013.

Mais nombre d'articles récents font aussi état de difficultés à répliquer les études publiées en sciences sociales. Certains s'alarment de taux de réplication exagérément bas, provoquant une « crise de confiance » (Pashler et Wagenmakers, 2012). Par exemple, parmi les 14 réplications tentées en psychologie par Nosek et Lakens (2014), neuf ne confirment pas du tout les résultats initiaux, et les cinq autres n'en confirment qu'une partie, dans des conditions spécifiques ou avec une taille d'effet plus petite. De la même manière, ayant travaillé sur 100 études publiées dans des revues de psychologie sociale, l'Open Science Collaboration Psychology (Nosek et al., 2015) observe un taux de réplication inférieur à 50%. Seulement 37% des réplications obtiennent des résultats significatifs et 47% des effets de taille originaux se retrouvent dans un intervalle de confiance à 95%. Citons enfin l'étude menée par Nature, qui montre que 70% des chercheurs en psychologie et en cancérologie ont déjà échoué à répliquer les résultats d'une étude antérieure (Baker, 2016). Certes, d'autres articles, notamment en psychologie et en économie comportementale, débouchent sur des conclusions moins pessimistes. Proposant d'autres manières de mesurer si le résultat est répliqué ou non (Goodman et al., 2016), ils rapportent des taux de réplication plus élevés (Camerer et al., 2016 ; Klein et al., 2014 ; Patil et al., 2016 ; Etz et Vandekerckhove, 2016) et remettent en cause la « *replication crisis* » (Fanelli, 2018). La question reste néanmoins posée : pourquoi les réplications ne conduisent pas toujours à des résultats convergents, et que faut-il en penser ?

De multiples raisons peuvent être citées, dont bien sûr l'impartialité et/ou l'expertise insuffisantes de certains chercheurs (Bench et al., 2017 ; Open Science Collaboration Psychology, 2015 ; Van Bavel et al., 2016). Mais il s'avère que même une recherche d'une qualité exemplaire peut produire des résultats empiriques non réplicables. Cela peut simplement résulter du hasard ou d'une erreur systématique (Nosek et al., 2015), mais pas lorsque les réplications non convergentes sont nombreuses. On pourra aussi mentionner la pression à la publication d'effets nouveaux et contrintuitifs, souvent subtils (Baker, 2016 ; Kruglanski, Factor et Jasko, 2018). Mais qu'en est-il des raisons inhérentes à la méthodologie du test ?

De nombreux auteurs s'accordent à dire que, en matière de recherche comportementale, même la simple « répétition » d'une expérimentation ne peut être strictement identique à l'expérimentation originale (Brandt et al., 2014 ; Fabrigar et Wegener, 2016 ; Rosenthal, 1991 ; Stroebe and Strack, 2014 ; Tsang and Kwan, 1999). Par-là, elles ne débouchent pas systématiquement sur des résultats convergents. C'est par exemple ce qu'ont observé Nosek et al. (2015), en dépit de la reprise fidèle des procédures et matériel initiaux, fournis et contrôlés par les auteurs eux-mêmes. Visant à répéter les effets de plusieurs études, le projet « *Many Labs* » (Klein et al., 2014), mené par 36 laboratoires indépendants, observe également des résultats relativement hétérogènes d'une étude à l'autre, en dépit d'un matériel identique. On peut attendre moins de convergence encore pour une réplication qui ne reproduit pas l'expérience originale à l'identique.

Quels éléments de la réplication peuvent entraver la convergence des résultats ? En réalité, la plupart des éléments du plan d'expérience (Brown et al., 2014 ; Klein et al., 2014 ; McShane et Böckenholt, 2017) : ceux qui sont décrits dans l'article original, mais qui ne sont pas toujours reproductibles à l'identique du fait de contraintes liées à l'environnement, mais aussi ceux dont on ne suspecte pas l'effet, qui ne sont ni contrôlés ni décrits (Brown et al.,

2014). Or, en psychologie, il est rare que les comportements ne résultent que d'un seul effet cognitif (Smets, 2018). Ils sont plutôt multi-déterminés, et les facteurs non contrôlés peuvent produire un bruit de fond ou un effet de plafond masquant l'effet étudié, surtout s'il est subtil (Kruglanski, Factor et Jasko, 2018). Des facteurs non contrôlés peuvent aussi interférer avec l'effet étudié. Pour reprendre l'exemple de Smets (2018), quand un sujet fait son choix, est-il influencé par ce qui lui semble familier (Status Quo), ou nouveau et différent (Nouveauté) ?

Dans le cadre des tests d'efficacité d'un Nudge, on considèrera les éléments suivants :

- Citons en premier lieu la (les) variable(s) manipulée(s). Lors d'un test d'efficacité de Nudge, l'exécution du prototype peut fortement varier, même à partir d'un même brief. Ce peut être parce que le brief ne décrit normalement que les éléments qui constituent le(s) levier(s) du Nudge, et ne peut détailler tout l'environnement. Ce peut être aussi du fait des orientations des personnes qui développent le prototype. Plusieurs expériences menées à la BVA Nudge Unit montrent que l'exécution d'un même Nudge peut varier selon les personnes impliquées dans la réalisation du projet (Soubils et Serin, 2018 ; Soubils et Niclas, 2017). Or, si l'on considère qu'en matière de Nudge, « *everything matters* » (Thaler et Sunstein, 2008), tous les éléments de l'environnement peuvent être source de variation. Par exemple, si l'on propose une poubelle de recyclage de journaux en forme d'ours polaire, quelle expression faciale donner à l'ours ? Si le designer la souhaite « neutre », il est néanmoins possible que l'ours soit perçu comme souriant, ou accusateur, ou triste… De la même manière, si le Nudge consiste en des instructions verbales, s'agira-t-il d'une voix féminine ou masculine ? et peut-on garantir que le ton, le timbre de voix ou l'accent n'auront aucun effet ? De même, place-t-on le Nudge à l'entrée ou à la sortie, à droite ou à gauche, à quelle distance ? Enfin, ne permettant pas toujours un dispositif identique, les contraintes propres à chaque contexte, peuvent également déboucher sur des exécutions variables. Par exemple, il est possible que la poubelle de recyclage ne puisse être mise à un mètre de la sortie du fait de la présence d'un distributeur de boisson.

- Il est souvent conseillé de tester les Nudges en effectuant un Essai Randomisé Contrôlé (ERC) (Haynes et al., 2013), mais le design peut varier. Certains choisiront un design de type avant/après, tandis que d'autres compareront les comportements d'un échantillon soumis au Nudge à ceux d'un échantillon contrôle (Soman, 2015). Si la première option peut faire craindre un effet d'histoire, certains environnements ne permettent parfois pas d'autre option. Un tel design peut être également intéressant s'il s'agit de tester un Nudge auprès d'une population familière avec l'environnement étudié, sans doute plus résistante au changement. En tout état de cause, une différence de design peut inverser les résultats.

- Appartenant au contexte, les variables externes du plan d'expérience peuvent, elles aussi, impacter le résultat des tests d'efficacité de Nudges. Il peut s'agir de l'environnement physique, social et temporel où le Nudge est mis en place. Au sein d'une même expérience, on peut essayer de contrôler ces variables de telle sorte qu'elles soient constantes avec et sans le Nudge testé. Par exemple, on peut systématiquement mesurer les comportements de recyclage dans une station de métro donnée, un jour et une heure donnés. On pourra peut-être ainsi supposer un environnement social identique avec et sans Nudge. A l'inverse, la similitude de toutes les variables externes entre une expérience et ses réplications est illusoire. Par exemple, quel résultat obtiendrait-on si on souhaitait mettre en place ce Nudge dans un espace qui n'est pas ouvert ce jour-là ? Le projet « *Many Labs* » a montré que l'utilisation d'une même étude, auprès d'un même échantillon, avec le même matériel, mais un jour différent, est source d'hétérogénéité (Klein et al. 2014). Et qu'obtiendrait-on si on testait le même Nudge aux mêmes jour et heure, mais dans une station différente du métro parisien, ou dans le métro d'une autre ville, ou ailleurs que dans un métro ? L'environnement social ne serait certainement pas comparable... Or il peut modérer l'efficacité du Nudge du fait de l'information et/ou de la pression supposée qu'il induit, même s'il n'a pas été envisagé comme un levier du Nudge lors

de l'idéation. Bien sûr, les expériences *in situ* plutôt qu'en laboratoire exposeront davantage à l'effet de variables non contrôlées, mais leur intérêt en terme de validité externe est à considérer.

- L'échantillon utilisé pour chaque test d'efficacité de Nudge peut également être source d'hétérogénéité des résultats. Outre les effets dus à la taille de l'échantillon, on considèrera aussi le mode de sélection et de recrutement (Brown et al. 2014). S'agit-il de participants concernés par le sujet, suffisamment pour s'intéresser un minimum au Nudge qui leur est soumis, mais pas trop pour éviter un effet plafond ? Ce peut être du fait de l'âge, de la culture, ou d'autres facteurs sociodémographiques. Par exemple, Gilbert et al. (2016) constatent qu'une étude menée aux US peut produire des résultats différents en Italie ou en Hollande du simple fait de la culture. Pettigrew (2018) observe que, d'un campus à l'autre au sein d'un même pays, les conclusions peuvent s'inverser. Enfin, s'il s'agit d'un échantillon étudiant, a-t-il été sollicité par un professeur, et comment ? Et quels objectifs, réels ou supposés, attribue-t-il à l'étude ?

- L'opérationnalisation de la (des) variable(s) dépendante(s) peut également altérer la convergence des résultats d'une étude à l'autre. Quels sont les KPIs à considérer pour tester l'efficacité du Nudge ? On mesure plus souvent des attitudes ou intentions plutôt que le comportement (Kruglanski, Factor et Jaśko, 2018). Comment la question est-elle posée ? Suggère-t-elle une réponse ? Laisse-t-elle imaginer ce que l'on recherche ? De même, les modalités de la variable dépendante peuvent aussi produire un effet plafond qui gommera l'effet à répliquer. Enfin, où se situe la question dans le questionnaire ? On sait notamment que demander au participant s'il est un homme ou une femme avant un exercice de maths fait chuter la performance des femmes du fait de la menace du stéréotype (Spencer, Steel et Quinn 1999).

Etude empirique

L'étude empirique ci-après observe, à titre exploratoire, en quoi des différences d'exécution et de plan expérimental peuvent conduire à des conclusions divergentes en matière de Nudge. 30 groupes d'étudiants se sont vus confier l'exécution et le test de 15 idées de Nudges incitant à des comportements en faveur du climat[1]. Aucune consigne ni indication ne laissait présager l'objet de la recherche aux étudiants et aux enseignants qui les encadraient. Les objectifs, leviers et prototypes de chaque Nudge étaient présentés dans une vidéo. Pour mieux les impliquer, les groupes pouvaient travailler sur le Nudge de leur choix.

Choisis par un seul groupe, quatre Nudges ont été exclus de l'analyse, de même que deux autres Nudges dont les tests n'étaient pas suffisamment aboutis. Les tests de neuf Nudges sont considérés dans cette étude, six ayant été choisis par 2 groupes, deux par 3 groupes et un par 4 groupes, pour un total donc de 22 tests.

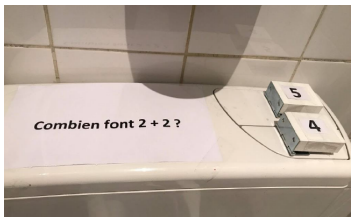
Les prototypes, plans d'expérience et résultats de chaque test ont ensuite été analysés par deux spécialistes des Nudges. Les désaccords ont été réglés par une discussion de convergence. Il suffit qu'un Nudge soit exécuté de manière différente par un groupe pour conclure à la possibilité d'exécuter différemment un même prototype à partir d'un brief commun. Par ailleurs, l'analyse des différentes réalisations doit permettre d'identifier les sources de variation possibles. La comparaison des designs de test pour un même Nudge a été menée selon les mêmes principes. Enfin, les mêmes analystes ont étudié les principales raisons – relatives au prototypage ou au design expérimental - conduisant à des résultats divergents.

Les résultats confirment que, à partir d'un même brief - une vidéo présentant un prototype du Nudge et une proposition de design de test - le prototypage et le design d'expérimentation varient sensiblement d'un groupe à un autre. Le tableau 1 montre par exemple un Nudge (*Stick n' Flush*, N.3) prototypé conformément au brief initial : inciter à appuyer sur le petit bouton de chasse d'eau (3L*) vs.* le gros bouton (6L) par le biais d'une

[1] Ces tests avaient initialement été créés par des étudiants dans le cadre du Nudge Challenge 2016 COP 21 organisé par Nudge France.

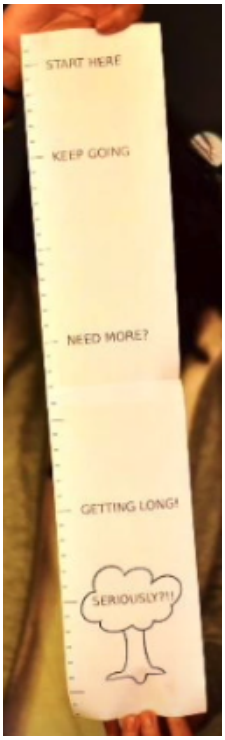
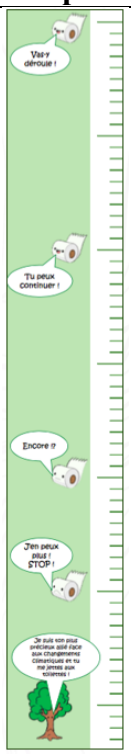
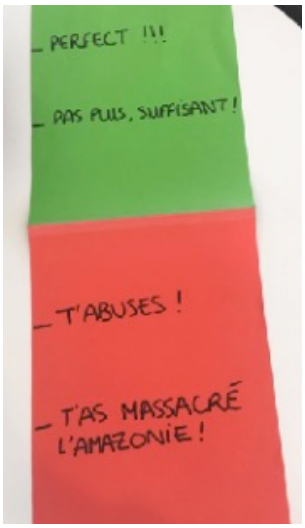
question inscrite sur les toilettes, et dont la bonne [mauvaise] réponse est inscrite sur le petit [gros] bouton. Au-delà de l'adaptation du Nudge à la chasse, qui varie d'un terrain d'expérimentation à l'autre, la mécanique du Nudge est inchangée par rapport au brief, proposant lui-même plusieurs exécutions (assumant que le Nudge repose sur la mécanique de la question posée, pas sur la nature de la question posée).

Tableau 1 – Prototypes similaires dans les groupes (Nudge *Stick n' Flush*, N.3)

Brief	Groupe 3.1	Groupe 3.2
		

Le tableau 2 montre à l'inverse le Nudge *Toilet Paper Ruler* (N.4, pour consommer moins de papier toilette), dont le prototype n'est pas conforme au brief. Le groupe 4.1 utilise des couleurs, dessins et messages différents, tandis que le groupe 4.2 recourt à des messages encore plus stigmatisants, plus courts, sans dessins et utilisant la couleur rouge. Tous les prototypes de tous les Nudges ont ainsi été comparés.

Tableau 2 – Prototypes différents dans les groupes (Nudges Toilet Paper Ruler, N.4)

Brief	Groupe 4.1	Groupe 4.2
		

La comparaison des designs expérimentaux portait sur les indicateurs d'efficacité (KPI), la durée des tests et le type de population testée. Le tableau 3 montre un exemple de Nudge, *The Mood Light Sticker* (N.9), dont l'objectif est d'économiser l'énergie. Les KPI sont différents d'un groupe à l'autre, de même que le temps d'exposition et la population testée.

Tableau 3 – Design de test différents dans les groupes (Nudge *Mood Light Sticker*, N.9)

Nudge	Groupe 9.1	Groupe 9.2
	KPI : KWH consommé Durée : 1 semaine Population : foyers	KPI : % de lumière éteinte Durée : 1 jours Population : étudiants

La comparaison des prototypes et designs expérimentaux pour l'ensemble des Nudges est synthétisée dans le tableau 4. Quels enseignements peut-on en tirer ?

Tableau 4 - Comparaison des tests.

Nudge		Type de Nudge (selon brief)	Nombre de tests	Tests partageant le (la) même…				Résultats conver-gents
				… prototype	… KPI	… durée	… population	
1	*Bread waste*	Objet, message et emplace-ment	4	3	3	0	3	non
2	***Food Waste***	**Objet et message**	**2**	**0**	**2**	**0**	**2**	**OUI**
3	***Stick n' Flush***	**Message**	**2**	**2**	**2**	**0**	**2**	**OUI**
4	*Toilet Paper Ruler*	Objet et message	3	0	n.c.	2	2	non
5	***Coffee lids***	**Emplace-ment**	**2**	**0**	**2**	**0**	**0**	**OUI**
6	*Trashball*	Objet	3	3	0	n.c.	2	non
7	*PutALid OnIt*	Objet	2	0	0	0	2	non
8	*bud project*	Objet	2	2	2	n.c.	0	non
9	*The mood light sticker*	Objet	2	0	0	0	0	non

On constate d'abord que, malgré des prototypes initiaux précis, seuls trois des neuf Nudges (N. 3, 6, et 8) sont conformes au brief et identiques entre eux. Un quatrième Nudge (N.1) a donné lieu à quatre tests, dont un présentait un prototype différent. Les cinq autres Nudges ont été prototypés de façon totalement différente. Les prototypes se ressemblent plus

lorsque le Nudge consiste en un objet physique plutôt qu'en une communication ou un message, souvent revus par les groupes. Sur les trois Nudges prototypés à l'identique, tous sont des objets. Pour autant, avoir un Nudge objet ne suffit pas à garantir un prototypage identique. Par exemple, consistant initialement en un objet, le Nudge N. 9 est parfois approximé par un message. Enfin, l'endroit où est placé ce prototype peut varier. Par exemple, dans le cadre d'un Nudge pour éviter le gâchis de pain en restauration scolaire (N.1), le Nudge est parfois placé en début de chaine, avant les entrées, ou en fin, juste avant la caisse, ce qui peut changer la nature du Nudge (Wansink et Hanks, 2013). Le design expérimental est également très variable. Les indicateurs d'efficacité des groupes testant un même Nudge sont identiques dans la moitié des cas (quatre Nudges sur huit), quelle que soit leur exécution. A l'inverse, la nature et la taille des échantillons, de même que les durées d'observation, diffèrent pour l'intégralité des Nudges.

Comme attendu, les différences observées en termes de prototypage et de design expérimental affectent la validation de l'efficacité des Nudges. Trois des neuf Nudges sont déclarés efficaces par tous les groupes. Nous réfléchissons ci-après aux éléments du test qui peuvent expliquer la convergence ou non des résultats.

Seul un des trois Nudges dont les conclusions sont identiques (Nudge N.3, *Stick n' Flush*) a été prototypé à l'identique par tous les groupes (tableau 5). Ceux obtenant des conclusions identiques ont été prototypés différemment. Par exemple, visant à réduire le gâchis de pain en restaurant collective, le Nudge *Bread Waste* (N.1) ajoutait un couvercle sur la panière pour augmenter la distance psychologique avec le pain. Certains groupes ont opté pour un couvercle transparent, d'autres non, d'autres ont ajouté un message (tableau 6), le tout induisant une distance psychologique sans doute variable, mais suffisante pour être efficace.

Tableau 5 - Prototypes et conclusions identiques (Nudges Stick n' Flush, N.3)

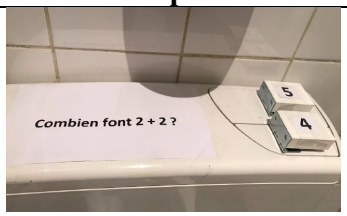
Groupe 3.1	Groupe 3.2
	

Tableau 6 - Prototypes différents et conclusions identiques (Nudge Bread Waste, N.1)

Groupe 1.1	Groupe 1.2
	

Qu'en est-il des différences au niveau du plan expérimental ? Tous les Nudges ayant été testés sur des durées différentes, aucune comparaison n'est possible concernant cet élément. Par ailleurs, tous les tests concluant de façon convergente ont des indicateurs d'efficacité

identiques (N.2, 3 et 5). La similarité des KPI est la seule variable commune aux Nudges ayant des conclusions identiques. Nous retrouvons ici le Nudge *Stick n' Flush* (N.3), dont l'efficacité est validée à partir du nombre d'appuis sur le petit bouton (Tableau 1). Pour autant, comme le montrent les résultats du Nudge *Bread Waste* (N.1), un même indicateur ne garantit pas la réplicabilité des résultats. Bien que mesurant tous deux la quantité de pain gaspillée par rapport au nombre de repas servis, les groupes 1.1 et 1.3 concluent à l'opposé (Tableau 7). Ainsi, des KPI identiques constituent une condition nécessaire à la convergence des résultats, mais pas suffisante.

Tableau 7 - KPI identiques et conclusions différentes (Nudges Bread Waste, N.1)

Groupe 1.1	Groupe 1.3
	
efficacité validée	efficacité non validée

Enfin, si deux Nudges sur les trois ayant des résultats identiques sont testés sur des échantillons tirés de populations-mères identiques (N.2 et 3), ce n'est pas le cas d'un autre (N.7), dont les résultats ne convergent pas, alors que tous les tests ont été menés sur des échantillons étudiants.

Pour conclure, la réplicabilité des résultats ne peut être garantie par aucun des critères retenus. Elle peut en revanche être facilitée par le recours à des échantillons tirés d'une même population-mère et à des indicateurs d'efficacité identiques, au-delà de la similarité des prototypes et de la durée de l'expérimentation.

Discussion

Ce travail exploratoire s'interrogeait sur la façon de tester l'efficacité d'un Nudge et sur l'influence de l'exécution et du design d'expérimentation sur la preuve de cette efficacité.

Certes, il ne nous a pas été possible de contrôler la totalité des éléments des expérimentations comparées dans cette étude, ni le savoir-faire des étudiants qui les ont menées. Si les tests ont été menés dans le cadre de cours d'études marketing, sous la supervision d'un enseignant, nous n'écartons pas quelques imperfections. Néanmoins, le nombre de tests menés et analysés permet de montrer combien un même Nudge, pourtant décrit à partir d'un brief précis, peut être exécuté et/ou testé de façons variables. Nous observons également combien la conclusion dépend du design experimental. Plusieurs éléments du design expérimental pouvant varier d'un test à l'autre, il est difficile d'identifier avec certitude ceux qui ont conduit à des conclusions divergentes, et c'est là une autre limite de ce travail. Cependant, pris dans leur ensemble, les résultats confirment les propos de Brown et al., (2014), Klein et al., (2014), ou McShane et Böckenholt (2017) au sujet des réplications : selon la manière de le prototyper et de le tester, un même Nudge peut être déclaré comme efficace ou non.

Que faire alors ? Au-delà des recommandations classiques quant à l'impartialité et au savoir-faire de l'expérimentateur, on recommandera d'abord d'avancer en parallèle l'idéation et l'exécution du Nudge qui doivent être pensées de manière intégrée et non linéaire (Govindarajan et Trimble, 2010). Si une dernière adaptation du prototype à l'environnement

peut ensuite être nécessaire, une co-création entre les équipes à l'origine de l'idée du Nudge et celles en charge de son prototypage permettra de tester un Nudge plus conforme à l'esprit initial. Or, l'idéation et le prototypage sont souvent conduits par des équipes différentes, et le passage de l'une à l'autre est souvent négligé. Plus généralement, le prototypage et les tests sont souvent moins dotés de moyen que ce qu'ils devraient (Soubils et Serin, 2018). Le design du test pourra également dépendre de la finalité du test, du degré d'avancement du projet, et des moyens à disposition. Dans une perspective académique, le chercheur souhaitera en premier lieu démontrer l'effet du Nudge, mais aussi en comprendre les mécanismes. On optera alors pour des *Randomized Controlled Trials* (essais randomisés contrôlés), visant à contrôler les variables autres que la (les) variable(s) indépendante(s). Répliquer les tests, avec des prototypes et designs voisins mais présentant des différences mineures (Serra, 2012), et précisément décrits (Brown et al. 2014), permettra de tester la robustesse des résultats, d'identifier des facteurs d'efficacité, voire des variables modératrices, et ainsi d'améliorer progressivement le dispositif (Smets, 2018). L'ensemble des tests pourra également donner lieu à des méta-analyses (Borenstein et al., 2009 ; Cooper, Hedges, and Valentine, 2009 ; Hedges and Olkin, 1985 ; Hunter and Schmidt, 2014 ; McShane et Bökenholt, 2017). De même, pour comprendre les mécanismes en jeu, on peut envisager des expériences manipulant les variables modératrices et/ou médiatrices précédemment identifiées a priori en sorte de valider plus rigoureusement leur effet. A l'inverse, dans une perspective professionnelle, l'objectif du test est essentiellement de prendre une décision, et dans des délais courts. La rigueur de la perspective académique, au-delà de son coût, peut décourager toute innovation. Dans un souci d'efficience, on optera plutôt pour des *Pragmatic Controlled Trials*, moins coûteux et plus proches de la réalité (Porzsolt et al., 2015 ; Morales, Amir et Lee, 2017). Tandis que la force des *Randomized Controlled Trials* est d'estimer un effet dans des conditions contrôlées, celles des *Pragmatic Controlled Trials* est de chercher ce qui marche, pour qui, et dans quelles circonstances (Dalziel et al., 2018). Mais dans ce souci d'efficience, on arbitrera aussi entre le prix de la fiabilité et le gain espéré. Il conviendra d'investir dans la mesure selon la nature de la décision et les actions qui découleront des KPI, conduisant souvent à prioriser en fonction des différents attendus possibles et des conséquences en termes d'action. On pourra aussi adopter une logique de « *test, learn and adapt* », généralement mise en œuvre dans les démarches de la Behavioral Insight Team anglaise (Haynes et al., 2012). Il s'agit d'un processus d'amélioration inférentielle continue et de maîtrise du risque, face une décision qui peut supporter une part d'incertitude, voire d'échec. Grace à de multiples itérations, sur peu d'individus mais dans la vraie vie, on obtiendra à la fin un prototype pré-testé et utilisable (comment un nouveau produit) avec une viabilité acceptable mais pas de garantie de succès (comme une start-up). Bien sûr, on veillera enfin à implémenter le Nudge de manière identique à ce qui a été fait lors des tests.

Conclusion

A partir des tests de neuf Nudges menés par 22 groupes d'étudiants, cette étude a démontré que les différences observées en termes de prototypage et de design expérimental ont des implications sur la validation de l'efficacité des Nudges testés. Dans certains cas, la réplicabilité des résultats d'un même Nudge peut être validée si les échantillons étudiés sont tirés d'une même population-mère et par le recours à des indicateurs d'efficacité identiques, quel que soit le degré de similitude des prototypes et de la durée de l'expérimentation. Néanmoins, pour une meilleure efficacité et réplicabilité, il peut être utile d'associer les équipes de conception et d'expérimentation du Nudge. De même, une approche d'expérimentation agile de type « *test and learn* » peut être adaptée à la finalité, professionnelle ou académique, du test.

Bibliographie

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*(7604), 452.

Bench, S. W., Rivera, G. N., Schlegel, R. J., Hicks, J. A., & Lench, H. C. (2017). Does expertise matter in replication? An examination of the reproducibility project: psychology. *Journal of Experimental Social Psychology*, *68*, 181-184.

Borenstein M., Hedges L.V., Higgins J.P.T., Rothstein H.R. (2009), Introduction to Meta-Analysis, Chichester, UK: Wiley.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication?. *Journal of Experimental Social Psychology*, *50*, 217-224.

Brown, S. D., Furrow, D., Hill, D. F., Gable, J. C., Porter, L. P., & Jacobs, W. J. (2014). A duty to describe: Better the devil you know than the devil you don't. *Perspectives on Psychological Science*, *9*(6), 626-640.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433-1436.

Cooper H. M., Hedges L.V., Valentine J.C. (2009), *The Handbook of Research Synthesis and Meta-Analysis*, New York: Sage.

Dalziel M. (2018), Why are there (almost) no randomised controlled trial-based evaluations of business support programmes?, *Palgrave Communications*.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PloS one*, *11*(2), e0149794.

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68-80.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to?. *Proceedings of the National Academy of Sciences*, *115*(11), 2628-2631.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). A Response to the Reply to our Technical Comment on "estimating the Reproducibility of Psychological Science".

Goepel, N., Svanhall, F., & Rahme, M. (2015). Strategic Recommendations for the Design of Nudges towards a Sustainable Society.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Sci. Transl. Med.* 8, 341ps12.

Govindarajan, V., & Trimble, C. (2010). *The other side of innovation: Solving the execution challenge*. Harvard Business Press.

Greendex. (2014). Consumer choice and the environment–a worldwide tracking survey.

Hausman, D. M., & Welch, B. (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy*, *18*(1), 123-136.

Haynes, L., Goldacre, B., & Torgerson, D. (2012). Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials| Cabinet Office.

Hedges L.V. & Olkin I. (1985), *Statistical Methods for Meta-Analysis*, Orlando, FL: Academic Press.

Hunter J.E. & Schmidt F.L. (2014), *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, Newbury Park, CA: Sage.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, *93*(5), 1449-1475.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability. *Social psychology*.

Kruglanski, A. W., Factor, A., et Jaśko, K. (2018). Is "behavior" the problem? *Social Psychological Bulletin*, 13(2).

Launiau, E., Bakoula, B. et Théry, C. (2017). Immeuble nudge, de l'architecture aux architectures de choix. *Guide de l'économie comportemental, ed. Labrador BVA.*

Mazar, N., & Ariely, D. (2015). Dishonesty in scientific research. *The Journal of clinical investigation*, *125*(11), 3993-3996.

McShane, B. B., & Böckenholt, U. (2017). Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research*, *43*(6), 1048-1063.

Morales A., Amir O., Lee L. (2007), Keeping It Real in Experimental Research—Understanding When, Where, and How to Enhance Realism and Measure Consumer Behavior, *Journal of Consumer Research*, (44), 465-76.

Nosek, B. A., & Lakens, D. (2014). Registered reports.

Nosek, B. A., Aarts A. A., Anderson C.J., Anderson, J. E., Kappes H. Barry (2015), Estimating the reproducibility of psychological science, Science, 349 (6251).

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

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science*, *7*(6), 528-530.

Patil P, Peng RD, Leek JT. (2016), What should researchers expect when they replicate studies? A statistical view of replicability in psychological science, *Perspect Psychol Sci*., 11:539–544.

Pettigrew, T. F. (2018). The emergence of contextual social psychology. *Personality and Social Psychology Bulletin*, 0146167218756033.

Porzsolt F. (2015), Efficacy and effectiveness trials have different goals, use,
Pragmatic and Observational Research, 6, 47–54.

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior and Personality*, *5*(4), 1.

Rosenthal R. (1991), "Replication in Behavioral Research," in *Replication Research in the Social Sciences*, ed. James W. Neulip, Newbury Park, CA: Sage, 1–30.

Sayre, F., & Riegelman, A. (2018). The reproducibility crisis and academic libraries. *College & Research Libraries*, *79*(1), 2.

Serra, D. (2012). Principes méthodologiques et pratiques de l'économie expérimentale: une vue d'ensemble. *Revue de philosophie économique*, *13*(1), 21-78.

Simon, H. (1957). *Models of Man*. New York, Wiley & Sons.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76-80.

Smets, K. (2018). There Is More to Behavioral Economics Than Biases and Fallacies, *Behavioral Scientist*.

Soman, D. (2015). *The last mile: Creating social and economic value from behavioral insights*. University of Toronto Press.

Soubils, M. L. et Niclas, J. (2017). Le Nudge au service de la propreté dans les trains OUIGO. *Guide de l'économie comportemental, ed. Labrador BVA.*

Soubils, M. L. et Serin, F. (2018). Le Nudge pour une meilleure perception des blocs sanitaire et des comportements plus vertueux. *Guide de l'économie comportemental, ed. Labrador BVA.*

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, *35*(1), 4-28.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59-71.
Sunstein, C. R. (2014). Nudging: a very short guide. *Journal of Consumer Policy*, *37*(4), 583-588.
Sunstein, C. R., & Thaler, R. H. (2008). *Nudge. The politics of libertarian paternalism*. New Haven.
Tsang, E. W., & Kwan, K. M. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management review*, *24*(4), 759-780.
Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454-6459.
Wansink, B., & Hanks, A. S. (2013). Slim by design: serving healthy foods first in buffet lines improves overall meal selection. *PloS one*, *8*(10), e77055.