HEC MONTRÉAL

Études sur la fiabilité, la validité et la sensibilité des échelles de mesure à un item en expérience utilisateur

par

Matthieu Cuvillier

Pierre-Majorique Léger

Sylvain Sénécal

HEC Montréal

Directeurs de recherche

Sciences de la gestion
(Spécialisation Expérience utilisateur dans un contexte d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion
(M. Sc.)

Janvier 2021

Résumé

Les échelles de mesure à un item telles que le « Net Promoter Score » sont régulièrement utilisées en expérience utilisateur. Elles permettent de mesurer les perceptions résultant d'une interaction de manière très rapide : en peu de temps, il est possible d'obtenir beaucoup de réponses sur plusieurs construits différents et cela à travers un grand nombre de répondants. Cet avantage pratique est pourtant teinté d'importantes faiblesses psychométriques. La recherche tend à pointer le fait qu'elles ne représentent pas totalement toute la richesse d'un construit, qu'elles manquent de fiabilité et de validité. Jusqu'à présent, peu de recherches dans le domaine de l'expérience utilisateur ont été réalisées sur ce sujet et aucun accord n'a été identifié au sujet de leur fiabilité et de leur validité. Aucun accord n'a été trouvé quant à leur utilisation ou non.

Ce mémoire tente de comprendre et d'explorer jusqu'à quel point le fait d'utiliser ce format d'échelles représente une manière fiable, valide et suffisamment sensible pour permettre de capter ce que vivent réellement les utilisateurs durant et après une interaction. Une étude en laboratoire a été menée auprès de 40 participants. Des corrélations ont été calculées entre les différentes mesures explicites et implicites de l'expérience utilisateur. Les résultats de cette étude s'avèrent partagés. Les directions des corrélations sont, en très nette majorité, contradictoires. Certaines corrélations sont significatives, d'autres ne le sont pas. Leur significativité semble différer selon des moments spécifiques de l'interaction (première et dernière impression), mais également selon la présence ou l'absence d'éléments particuliers présents sur les interfaces testées. Il semble apparaître, aussi, que cette significativité est plus fréquente lorsque les participants interagissent avec des éléments considérés plutôt comme hédoniques. Nos résultats mettent en avant le fait que ce type d'échelles ne mesurent pas la perception d'une expérience dans sa totalité. Les limites de ces mesures sont à prendre en compte lorsque nous sommes amenés à utiliser ces échelles.

Mots clés : « Affective Slider »; « CES »; « CSAT »; Expérience utilisateur; Fiabilité; Mesures à un item; « NPS »; Psychométrie; Sensibilité; Validité

Méthodes de recherche : Expérimentation; Recherche quantitative

Table des matières

Résumé iii
Table des matières v
Liste des tableaux et des figures vii
Tableaux vii
Chapitre 1 vii
Chapitre 2 vii
Figures vii
Chapitre 1 vii
Chapitre 2 vii
Liste des abréviations ix
Avant-propos xi
Remerciements xiii
Introduction 1
Mise en contexte 1
Objectifs de l'étude et questions de recherche 2
Structure du mémoire 5
Informations sur le chapitre 1 5
Informations sur le chapitre 2 6
Contributions et responsabilités personnelles 7
Références 8
Chapitre 1 Revue de littérature 11
1.1 Introduction 11
1.2 Validation et méthodes de construction d'échelles de mesure 12
1.3 Mesurer les perceptions à l'aide d'échelles à un item : avantages et inconvénients 21
1.4 Mesurer les perceptions à l'aide d'échelles à un item en UX 25
1.5 Conclusion 34
Références 35
Abstract 43

2.1 Introduction 43
2.2 Litterature review & hypothesis development 46
2.3 Method 53
2.4 Results 59
2.5 Discussion 65
Appendices 71
References 72
Conclusion 81
Contributions théoriques 84
Références 87
Bibliographie 89

Liste des tableaux et des figures

Tableaux

Tableau 1. Contributions de l'étudiant 7

Chapitre 1

Tableau 1.1. Mesures utilisées dans notre recherche 28

Tableau 1.2. Sélection des principaux antécédents et conséquences des quatre mesures utilisées dans notre étude 33

Chapitre 2

Table 2.1. Psychometric measures 57

Table 2.2. Descriptive statistics of the psychometric variables 60

Table 2.3. Descriptive statistics of the implicit measures without considering the website attributes 60

Table 2.4. Descriptive statistics of the implicit measures during the different contexts of use: hedonic ("H") and utilitarian ("U") 61

Figures

Chapitre 1

Figure 1.1. Schématisation du NPS 29

Chapitre 2

Figure 2.1. Data collection procedure 55

Figure 2.2. Example of AOIs drawn in a website used 58

Figure 2.3. Overview of the correlations between our variables (H1) 62

Figure 2.4. Overview of the correlations between our variables considering AOIs (H2) 64

Liste des abréviations

CES	Customer Effort Score
CSAT	Customer SATisfaction Score
NPS	Net Promoter Score
UX	Expérience utilisateur

Avant-propos

L'autorisation de rédiger ce mémoire par articles a été obtenue par la direction du programme de M. Sc. de HEC Montréal. Ce mémoire a donc été rédigé sous la forme d'une revue de littérature et d'un article de recherche. L'accord des coauteurs a également été obtenu pour que ce dernier soit présenté dans ce mémoire. Par ailleurs, le comité d'éthique en recherche (CER) a également approuvé ce projet de recherche.

Remerciements

Je tiens tout d'abord à remercier Pierre-Majorique Léger et Sylvain Sénécal, mes deux codirecteurs de recherche. Votre support ainsi que vos conseils m'ont été d'une aide plus que précieuse dans l'élaboration de ce mémoire. Merci de m'avoir fait grandir en me donnant la chance d'intégrer le Tech3Lab.

Merci aussi à toute l'équipe du Tech3Lab pour votre soutien durant toute la collecte de données.

Je ne pourrai pas rédiger ces remerciements sans mentionner l'aide et les nombreux conseils que j'ai pu recevoir de la part de deux statisticiens : Shang-Lin Chen et Carl Saint-Pierre. Un mot : merci! Merci pour votre patience et votre pédagogie.

Un grand merci à mes proches, ma famille et mes amis. Un merci tout particulier est à accorder à ma maman qui a toujours été présente pour me soutenir dans mes différents projets et celui-ci en fait plus que partie. Merci aussi à Guillaume d'avoir fait preuve de patience et surtout pour m'avoir supporté tout au long de ma maîtrise.

Merci aussi à mes collègues et amis, Marion et Benjamin. Merci pour votre bonne humeur et vos conseils durant ces longues heures de rédaction à la bibliothèque. Ce mémoire n'aurait probablement pas été vécu de cette manière sans vous.

Je remercie, finalement, le Conseil de Recherche en Sciences Naturelles et Génie (CRSNG), Prompt et la Chaire UX pour leur contribution financière à cette étude.

Introduction

Mise en contexte

150... Selon le Wall Street Journal, le NPS (Net Promoter Score, Reichheld, 2003) aurait été mentionné 150 fois sur plusieurs milliers de conférences téléphoniques d'analystes financiers en 2018 (Safdar et Pacheco, 2019). Le sujet de ces conférences portait sur leurs résultats financiers globaux de 50 sociétés. Selon ces mêmes auteurs, ce nombre serait quatre fois plus important qu'en 2013. Pointilist, une entreprise développant un logiciel de parcours client, a effectué un sondage en 2019. Sur 700 professionnels œuvrant dans le domaine de l'expérience client, plus de la moitié des répondants considère le NPS comme étant l'une des métriques les plus importantes à leurs yeux (Pointilist, 2019). Celle-ci arrive bien devant des métriques plus traditionnelles comme le taux de conversion ou encore la valeur vie client. Son format est particulier, une seule et unique question : "*Quelle est la probabilité que vous recommandiez [nom de la compagnie] à un.e ami.e / collègue / membre de famille?*". En psychométrie, on appelle les mesures de ce type : des échelles à un item. D'un point de vue pratique, elles possèdent de nombreux avantages. On leur reconnaît notamment une rapidité d'exécution en matière de collecte de données (Straub, Boudreau et Gefen, 2004). Leur format leur donne également un gros avantage : le fait qu'elles ne soient pas contaminées par d'autres items, en comparaison aux échelles plus longues (Nagy, 2002). Elles restent néanmoins plus que déconseillées par le milieu scientifique de par leurs faiblesses psychométriques : un portrait incomplet de toute l'étendue d'un construit (Baumgartner et Homburg, 1996) ou encore une fiabilité très difficile, voire impossible à calculer (Churchill, 1979). Il n'existe, à notre connaissance, aucun réel consensus quant à l'utilisation systématique ou non de ce type d'échelles : sur le terrain, les praticiens semblent s'accorder pour s'en servir régulièrement alors que la science recommande le contraire. Certains auteurs semblent les déconseiller dans tous les cas, d'autres ont une opinion inverse et tendent à les recommander. Pour d'autres encore, cela dépend du contexte ou du construit d'intérêt.

L'expérience utilisateur (UX) se définit comme étant un ensemble de perceptions et de réactions qui résultent d'une interaction entre une personne et un système, un produit ou un service (Organisation internationale de normalisation et Commission électrotechnique, 2016). Il s'agit ici d'analyser et de comprendre les multiples réactions émotionnelles, les comportements lors d'une interaction. Traditionnellement, ces réactions et comportements se mesurent de deux manières différentes, mais pour certains (Li, Walters, Packer et Scott, 2018; Tomarken, 1995) complémentaires (l'une permettant de comprendre et jusqu'à un certain point expliquer ce que l'autre ne comprend pas) : les méthodes de type implicites (psychophysiologiques) et les méthodes explicites (entrevues et questionnaires).

Les méthodes implicites font appel aux réactions naturelles produites par notre corps lorsque nous sommes face à un stimulus. Ces réactions peuvent se produire automatiquement, de manière incontrôlée (par exemple, l'augmentation ou la diminution du rythme cardiaque), ou contrôlée (mouvement des yeux). Elles ont pour avantage notamment de mesurer une expérience en temps réel, sans avoir besoin d'interrompre le flux d'une interaction (Ganglbauer, Schrammel, Deutsch et Tscheligi, 2011). Elles ont pour principal inconvénient de nécessiter un matériel spécifique afin d'être mises en place. Les méthodes explicites, de leur côté, ont l'avantage de permettre une certaine rétroaction de la part d'un utilisateur, une conscientisation de l'émotion ressentie lorsqu'il ou elle interagit avec une interface. Mais elles sont sujettes à plusieurs biais.

Objectifs de l'étude et questions de recherche

L'étude de la relation entre l'expérience vécue et l'expérience perçue n'est pas nouvelle. Citons par exemple les travaux de Lang, Greenwald, Bradley et Ham (1993) qui, après avoir fait visionner un certain nombre d'images à plusieurs personnes, ont confronté leurs réponses à leurs réactions psychophysiologiques. Ces auteurs avaient trouvé plusieurs liens entre l'émotion vécue, ainsi que son intensité, et l'émotion perçue (Lang, Greenwald, Bradley et Ham., 1993). D'autres études semblent aller dans le même sens (Ortiz de Guinea, Titah et Léger, 2013; Le Pailleur, Huang, Léger et Sénécal, 2020). Il est, dès lors, possible que ces deux méthodes corrèlent entre elles. Une première question en découle :

« *Dans quelle mesure l'expérience utilisateur vécue est-elle en relation avec l'expérience utilisateur perçue?* »

Demander à une personne de faire une rétroaction de l'expérience qu'elle vient de vivre fait appel à sa mémoire. Et la mémoire s'altère avec le temps. Elle forme généralement une courbe : nous nous souvenons mieux des éléments s'étant déroulés au début et à la fin d'une interaction; le milieu est en général plus facilement oublié (Murdock, 1962). Certaines recherches tendent à montrer une influence de ces effets sur la relation entre l'expérience vécue et l'expérience perçue. Pour plusieurs auteurs, il semblerait exister surtout des effets de première impression (DiGirolamo et Hintzman, 1997; Lindgaard, Fernandes, Dudek et Brown, 2006), pour d'autres, ce sont plutôt des effets de dernière impression (Hassenzahl et Sandweg, 2004; Bergeron, Fallu et Roy, 2008) et pour d'autres encore, ce sont à la fois des effets de première et de dernière impression, cela dépend du construit d'intérêt (Lourties, Léger, Sénécal, Fredette et Chen, 2018; Murphy, Hofacker et Mizerski, 2006). Dès lors, une deuxième question se pose :

« Dans quelle mesure la relation en l'expérience utilisateur vécue et l'expérience utilisateur perçue est-elle différente à différents moments d'une interaction? »

La force d'une mesure peut se traduire par sa validité, sa fiabilité, mais également par sa sensibilité. La sensibilité correspond au degré de captation des différences (Lewis, 2002). En d'autres termes, une mesure devrait être capable de distinguer différentes valeurs émanant tant des différentes manipulations que du système dont il est question (Lewis, 2002). Il semblerait que les éléments constituant un système, une interface, ont une influence sur l'expérience qui en découle. Plusieurs auteurs se sont intéressés à ce sujet et mettent en avant le fait que les caractéristiques propres à une interface jouent un rôle sur le comportement d'une personne lorsqu'elle interagit avec cette dernière (Cyr, Head, Larios et Pan, 2009). Par exemple, Cai & Xu (2011) ont remarqué que l'esthétique d'une interface augmente le plaisir résultant de l'interaction. Les mesures à un item devraient, de ce fait, être suffisamment fortes, suffisamment sensibles que pour détecter ces différences dans les interfaces. Une dernière question de recherche se pose :

« *Dans quelle mesure les échelles à un item sont-elles suffisamment sensibles pour distinguer l'expérience utilisateur à travers différents contextes d'utilisation, tels que des contextes hédoniques ou utilitaires?* »

Afin de répondre à nos trois questions de recherche, nous avons réalisé une étude en laboratoire avec 40 participants sur dix interfaces différentes. L'étude s'est déroulée entre décembre 2019 et janvier 2020 et a été approuvée par le comité d'éthique de HEC Montréal (CER). Les participants devaient effectuer trois tâches sur deux sites internet d'institutions financières canadiennes différentes. Ces tâches étaient variées et consistaient principalement à de la navigation et de la complétion de formulaires.

Nous avons évalué l'expérience vécue à l'aide de différents outils : la valence émotionnelle a été mesurée grâce à un logiciel d'analyse de reconnaissance des mouvements des muscles faciaux (FaceReader© version 6.0, Noldus, Wageningen, Pays-Bas). L'activation émotionnelle a été mesurée en utilisant deux méthodes différentes : l'activité électrodermale et l'activité cardiaque par le logiciel Acqknowledge© (Biopac, Goleta, USA). Enfin, la charge cognitive a été mesurée en analysant la dilatation moyenne de la pupille à l'aide du logiciel Tobii Studio© (Stockholm, Suède).

L'expérience perçue a été mesurée à l'aide de quatre mesures à un item. Le niveau de plaisir perçu et l'activation émotionnelle ont été mesurés grâce à l'Affective Slider (Betella et Verschure, 2016). L'effort perçu par Customer Effort Score (Dixon, Freeman et Toman, 2010). Nous avons mesuré le niveau de satisfaction en utilisant le Customer SATisfaction Score[1]. Et enfin, la rétention a été mesurée à l'aide du NPS (Reichheld, 2003).

[1] Cette mesure n'a pas réellement été créée par une personne en particulier. Son acronyme semble plutôt provenir de l'industrie. Plusieurs auteurs ont cependant mesuré la satisfaction de cette manière depuis de nombreuses années, parfois sur une échelle de Likert à 5, 7 ou encore 11 points.

Structure du mémoire

Ce mémoire par articles est structuré en deux parties : un chapitre de type revue de littérature et un article. Le premier chapitre aborde les fondements théoriques sur lesquels se base la psychométrie. Il pose les racines de notre recherche proposée dans l'article. Le chapitre suivant est un article à travers lequel nous proposons une exploration de la relation entre l'expérience vécue et l'expérience perçue lorsqu'elle est reportée à l'aide d'une échelle à un item.

Informations sur le chapitre 1

Le chapitre 1 est une revue de littérature sur les échelles à un item. Il pose les bases justifiant la recherche effectuée dans le chapitre suivant. Il est divisé de cette manière : une première partie sur la théorie propre à la psychométrie, une deuxième partie sur les différents avantages et inconvénients d'utiliser les échelles à un item et la troisième partie s'intéresse plus spécifiquement aux échelles à un item très utilisées en UX.

La première section propose de s'intéresser aux bases de la théorie émanant de la psychométrie. Elle est divisée en trois sous-sections, à commencer par la fiabilité des échelles de mesure. Nous y abordons les principales méthodes pour estimer la fiabilité d'une échelle : la cohérence interne, la bissection, la méthode test-retest et l'utilisation des formes alternatives. La deuxième section porte sur la validité. Comme pour la section précédente, plusieurs méthodes y sont mentionnées et expliquées afin de mieux les appréhender : la validité de contenu, la validité prédictive et la validité de construit. La dernière section explique quelles sont les principales méthodologies utilisées pour développer une échelle : la technique classique et une méthode plus contemporaine.

La deuxième partie de ce chapitre met en avant les principaux avantages et inconvénients d'utiliser des échelles à un item pour évaluer une interaction. Cette partie met en lumière le fait que les échelles de ce type sont rapides, directes et peu coûteuses, bien qu'elles souffrent de nombreuses faiblesses comme une fiabilité difficile voire impossible à calculer, une faible validité et un certain manque d'adaptation pour certains construits dont l'abstraction peut être forte.

La dernière partie de la revue de littérature se veut plus pratique et propose un tour d'horizon de plusieurs échelles utilisées en expérience utilisateur. Nous y apprenons premièrement quelle est leur utilisation globale dans l'industrie. Ensuite, nous proposons une explication des quatre mesures à un item utilisées dans le chapitre suivant. Enfin, nous proposons une sélection des principaux antécédents et conséquences des mesures précédemment citées.

Informations sur le chapitre 2

Le chapitre 2 est un article en préparation de soumission à la revue Computers in Human Behavior Reports (ISSN: 2451-9588). Il s'agit d'une revue s'intéressant aux interactions humaines avec les ordinateurs. Cet article tente de répondre aux questions de recherche susmentionnées : nous y étudions à quel point l'expérience vécue est-elle en relation avec l'expérience perçue, cela à travers différents moments d'une interaction ainsi qu'à travers plusieurs attributs propres à chaque système.

Dans cet article, nous observons que l'expérience perçue n'est pas systématiquement en relation avec l'expérience vécue. Les corrélations qui y ont été calculées se comportent négativement dans la majorité des cas. Certaines sont significatives, d'autres ne le sont pas. Cela dépendrait du construit d'intérêt, du moment de l'interaction ou d'éléments spécifiques aux interfaces utilisées. Cet article met en avant les limites desquelles les professionnels devraient tenir compte lorsqu'ils estiment avoir besoin de les utiliser.

Contributions et responsabilités personnelles

Tableau 1. Contributions de l'étudiant

Étape	Contributions de l'étudiant
Design expérimental	Le design expérimental a été effectué en amont par l'équipe du Tech3Lab. L'étudiant y a collaboré en amenant les mesures utilisées dans son mémoire : 75%
Stimuli	Les stimuli ont été discutés et mis en place par l'équipe du Tech3Lab et l'étudiant. L'étudiant était en charge d'en vérifier leur qualité ainsi que leur fonctionnement : 75%
Recrutement	L'étudiant s'est chargé lui-même du recrutement et a été assisté par l'équipe du Tech3Lab : 50%
Prétests et collectes	L'étudiant a été présent et a modéré tous les prétests et toutes les collectes de données : 100%
Extraction et transformation des données	L'étudiant s'est chargé d'extraire toutes les données. Plusieurs aires d'intérêts ont été codifiées par l'étudiant sur différentes pages utilisées par les participants. La triangulation des données a été effectuée par l'équipe du Tech3Lab : 100%
Analyses statistiques	Les analyses ont été préparées par l'étudiant, supporté par les statisticiens du laboratoire. Les tests statistiques ont été réalisés par l'étudiant. Les analyses ont été effectuées par l'étudiant, supporté par les deux statisticiens : 90%
Rédaction	L'étudiant était en charge de l'entièreté de la rédaction : 100%

Références

Baumgartner, Hans et Christian Homburg (1996). « Applications of structural equation modeling in marketing and consumer research: A review », *International Journal of Research in Marketing*, vol. 13, no 2, p. 139-161.

Bergeron, Jasmin, J. M. Fallu et Jasmin Roy (2008). « Une comparaison des effets de la première et de la dernière impression dans une rencontre de vente », *Recherche et Applications en Marketing*, vol. 23, no 2, p. 19-36.

Betella, Alberto et Paul F. M. J. Verschure (2016). « The affective slider: A digital self-assessment scale for the measurement of human emotions », *PLoS ONE*, vol. 11, no 2.

Cai, Shun et Yunjie Xu (2011). « Designing not just for pleasure: Effects of web site aesthetics on consumer shopping value », *International Journal of Electronic Commerce*, vol. 15, no 4, p. 159-188.

Churchill, Gilbert A. (1979). « A paradigm for developing better measures of marketing constructs », *Journal of Marketing Research*, vol. 16, no 1, p. 64-73.

Cyr, Dianne, Milena Head, Hector Larios et Bing Pan (2009). « Exploring human images in website design: A multi-method approach », *MIS Quarterly: Management Information Systems*.

DiGirolamo, Gregory J. et Douglas L. Hintzman (1997). « First impressions are lasting impressions: A primacy effect in memory for repetitions », *Psychonomic Bulletin and Review*, vol. 4, no 1, p. 121-124.

Dixon, Matthew, Karen Freeman et Nicolas Toman (2010). « Stop trying to delight your customers », *Harvard Business Review*, vol. 88, no 7-8.

Ganglbauer, Eva, Johann Schrammel, Stephanie Deutsch et Manfred Tscheligi (2011). « Applying psychophysiological methods for measuring user experience: Possibilities, challenges and feasibility », *Human-Computer Interaction. INTERACT 2011 (Lecture Notes in Computer Science)*.

Hassenzahl, Marc et Nina Sandweg (2004). « From mental effort to perceived usability: Transforming experiences into summary assessments », communication présentée au *CHI'04*, 2004, Vienna, Austria.

Lang, Peter J., Mark K. Greenwald, Margaret M. Bradley et Alfons O. Hamm (1993). « Looking at pictures: Affective, facial, visceral, and behavioral reactions », *Psychophysiology*, vol. 30, no 3, p. 261-273.

Le Pailleur, Félix, Bo Huang, Pierre Majorique Léger et Sylvain Sénécal (2020). « A new approach to measure user experience with voice-controlled intelligent assistants: A pilot study », communication présentée au *HCII 2020*, Copenhagen, Denmark.

Lewis, James R. (2002). « Psychometric evaluation of the pssuq using data from five years of usability studies », *International Journal of Human-Computer Interaction*, vol. 14, no 3-4, p. 463-488.

Li, Shanshi, Gabby Walters, Jan Packer et Noel Scott (2018). « A comparative analysis of self-report and psychophysiological measures of emotion in the context of tourism advertising », *Journal of Travel Research*, vol. 57, no 8, p. 1078-1092.

Lindgaard, Gitte, Gary Fernandes, Cathy Dudek et J. Browñ (2006). « Attention web designers: You have 50 milliseconds to make a good first impression! », *Behaviour and Information Technology*, vol. 25, no 2, p. 115-126.

Lourties, Sébastien, Pierre Majorique Léger, Sylvain Sénécal, Marc Fredette et Shang Lin Chen (2018). « Testing the convergent validity of continuous self-perceived measurement systems: An exploratory study », communication présentée au *HCII 2018*, Las Vegas, USA.

Murdock, Bennet B. (1962). « The serial position effect of free recall », *Journal of Experimental Psychology*, vol. 64, no 5, p. 482-488.

Murphy, Jamie, Charles Hofacker et Richard Mizerski (2006). « Primacy and recency effects on clicking behavior », *Journal of Computer-Mediated Communication*, vol. 11, no 2, p. 522-535.

Nagy, Mark S. (2002). « Using a single-item approach to measure facet job satisfaction », *Journal of Occupational and Organizational Psychology*, vol. 75, no 1, p. 77-86.

Organisation internationale de normalisation et internationale Commission électrotechnique (2016). *Systems and software engineering : Systems and software quality requirements and evaluation (square) : Measurement of system and software product quality = ingénierie des systèmes et du logiciel : Exigences de qualité et évaluation des systèmes et*

du logiciel (square) : Mesurage de la qualité du produit logiciel et du système, ISO/IEC, (1st ed.), c. viii, 45 p.

Ortiz de Guinea, Ana, Ryad Titah et Pierre Majorique Léger (2013). « Measure for measure: A two study multi-trait multi-method investigation of construct validity in is research », *Computers in Human Behavior*, vol. 29, no 3, p. 833-844.

Pointilist (2019). *State of customer journey management & cx measurement*, 31-31 p.

Reichheld, Frederick F. (2003). « The one number you need to grow », *Harvard Business Review*, vol. 81, p. 46-54+124.

Safdar, By Khadeeja et Inti Pacheco (2019). « The dubious management fad sweeping corporate america », *Wall Street Journal*, p. 1-10.

Straub, Detmar, Marie-Claude Boudreau et David Gefen (2004). « Validation guidelines for is positivist research », *Communications of the Association for Information Systems*, vol. 13, p. 380-427.

Tomarken, Andrew J. (1995). « A psychometric perspective on psychophysiological measures », *Psychological Assessment*, vol. 7, no 3, p. 387-395.

Chapitre 1
Revue de littérature

1.1 Introduction

L'expérience utilisateur est définie comme étant les « *perceptions et réponses d'une personne résultant de l'utilisation et/ou de l'utilisation prévue d'un système interactif...* » (ISO, 2016 – traduction libre). Cette définition possède une dimension importante : les perceptions. Ces dernières sont capitales dans l'utilisation d'une interface et nécessitent que l'on s'y intéresse. Elles sont en général évaluées à l'aide de diverses échelles, chacune représentant une ou plusieurs facettes de ces perceptions. Les techniques employées proviennent principalement de la psychologie, et la science qui permet notamment de les développer ainsi que d'en déterminer leur validité est la psychométrie.

La validation des instruments est un point très sensible, particulièrement discuté et bien documenté à travers la littérature. En 1989, Straub affirmait que la validation des construits n'était pas correctement effectuée dans le domaine des technologies de l'information (Straub, 1989). Quinze ans plus tard, en 2004, ce même auteur constatait une certaine amélioration, mais avançait que cette faiblesse était toujours d'actualité (Straub, Boudreau, Gefen, 2004). Plus récemment, en 2009, une étude s'est intéressée à la validation des mesures perceptuelles. Le constat est plus positif : à travers une analyse de différentes études publiées dans le Journal of the American Society for Information Science and Technology entre 1982 et 2007, Kim (2009) constate une certaine amélioration. La validation des échelles de mesure perceptuelles se fait de plus en plus présente et naturelle, même si du travail reste encore à effectuer (Kim, 2009).

Vieille de plus de cent ans, la validation prendrait ses racines au début du XXe siècle (André, Loye et Laurencelle, 2016). Ces derniers citent notamment les travaux de Binet en affirmant que, sans citer le terme de « validité », l'auteur se questionnait déjà sur la pertinence des échelles utilisées en psychologie. La théorie classique propose d'emprunter deux chemins, deux directions se distinguant par leurs différentes formes,

mais qui pourtant sont plus que complémentaires : la fiabilité et la validité. Ces deux notions sont discutées à travers le point suivant.

1.2 Validation et méthodes de construction d'échelles de mesure

Avant de comprendre la manière dont les mesures des construits que nous utilisons sont développées et validées, il apparaît important de comprendre ce qu'est un construit. Il consiste en une idée abstraite qui, dans le cas d'une enquête, est mesurée à l'aide d'une ou plusieurs questions (Dew, 2008). C'est un concept, une abstraction formée d'élément(s) précis qui la définissent (Kerlinger et Lee, 2000). Il possède cependant une dimension plus particulière puisqu'il est intentionnellement et expressément inventé, créé de toutes pièces, pour un usage scientifique très précis (Kerlinger et Lee, 2000). C'est, de ce fait, une idée qui n'existe pas de manière naturelle, mais qui a été concrétisée dans le but de pouvoir l'étudier. Cette définition implique qu'un construit n'est pas observable directement. L'étude scientifique de celui-ci implique, en particulier, une certaine opérationnalisation (Riedl, Davis et Hevner, 2014) : il s'agit ici d'une manière de le mesurer pouvant faire référence à plusieurs niveaux analytiques. La concrétisation d'un construit est donc réalisable grâce à l'utilisation d'une mesure (Riedl, Davis et Hevner, 2014). Si son étude implique l'utilisation de certaines mesures, sa définition en implique plutôt la prise en compte de ses différentes dimensions (Dew, 2008). À titre d'exemple, un construit très largement étudié à travers la littérature est la satisfaction au travail. Churchill, Ford et Walker (1974) ont identifié sept dimensions permettant de définir ce dernier, et plus spécifiquement au sujet de la satisfaction au travail des vendeurs industriels : les tâches effectuées, la relation avec les collègues, le management, la politique de l'entreprise et le support offert, le salaire, les promotions et les opportunités d'avancement ainsi que les clients. Ces sept dimensions sont interreliées et permettent, lorsqu'elles sont mises en commun de définir ce construit particulier. Il apparaît donc évident qu'un construit n'existe pas seul; l'importance de tenir compte des différents éléments et de ses dimensions (qui elles-mêmes possèdent leurs propres dimensions) est claire. Les construits sont donc présents dans des schémas théoriques et sont liés, jusqu'à un certain point, à d'autres construits (Kerlinger et Lee, 2000).

Il est possible que certains construits soient influencés par des entités sous-jacentes. Ces entités sont considérées comme latentes et, dès lors, non directement observables. Il n'existe pas réellement de manière de les observer, elles se situent en quelque sorte, « sous » le construit mesuré (Kerlinger et Lee, 2000). Elles existent, mais ne sont pas capturables directement. Selon Straub, Boudreau et Gefen (2004), il n'y a aucune mesure immédiate ou évidente pour capturer l'essence d'un construit. Selon eux, cette latence existe bel et bien et il faudrait en tenir compte lorsque vient le temps d'étudier un construit. Borsboom, Mellenbergh et Van Heerden (2003) suggèrent une approche tantôt mathématique, en les considérant comme résultant d'une simple régression; tantôt philosophique et soutiennent par la même occasion que la science devrait s'intéresser à la définition même d'une variable latente qui, selon eux, n'est pas propre à toute une population, mais bien résultante des différences émanant de chaque individu.

1.2.1. Fiabilité des mesures

« *Les personnes fiables (…) sont celles dont le comportement est cohérent, fiable, prévisible (…). Les personnes peu fiables, en revanche, sont celles dont le comportement est beaucoup plus variable* » (Kerlinger et Lee, 2000, p. 642 – traduction libre). C'est à travers cette image que la définition de la fiabilité prend tout son sens. Elle consiste à évaluer la force d'une mesure perceptuelle. Cronbach (1951) affirme que la fiabilité est une question de précision. Plus concrètement, estimer la fiabilité d'une échelle consiste à en évaluer sa stabilité à travers le temps; on parle ici d'une certaine répétabilité (Kerlinger et Lee, 2000 ; Nunnally et Bernstein, 1994) : si un chercheur décide d'effectuer plusieurs enquêtes différentes sur le même sujet en utilisant la même mesure, les résultats obtenus devraient être équivalents. Déterminer la fiabilité consiste également à tenter de connaître le degré d'erreur d'une mesure (Kerlinger et Lee, 2000). C'est aussi l'une des définitions proposées par Nunnally et Bernstein en 1994 : la fiabilité, c'est l'absence d'erreurs aléatoires. Les auteurs ajoutent également : « *Cette définition implique l'homogénéité du contenu des tests contenant plusieurs items et une cohérence interne ou des corrélations élevées entre les composantes de la mesure globale…* » (Nunnally et Bernstein, 1994, p.213-214 – traduction libre).

De manière plus concrète, estimer la fiabilité d'une mesure consiste à en calculer le degré de variabilité des réponses. Habituellement, les corrélations à travers les différents items d'une échelle sont estimées afin d'en déterminer leur fiabilité. L'un des indicateurs les plus utilisés est le coefficient alpha de Cronbach, parfois à tort (Hogan, Benjamin et Brezinski, 2000). Ce coefficient dans sa globalité permet de calculer les corrélations à travers les différents items d'une échelle (Straub, 1989). Il varie habituellement entre 0 et 1 et, même s'il n'existe pas de réel consensus au sujet du coefficient moyen désiré, la littérature, se basant très largement sur les travaux de Nunnally et Bernstein (1994), s'accorde pour dire qu'un coefficient supérieur à 0.7 suggère qu'une mesure est fiable. Cependant, plus le nombre d'items propres à un questionnaire augmente, plus ce coefficient sera élevé. Cela implique donc, paradoxalement, qu'un coefficient supérieur à 0.9 ne paraît pas être un bon indicateur de la fiabilité. En effet, un coefficient très élevé indiquerait une certaine redondance entre les différents items d'une mesure : plusieurs items mesurant le même aspect (Streiner, 2003). Même si le coefficient moyen conseillé devrait s'élever au minimum à 0,70, cette affirmation n'est pas valable dans chaque cas. Par exemple, Nunnally et Bernstein (1994) indiquent que ce coefficient devrait être plus élevé (0.80) si une étude cherche à comparer les différences entre les individus. Cette règle n'est, de ce fait, pas à suivre de manière stricte, mais bien à utiliser en connaissance de cause et en ayant conscience des objectifs de l'étude.

Plusieurs approches sont possibles pour estimer la fiabilité d'une mesure. Citons notamment les travaux de Kerlinger et Lee (2000), Straub, Boudreau et Gefen (2004) ainsi que Nunnally et Bernstein (1994) qui les répertorient. Ces méthodes sont généralement : la cohérence (ou consistance) interne, la bissection (« split-half »), test-retest, l'utilisation des formes alternatives ou équivalentes et la fiabilité unidimensionnelle.

La cohérence interne est probablement l'approche la plus utilisée quand il s'agit de vérifier la fiabilité d'une mesure. Il s'agit ici de la « *mesure dans laquelle la performance d'un item dans un instrument constitue un bon indicateur de la performance de tout autre item du même instrumen*t » (DeVon, Block, Moyle-Wright, Ernst, Hayden, Lazzara, Savoy et Kostas-Polston, 2007, p. 162 – traduction libre). Cette méthode consiste

à calculer les corrélations moyennes entre les différents items constituant un instrument de mesure. Le coefficient généralement utilisé est l'alpha de Cronbach, décrit plus haut. Ce coefficient implique une certaine cohérence entre tous les items : chacun doit être évalué sur le même type d'échelle (Straub, Boudreau et Gefen, 2004). L'un des avantages dans le fait d'utiliser cette méthode est qu'il ne nécessite qu'une seule passation de la mesure pour pouvoir le calculer (Devon et al., 2007). La méthode de bissection (ou split-half en anglais) en est une déclinaison. Il s'agit, également, de calculer la corrélation moyenne entre chaque item d'un même instrument de mesure. La différence réside ici dans le fait de séparer l'instrument en deux parts égales et de vérifier à quel point les items corrèlent entre eux (Nunnally et Bernstein, 1994). Un problème de taille réside pourtant dans cette approche : les résultats obtenus sont influencés par la manière dont l'échelle est scindée (Peter, 1979). L'auteur propose, pour remédier à cet inconvénient, de déterminer la fiabilité moyenne pour toutes les manières possibles de séparer la mesure; une solution paraissant plus que fastidieuse. Une autre solution possible est de séparer la mesure de manière aléatoire (Imbault, Shore et Kuperman, 2018). L'approche utilisant la technique « test-retest » consiste à faire passer deux fois le même test à la même personne, cela à des intervalles de temps différents. La corrélation entre les deux passations est ensuite calculée. L'inconvénient de cette approche est que les répondants risquent de se souvenir des réponses précédemment indiquées et de noter une seconde fois la même réponse (Cook, Campbell et Day, 1979). L'utilisation des formes alternatives ou équivalentes semble plutôt rare (Straub, Boudreau et Gefen, 2004). L'idée derrière cette approche est d'administrer deux mesures différentes, mais très proches l'une de l'autre, à deux intervalles de temps pour évaluer le même construit, et d'en comparer les résultats (DeVon et al., 2007). Cette méthode possède sensiblement les mêmes inconvénients que la méthode test-retest : le fait que les répondants puissent se souvenir de leurs réponses passées. Enfin, la dernière approche est la fiabilité unidimensionnelle. Cette forme est plus complexe et prend tout son sens dans la considération des construits latents. En effet, il s'agit ici de mesurer à quel point chaque dimension d'un questionnaire est capable de mesurer chaque construit latent (Straub, Boudreau et Gefen, 2004).

1.2.2. Validité des mesures

Si la fiabilité permet de savoir si une échelle mesure quelque chose, rien n'indique que cette dernière mesure réellement le construit désiré. La fiabilité est nécessaire, mais insuffisante (Cook et Beckman, 2006). Si un chercheur souhaite savoir si le questionnaire utilisé mesure correctement le construit d'intérêt, il doit faire appel à la validité. La validité est une notion large. Elle revêt plusieurs définitions différentes, plusieurs dimensions différentes. La littérature s'accorde globalement pour affirmer qu'il existe trois différents types de validité. La validité de contenu, la validité prédictive et la validité de construit.

La validité de contenu est basée sur l'idée que chaque propriété psychologique possède son propre univers théorique, lui-même étant découpé en plusieurs éléments théoriques, appelés items (Kerlinger et Lee, 2000). Derrière cela, la validité de contenu prend son sens dans le fait de créer une mesure la plus représentative possible de tout l'univers disponible d'un construit (Riedl, Davis et Hevner, 2014). C'est à travers cette définition que réside le point clé de la validité de contenu : connaître l'étendue d'un construit. Il est d'ailleurs presque impossible de déterminer avec précision chaque dimension d'un construit mesuré, puisque « *l'univers des items est lui-même indéterminé* » (Straub, Boudreau et Geffen, 2004, p. 387 - traduction libre). Il apparaît très difficile de déterminer l'étendue réelle des attributs propres aux comportements, aux croyances ou aux attitudes (DeVellis, 2003). Pourtant, Haynes, Richard et Kubany (1995), proposent un guide qui se veut à la fois qualitatif et quantitatif, composé de 7 points clés pour tenter d'estimer au mieux la validité de contenu. Une importante part de subjectivité y subsiste pourtant. Cette validité fait, en effet, appel à l'avis de juges et d'experts du domaine d'intérêt, pour lesquels il est demandé de noter chaque item selon sa pertinence relative au construit mesuré (Straub, Boudreau et Gefen, 2004). Cette subjectivité est un point récurrent dans la littérature et soulevé notamment par Kerlinger et Lee (2000) qui indiquent que chaque item est jugé pour sa pertinence présumée à la propriété mesurée. La méthode couramment utilisée gravite autour de deux points principaux : la revue de littérature (selon les théories développées propres à chaque comportement étudié) et l'analyse d'experts. C'est une méthode peu évidente, quoique facilitée grâce à un bon

échantillonnage (DeVellis, 2003), mais qui revêt une importance capitale. Importance soulignée par Haynes, Richard et Kubany (1995) qui mettent en garde : « *Les données obtenues à partir d'un instrument non valide peuvent surreprésenter, omettre ou sous-représenter plusieurs facettes d'un construit et refléter des variables hors du domaine du construit.* », p. 240 (traduction libre).

Le deuxième type de validité est directement quantitatif. Il s'agit de la validité prédictive. Cette approche est purement statistique et tente d'établir la relation entre les mesures et les construits en démontrant une corrélation ou une certaine force de prédiction d'une variable (Straub, Boudreau, Gefen, 2004). DeVellis (2003), précise également que cette validité ne requiert pas nécessairement de relation de causalité entre plusieurs variables. Ce n'est pas, d'après l'auteur, la relation qui importe, mais bien la force de cette relation.

Le troisième type de validité est la validité de construit. Cette dernière est particulièrement importante puisqu'elle permet de savoir si les mesures choisies s'accordent ensemble pour mesurer l'essence du construit étudié (Straub, Boudreau, Gefen, 2004). Cette forme possède plusieurs indicateurs différents. Les principaux sont la validation discriminante et son opposée : la validation convergente, la validité factorielle et la validité nomologique (Straub, Boudreau et Gefen, 2004). La validité discriminante teste la liaison entre la mesure en question et des items qui n'ont pas de lien théorique. En d'autres termes, la validité discriminante signifie que l'on peut différencier le construit d'autres construits qui peuvent être similaires (Kerlinger et Lee, 2000). La validité convergente consiste à connaître à quel point différentes mesures d'un même construit convergent (DeVon et al., 2007). Si les formes de validité convergente et discriminante prennent leur sens lorsque plusieurs méthodes sont utilisées pour mesurer un même construit, la validité factorielle, elle, semble plus adaptée lorsqu'une seule méthode a été utilisée (Straub, Boudreau et Gefen, 2004). Cette dernière s'intéresse à la mesure dans laquelle les dimensions théoriques d'un construit sont englobées dans une échelle (Lewis, Templeton et Byrd, 2005). Les méthodes classiques d'analyse factorielle y sont ici utilisées : notamment les modèles d'équations structurelles. Nous avons précédemment (cf. 1.2.) discuté des construits considérés comme latents et avons suggéré qu'il n'existait

aucune manière directe de les observer. C'est à travers la validité factorielle que leur considération prendrait son sens (Gefen et Straub, 2005). Enfin, la validité nomologique fait appel à d'autres échelles validées au préalable. Concrètement, il s'agit ici de vérifier à quel point la nouvelle échelle corrèle avec une autre, qui elle-même mesure d'autres construits théoriquement représentatifs de notre construit d'intérêt (MacKenzie, Podsakoff et Podsakoff, 2011).

1.2.3. Méthodes de développement d'échelles de mesure

Il n'existe pas de réel consensus quant à une vraie méthode officielle à utiliser pour développer une échelle de mesure. Nous pouvons aisément dire que deux écoles ont tendance à s'affronter : les partisans des mesures à plusieurs items et les partisans des mesures à un item. Lors d'un article publié en 1979, Churchill mettait en avant un certain manque de méthode lorsqu'il s'agit de développer un instrument pour mesurer les attitudes. Selon lui, il manque un cadre formel permettant de baliser les éléments nécessaires à la construction des échelles de mesure. Selon ses dires, il existe un retard de vingt ans entre le domaine de la psychologie et celui du marketing : « …*la littérature psychologique est dispersée. Les notions sont disponibles en de nombreux morceaux dans des sources variées. Il n'existe pas de cadre général que le spécialiste du marketing puisse adopter pour aider à organiser les nombreuses définitions et mesures de fiabilité et de validité en un tout intégré, de sorte qu'il soit évident de savoir quoi utiliser et quand le faire.* » (Churchill, 1979, p. 65 (traduction libre). À travers cet article, Churchill posait les bases de ce qui était une méthodologie pour développer les échelles de mesure. Plusieurs auteurs ont adopté la même procédure, à la lettre ou quelque peu modifiée (Moore et Benbasat, 1991 ; Lewis, Templeton et Byrd, 2005 ; DeVellis, 2003). Méthodologie qui, à l'heure actuelle, est encore très largement utilisée. À noter que cette méthode, comme l'indique l'auteur, est valable uniquement pour les échelles contenant plusieurs items. Une procédure plus récente s'intéressant au développement d'échelles à un item existe (Rossiter, 2002) et sera développée plus tard dans cette partie.

À travers la littérature, trois étapes majeures sont à prendre en compte pour développer un instrument de mesure : spécifier le domaine du construit, construire l'instrument et tester l'instrument. Ces phases ne sont pas définies de la même manière par toute la communauté scientifique et il existe des sous-étapes que certains auteurs ont décidé d'ailleurs de regrouper.

La première phase dans le processus de développement d'un instrument de mesure est commune et consiste à spécifier le domaine du construit mesuré. Cette étape, purement théorique, est primordiale puisqu'elle permet de délimiter avec le plus de précision possible le domaine attenant au construit d'intérêt (Churchill, 1979). La procédure se veut plutôt qualitative et se base en grande majorité sur la revue de littérature et les différentes définitions qui y sont proposées. Il est également possible de faire appel à un panel d'experts du domaine ou encore utiliser la méthode du groupe de discussion (Lewis, Templeton et Byrd, 2005), qui consiste à organiser une ou plusieurs séances en groupe composé d'experts proposant plusieurs idées. Il apparaît important, à travers cette première phase, de catégoriser chaque item provenant des différents instruments (Moore et Benbasat, 1991). Cette étape amène à trois niveaux d'information (Lewis, Templeton et Byrd, 2005) : les prémisses, spécifiant ce que cherche à mesurer le construit, la définition conceptuelle, donnant les clés pour définir le construit dans son ensemble et une liste de dimensions propres au construit d'intérêt. Cette étape permet donc de générer un premier échantillon d'items contenus dans l'instrument et se rapproche fortement de la validité de contenu décrite plus haut (Moore et Benbasat, 1991). En effet, l'utilisation d'experts du domaine mesuré ainsi que les objectifs de cette mesure permettent d'obtenir une première idée sur l'ensemble des items représentant le construit d'intérêt.

Les fondations de l'instrument étant posées, la deuxième phase se veut plus pratique et s'intéresse à la construction de l'instrument en lui-même. Il s'agit ici de créer l'instrument, ses dimensions et ses items sur base des définitions obtenues au préalable. Il est également important de déterminer le format des questions. DeVellis (2003) en répertorie plusieurs et propose, probablement un des plus utilisés : l'échelle de Likert. L'instrument (ainsi que sa forme) étant créé, la littérature suggère de le prétester afin d'en connaître ses forces et faiblesses. Habituellement, les participants sont constitués

d'experts (Moore et Benbasat, 1991) et/ou d'autres volontaires (Lin, Gregor et Ewing, 2008). Kim (2009) suggère également un test pilote, qui se veut la suite du prétest sur un panel plus large. Les analyses découlant de ce(s) premier(s) test(s) permettent d'évaluer la validité de construit (Moore et Benbasat, 1991), notamment le calcul de la convergence et de la divergence. Pour certains auteurs, cette étape se nomme le développement de l'échelle (Moore et Benbasat, 1991 ; Lewis, Templeton et Byrd, 2005)) ; pour d'autres elle se nomme l'analyse exploratoire (Lin, Gregor et Ewing, 2008).

La dernière phase de la construction d'un instrument de mesure est plutôt statistique et cherche à tester l'instrument en question sur le terrain. Il s'agit ici de confirmer la validité globale ainsi que la fiabilité de l'échelle (Lin, Gregor et Ewing, 2008). Les méthodes sont sensiblement les mêmes que celles décrites plus haut, à l'exception qu'elles sont appliquées sur un panel plus important. Le but étant, également, d'obtenir la mesure la plus représentative possible du construit étudié. Si Lin, Gregor et Ewing (2008), suggèrent ici une analyse confirmatoire uniquement, Lewis, Templeton et Byrd (2005) voient plutôt cette dernière phase comme étant à la fois exploratoire, mais aussi confirmatoire.

Un élément souvent mis en avant par Churchill (1979) dans son article proposant une procédure pour développer un instrument de mesure est l'itération. Pour l'auteur, la fiabilité doit se calculer très tôt, après la première collecte de données (prétest). S'il s'avère que la mesure ne possède pas des indices de fiabilité suffisants, il est important de revenir en arrière et de reprendre les fondations.

En parallèle à la théorie classique proposée par les auteurs ci-dessus, il existe d'autres méthodes pour développer une échelle de mesure. Notamment la procédure C-OAR-SE (Rossiter, 2002). Le fondateur de cette procédure s'est basé sur la littérature classique en psychométrie et a décidé de l'adapter plus particulièrement au domaine du marketing. À travers cette nouvelle procédure, Rossiter (2002) estime que l'élément clé dans la création d'une échelle est la validité de contenu. Selon lui, si la validité de contenu est effectuée correctement et en profondeur, le reste n'est pas nécessairement utile. Tout repose donc sur les deux premières phases développées par les auteurs précédents. Il

affirme que l'évaluation effectuée par les experts du domaine d'intérêt ainsi que celle effectuée par les cibles du questionnaire suffirait à estimer la validité de contenu. Pour lui, les autres formes de validité ainsi que la fiabilité n'apportent que peu de preuves de la qualité d'un instrument de mesure. Également, l'argument selon lequel une échelle à un item ne devrait pas être utilisée peu importe les circonstances n'est pas valable. L'auteur, dans sa procédure, propose notamment un cadre permettant de définir avec une certaine précision le construit d'intérêt, en le classant selon deux termes principaux : concret ou abstrait. Si le construit est considéré comme étant concret, il pourrait se mesurer à l'aide d'une échelle à un item; s'il est abstrait, l'auteur recommande les échelles classiques.

1.3 Mesurer les perceptions à l'aide d'échelles à un item : avantages et inconvénients

1.3.1. Avantages

Le premier avantage est logique et découle du format de ces échelles. Puisqu'une question permet d'analyser un construit en particulier, il apparaît évident que les questionnaires ne contenant qu'un item sont rapides à utiliser pour les répondants. Cette rapidité a notamment été mise en avant par certains théoriciens classiques en psychométrie et plus particulièrement Straub, Boudreau et Gefen (2004), qui affirment une réduction de la charge mentale impliquée. Cette charge est, selon Drolet et Morrison (2001), plus importante dans les mesures contenant plusieurs items. Ces derniers auteurs avancent que plusieurs items peuvent mener à une forme de fatigue, à de l'inattention et à de l'ennui. Ces états émotionnels peuvent mener à un comportement de réponse inapproprié. Les auteurs appuient leurs propos en donnant l'exemple d'une étude réalisée au préalable à travers laquelle un questionnaire à plusieurs items mesurés à l'aide d'une échelle de Likert était administré à un panel de personnes. Une partie des participants à cette étude avaient rempli le questionnaire correctement (en choisissant plusieurs réponses différentes). L'autre avait choisi la même réponse pour chaque question. Les auteurs en ont conclu que, pour ces participants, un certain ennui était présent et que ces derniers avaient dès lors adopté un comportement de réponse inapproprié. Faire le choix entre plusieurs formats de questionnaires représenterait le

compromis entre la richesse de la quantité d'informations différentes obtenues à l'aide des échelles à 1-item et la fiabilité obtenue grâce aux questionnaires classiques (Drolet et Morrison, 2001).

Le deuxième avantage découlant de l'utilisation de ces mesures est fortement lié au premier. Si un construit est représenté par une seule et unique question, il est évident que la redondance, et donc, la contamination des autres critères soit très fortement réduite, voire nulle. Les échelles contenant plusieurs items souffriraient d'un certain manque de validité apparente (Wanous, Reichers et Hudy., 1997). Plus précisément, les personnes à qui nous administrons un questionnaire peuvent avoir l'impression d'une répétition des autres éléments dudit questionnaire. Il est vrai que, afin de représenter le plus fidèlement possible l'étendue d'un construit, les chercheurs ont tendance à indiquer plusieurs items paraissant très similaires, cela afin d'observer le plus de variabilité possible entre chaque élément d'un questionnaire. On peut simplement avoir l'impression de répondre plusieurs fois à la même question de manière détournée et ces items finissent par se contaminer l'un et l'autre (Wanous, Reichers et Hudy, 1997). Également, la théorie classique permettant de calculer la fiabilité d'une mesure est très largement basée sur l'alpha de Cronbach. L'une des règles est que, plus on augmente le nombre d'items d'un questionnaire, plus ce coefficient sera élevé. Boyle (1991) en soulève l'un des risques inhérents : cette théorie est fragile, l'augmentation du nombre d'items peut indiquer une certaine redondance. Cette redondance a un impact négatif sur le comportement de réponse adopté par un répondant.

Le troisième avantage est purement économique. De manière globale, créer une échelle de mesure est un processus long et fastidieux. Il est important de respecter un certain nombre d'étapes au risque de créer une mesure inutile, ne représentant aucun construit et instable. Ces étapes sont parsemées de plusieurs tests et analyses différents, étalés sur quelques mois. De manière très logique, cet étalement a un impact financier. L'impact économique lié à l'utilisation des mesures à un item réside également dans le côté pratique : plus le questionnaire est court, plus les coûts associés à sa passation sont réduits (Wanous, Reichers et Hudy, 1997). Cet argument est aussi partagé par Bergkvist et Rossiter (2007) qui ajoutent que les coûts liés à l'analyse des données provenant d'un

questionnaire de ce format sont réduits. Cet avantage financier compenserait, selon Sarstedt et Wilczynski (2009), les inconvénients théoriques associés à l'utilisation des échelles à un item.

1.3.2. Inconvénients

Bien qu'utiliser une échelle à un item puisse être avantageux, les inconvénients associés à leur utilisation sont pourtant bien présents. Ils sont en grande majorité conceptuels. Le premier inconvénient réside dans la représentativité d'un construit à travers une mesure. Un argument assez récurrent est qu'une mesure à un item ne représenterait pas l'entièreté d'un construit. Elle ne contiendrait que peu d'informations au final. Plus le nombre d'items est important dans une mesure, plus les chances de représenter l'ensemble des facettes définissant un construit augmentent (Baumgartner et Homburg, 1996). En fait, de manière très logique et évidente, les échelles contenant plusieurs items permettent de capter plus d'informations, avec plus de précision que celles n'en contenant qu'un (Bergkvist et Rossiter, 2007). Le même argument revient également chez Jacoby (1978) qui insiste sur le fait que « *Comme c'est le cas pour des concepts comme la personnalité et l'intelligence, la plupart de nos concepts fondamentaux (…) sont multiples et complexes.* » (Jacoby, 1978, p. 93 – traduction libre). Il apparaît dès lors très évident d'utiliser plusieurs items pour étudier chaque dimension, chaque facette propre à un construit.

Le deuxième inconvénient indique que l'utilisation d'une échelle à un item n'est pas adaptée lorsque le construit mesuré est de type « abstrait » ou « complexe ». Rossiter (2002), lorsqu'il a mis en avant une nouvelle procédure pour créer une échelle de mesure proposait de classer les construits selon leur complexité. Deux grands types de construits peuvent ainsi émerger : ceux considérés comme concrets et ceux considérés comme étant abstraits. Selon l'auteur, les construits dits « abstraits » ne devraient pas être étudiés à l'aide d'une échelle à 1 item. Cet argument a également été avancé quelques années plus tôt par Wanous, Reichers et Hudy (1997) qui, de leur côté, proposent plutôt trois degrés de complexité : des construits très simples (ex. : les attentes) aux construits très abstraits (ex. : la personnalité), avec une catégorie au centre.

Le troisième inconvénient réside dans l'essence de la psychométrie. Selon la théorie classique, il est plus que primordial de respecter la validité et la fiabilité d'une mesure. Ces deux derniers éléments sont habituellement calculés grâce aux méthodes indiquées plus haut dans ce chapitre. L'erreur de mesure associée à leur utilisation peut être très importante (Churchill, 1979). Si nous administrons successivement le même questionnaire à la même personne (méthode appelée « test-retest »), à des intervalles de temps différents, il serait très peu probable que la position des réponses sur l'échelle soit identique (Churchill, 1979). Il s'agit ici d'un problème de fiabilité, comme expliqué au point 1.2.1. Il est d'ailleurs d'usage de calculer la corrélation entre les différents items pour en estimer la fiabilité. Cette méthode apparaît très difficile, voire impossible, lorsqu'il s'agit de l'évaluer pour les échelles à un item, si l'on respecte totalement la théorie classique (Churchill, 1979). Au sujet de la validité, plusieurs auteurs ont tenté de l'évaluer pour les échelles à un item, en suivant la procédure développée par Wanous et Reichers (1996). Il s'agit ici d'une procédure qui s'intéresse à calculer la corrélation moyenne entre une mesure contenant un item et une mesure en contenant plusieurs.

Enfin, plusieurs biais peuvent exister lorsqu'une personne répond à un questionnaire. Probablement l'un des plus connus est le biais de méthodes communes. Il s'agit ici du degré auquel les résultats et surtout les corrélations obtenus suite à la passation d'un questionnaire sont altérés à cause de la méthode utilisée et non à cause des différences individuelles (Kamakura, 2010). Plus concrètement, ce biais apparaît lorsque les réponses données à un questionnaire s'influencent entre elles. Il existe plusieurs manières de contourner ce phénomène. Dans une revue de littérature publiée en 2003, Podsakoff, Mackenzie, Lee et Podsakoff ont répertorié différentes causes pouvant expliquer ce biais. Elles sont multiples et, dans notre cas, il apparaît important de se concentrer sur certaines d'entre elles. Le biais de méthodes communes peut être dû, notamment, aux caractéristiques des items. En effet, les auteurs mettent en avant le fait que certaines questions pourraient être mal comprises par les participants et amèneraient une certaine ambiguïté. Selon eux, la complexité des termes employés dans un questionnaire peut avoir une influence sur les réponses données par les participants. Et ce degré de difficulté peut parfois être dû à la complexité et l'abstraction du construit étudié. Cette théorie rejoint l'un des arguments contre l'utilisation des mesures à un item. C'est

le cas de Straub (1989) par exemple qui mettait cet argument en avant pour justifier l'utilisation des échelles à plusieurs items. Pourtant, cette cause ne semble pas propre à ce type d'échelles selon Podsakoff et al. (2003), elle apparaît, peu importe la taille de l'échelle. Le format du questionnaire possède également une influence, selon les auteurs, qui indiquent que, plus le questionnaire possède d'éléments, plus le risque de se souvenir des réponses précédentes est diminué. Un autre biais tout aussi connu est le biais de monométhode. Il est très proche de celui de la méthode commune ; la différence réside ici dans le simple fait qu'utiliser une seule et unique méthode risque de fausser les résultats observés puisque rien n'indique que cette méthode capture correctement ce que nous cherchons à mesurer. En 2013, Ortiz de Guinea, Titah et Léger ont tenté de contourner le problème en utilisant deux méthodes différentes pour mesurer le même construit : une méthode implicite (neurophysiologique) et une méthode explicite (questionnaire). Les résultats suggèrent une certaine différence entre les construits considérés comme « primitifs » (l'activation émotionnelle) et les construits considérés comme « complexes » (l'engagement et la charge cognitive) (Ortiz de Guinea, Titah et Léger, 2013, p. 841). Les construits primitifs ne sembleraient pas souffrir du biais de monométhode, à l'inverse des construits plus complexes.

1.4 Mesurer les perceptions à l'aide d'échelles à un item en UX

Cette dernière section met en avant une sélection des principales mesures à un item utilisées dans le domaine de l'expérience utilisateur (UX) : l'Affective Slider, le NPS, le CSAT et le CES. La première partie explique leur célébrité à travers l'industrie, nous pouvons y remarquer une certaine domination de ce format d'échelles. La deuxième partie les présente concrètement en expliquant leur libellé, la manière de les calculer et les construits auxquels elles sont reliées. La troisième et dernière partie met en avant une sélection de leurs principaux antécédents et conséquences afin de mieux les comprendre.

1.4.1. Principales données concernant l'utilisation des échelles à un item dans l'industrie

Notre société contemporaine est en constante évolution, en témoignent les nombreuses prouesses technologiques de ces vingt dernières années. Le besoin de tester rapidement avec un grand nombre de données est bel et bien présent. Ces échelles ne

contenant qu'une question en sont l'un des reflets : leur rapidité d'exécution et d'analyse répondent parfaitement à cette philosophie, au détriment parfois de leur adéquation et de leur qualité. Elles ne sont, en effet, pas adaptées dans toutes les situations, cela dépend du degré de complexité d'un construit (Bergkvist et Rossiter, 2007). Nous pouvons affirmer que les besoins entre les scientifiques et les praticiens sont, dans une certaine mesure, opposés : les premiers cherchent généralement à obtenir l'échelle s'approchant le plus possible d'une parfaite représentation du construit d'intérêt. Les derniers ont besoin d'un grand nombre de données, rapidement, afin de prendre plusieurs décisions d'affaires, tout en acceptant cette part d'imperfection. Le choix pour les praticiens, sur le terrain, est donc cornélien : suivre la théorie classique inhérente à la psychométrie et administrer des échelles à plusieurs items ou suivre la voie la plus pratique, bien que plus risquée et administrer plusieurs échelles différentes à un item, parfois incomplètes. Il semblerait que la tendance soit au second argument.

Lors d'une enquête réalisée en 2019 par Pointilist®, une entreprise fournissant un logiciel d'analyse de parcours clients, auprès de 700 professionnels de l'expérience client, plus de la moitié (58,5%) des répondants classent le *NPS* (Net Promoter Score) (Reichheld, 2003, une échelle à un item prétendant mesurer l'intention de recommander et donc, la rétention), comme étant la métrique la plus importante à leurs yeux (Pointilist, 2019). Cette dernière mesure est suivie de près par le *CSAT*[2] (Customer SATisfaction score, également une mesure à un item prétendant ici évaluer la satisfaction). Le *CES* (Customer Effort Score, Dixon, Freeman et Toman, 2010, un item pour mesurer la charge cognitive perçue), lui, arrive en quatrième position. Ces métriques (expliquées dans la section suivante) se situent d'ailleurs au-dessus d'une des plus connues : le retour sur investissement. Ces données suggèrent donc qu'offrir une expérience satisfaisante semble être une des priorités et primerait sur différentes métriques directement liées aux aspects financiers. À travers un sondage réalisé en 2018 par Thompson, Scheibenreif et Chiu, il semblerait que le *CSAT* soit la métrique la plus considérée en comparaison des autres métriques. Sur 208 professionnels de l'expérience client, le *CSAT* domine : 62% des

[2] Cette mesure n'a pas réellement été créée par une personne en particulier. Son acronyme semble plutôt provenir de l'industrie. Plusieurs auteurs ont cependant mesuré la satisfaction de cette manière depuis de nombreuses années, parfois sur une échelle de Likert à 5, 7 ou encore 11 points.

répondants l'indiquent comme l'une des mesures les plus utilisées dans leur compagnie (Thompson, Scheibenreif, Chiu, 2020). Dans cette enquête, le *CES* arrive en septième position (30% des répondants) et le NPS en onzième position (25% des répondants). Pour terminer, la compagnie Incite Group a de son côté également réalisé une enquête, la même année. Leurs intentions étaient quelque peu différentes : connaître les métriques utilisées par les compagnies pour mesurer le retour sur investissement. Leurs constats sont similaires : le *CSAT* et le *NPS* arrivent en tête de liste avec respectivement 55,5% et 36,7% pour la relation B2C (Kees, 2020).

1.4.2. Tour d'horizon des quatre plus importantes mesures à un item en UX[3]

Le point précédent mettait en lumière l'importance de l'utilisation des échelles à un item dans l'industrie. Pourtant, il existe une quantité probablement plus importante d'autres échelles, plus longues. Leur format, leur taille, leur mode de passation, leur type et le construit qu'elles ambitionnent de mesurer diffèrent tous. Certaines sont adaptées à tous types d'interfaces, d'autres sont plus précises et ne sont applicables que pour un type d'interface en particulier. L'utilisabilité globale d'un système peut être par exemple mesurée à l'aide du questionnaire très connu *SUS* (pour System Usability Scale) à travers trois dimensions : l'efficacité, l'efficience et la satisfaction (Brooke, 2013; 1996) ou encore *l'Attrakdiff*, à travers les qualités et hédoniques et pragmatiques perçues (Hassenzahl, Burmester et Koll, 2013; Lallemand, Koenig, Gronier et Martin, 2015). D'autres se veulent plus précises et très contextuelles : s'intéressant par exemple à un type d'interface en particulier. C'est, à titre d'exemple, le cas du questionnaire *GUESS* (pour Game User Experience Satisfaction Scale), utilisé dans le milieu des jeux vidéo à travers neuf dimensions (Phan, Keebler et Chaparro, 2016) ou encore le très connu *WebQual*, développé en 2007 et propre au domaine du commerce électronique (Loiacono, Watson et Goodhue, 2007), pour n'en citer qu'une infime partie[4].

[3] Notre recherche s'intéresse en particulier aux construits reliés aux émotions. Nous avons donc volontairement écarté les échelles mesurant des construits propres aux systèmes (efficience, facilité d'utilisation, utilité perçue, etc.).

[4] Nous vous invitons à consulter les ressources suivantes, très complètes, à travers lesquelles différentes échelles utiles dans notre domaine sont présentées : « Méthodes de Design UX, 30 méthodes

Tableau 1.1. Mesures utilisées dans notre recherche. Il s'agit ici d'un résumé des principales métriques considérées en expérience utilisateur.

Mesure	Libellé	Référence
Affective Slider	Quelle est votre émotion vis-à-vis de votre expérience? Veuillez déplacer le curseur afin de représenter votre niveau de plaisir (malheureux-heureux).  Veuillez déplacer le curseur afin de représenter le niveau d'intensité de votre émotion (calme-excité) 	Betella, Verschure, 2016
Net Promoter Score (NPS)	En vous basant sur cette expérience, quelle est la probabilité que vous recommandiez [nom de la compagnie] à un.e ami.e / collègue / membre de famille? Merci de donner une réponse allant de 0 à 10 (0 = extrêmement improbable; 10 = extrêmement probable).	Reichheld, 2003
Customer SATisfaction score (CSAT)	Veuillez indiquer votre niveau de satisfaction sur l'expérience que vous avez vécue aujourd'hui. Merci de donner une réponse allant de 1 à 5 (1 = très insatisfait; 5 = très satisfait).	[5]
Customer Effort Score (CES)	Quel est, selon-vous, le niveau d'effort que vous avez dû déployer pour faire [nom de la tâche]? Merci de donner une réponse allant de 1 à 5 (1 = très faible; 5 = très fort).	Dixon, Freeman et Toman, 2010

L'une des plus anciennes mesures de ce format (et très probablement la plus connue) est le *Self Assessment Manikin Scale* (SAM Scale, Bradley et Lang, 1994). Cette échelle a été à l'origine développée pour mesurer différentes émotions et comporte trois dimensions, chacune mesurée à l'aide d'une seule et unique question : le plaisir, l'activation émotionnelle (comprenez l'intensité de l'émotion) et la perception du niveau de contrôle. Chaque item y est imagé par plusieurs pictogrammes représentant différents humains sur une échelle de 9 points. Une vingtaine d'années plus tard, en 2016, Betella et Verschure mettaient en place une variante « … *exploitant les normes contemporaines de conceptions pour les interfaces utilisateurs ainsi que les représentations graphiques métacommunicatives modernes des émotions*… » (Betella et Verschure, 2016, p. 3 -

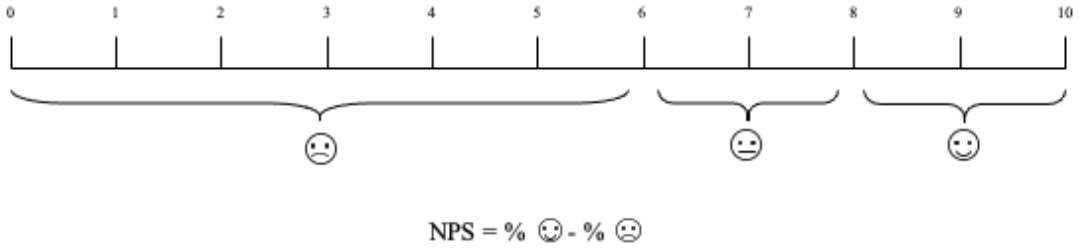
fondamentales pour concevoir des expériences optimales », de Carine Lallemand, Guillaume Gronier et Marc Dugué (2018); « Measuring the User Experience, 2nd Edition », de Tom Tullis et Bill Albert

[5] Cf. point 1.4.1. Principales données concernant l'utilisation des échelles à un item dans l'industrie.

traduction libre), qu'ils ont nommée *Affective Slider*. Celle-ci se veut donc mieux adaptée à l'actualité et mesure uniquement les deux premières dimensions proposées par Bradley et Lang, à savoir : le niveau de plaisir et l'activation émotionnelle perçus, mesurés à travers un curseur allant de 0 à 100 points. Cette dernière spécificité amène à une plus grande variabilité des réponses obtenues lors de sa passation.

En 2003, Reichheld publiait un article dans lequel il mettait au point une mesure très particulière prétendant être « LA » métrique dont les entreprises devaient absolument se servir pour en évaluer leur santé : le *Net Promoter Score* (NPS, Reichheld, 2003). Cette échelle a pour vocation de mesurer l'intention de recommandation, évaluée entre 0 et 10. Sa première spécificité réside dans le classement des répondants : ceux ayant évalué leur intention entre 0 et 6 sont appelés détracteurs, ceux pour qui l'intention de recommander se situe entre 7 et 8 sont considérés comme passifs (et ne sont pas pris en compte dans les résultats) et la classe au-dessus est considérée comme étant les promoteurs. Sa seconde spécificité réside dans son calcul, qui est un ratio entre le nombre de détracteurs et le nombre de promoteurs. La figure suivante (***Figure 1.1***) propose une schématisation de cette mesure. Le pourcentage obtenu suite à cette formule devrait indiquer où se situe une entreprise par rapport à elle-même, mais également par rapport à ses concurrents.

Figure 1.1. Schématisation du NPS



La satisfaction est un construit mesuré depuis de nombreuses années. Citons, à titre d'exemple, les travaux de Tse et Wilton (1988) ou encore ceux de Peterson et Wilson (1992). Il est possible et courant de la mesurer à travers une seule question (Bendle, Farris, Pfeifer et Reibstein, 2016). D'ailleurs, il existe plusieurs études mettant en avant une certaine force de cette mesure lorsqu'elle est administrée de cette manière (Nagy, 2002; Wanous, Reichers, 1996). Bien que ne portant pas ce nom initialement, le *CSAT* est une mesure bien connue et utilisée. Il s'agit d'une échelle à un item dont le but est de mesurer la satisfaction résultante, notamment, d'une interaction.

Enfin, il est également courant, lorsqu'il s'agit de mesurer les perceptions résultant d'une interaction, d'évaluer l'effort mental. En 1985, Zijlstra et Van Doorn ont créé le *Subjective Mental Effort Questionnaire* (SMEQ). Sous la forme d'un item, la graduation se fait ici de 0 à 150. 25 ans plus tard, Dixon, Freeman et Toman développent une version beaucoup plus courte, à 5 points, nommée *Customer Effort Score* (CES). Les auteurs affirment que cette dernière mesure « *surpasse les deux* [CSAT et NPS] … ». (Dixon, Freeman et Toman, 2010, p. 7 - traduction libre).

1.4.3. Antécédents et conséquences

À travers notre recherche, nous nous sommes intéressés à trouver les principaux antécédents (éléments permettant de prédire les réponses à ces questions) et conséquences (ce qu'elles permettent de prédire) de ces cinq mesures. Nos principaux mots-clés étaient des termes tels que « *net promoter score antecedents outcomes* », « *single item satisfaction key drivers* » ou encore « *self reported mental load predictors* ». Une première observation nous montre la richesse de la recherche pour les deux métriques les plus considérées : le *CSAT* et le *NPS* (Pointilist, 2019). Le **tableau 1.2** propose une sélection des principaux antécédents et conséquences des cinq mesures citées dans le point précédent.

Bien que les études des antécédents et des conséquences de la perception de la satisfaction soient nombreuses (Churchill et Surprenant, 1982 ; Tse et Wilton, 1988), il n'en existe à notre connaissance que très peu ayant tenté de le faire spécifiquement pour le *CSAT.* Cependant, plusieurs études semblent avoir comblé ce besoin en mesurant la satisfaction à l'aide d'une échelle à un item. Il s'avère que la qualité perçue d'un service fourni est l'un des points clés influençant grandement la satisfaction. L'une des publications ayant probablement eu le plus d'impact est celle de Cronin et Taylor, en 1992. À travers une étude réalisée sur base de questionnaires, les auteurs ont découvert que le lien entre la qualité du service perçue et la satisfaction apparaissait comme fort et évident (Cronin et Taylor, 1992). D'autres recherches semblent aller dans ce même sens et indiquent que, à différents degrés, ce construit aurait une nette influence sur la satisfaction mesurée par une échelle à un item (Kumar et Mittal, 2015; Ishaq, 2011; Zhou, Wang, Shi, Zhang, Zhang et Guo, 2019; De Pechpeyrou et Nicholson, 2019; Bharadwaj

et Matsuno, 2006). D'autres auteurs ont mis en avant une certaine force de prédiction de la part de la confiance perçue (Bharadwaj et Matsuno, 2006). Il semblerait également que les attributs propres à un système permettraient de prédire la satisfaction. Tractinsky, Katz et Ikar (2000) ont remarqué, à l'aide d'une étude réalisée sur des machines distributrices de billets, que « *... l'apparence affecte les perceptions de l'apparence, l'utilisabilité réelle affecte les perceptions de l'utilisabilité, et cette combinaison des deux affecte la satisfaction du système.* » (Tractinsky, Katz et Ikar, 2000, p.140 – traduction libre). Ces éléments apparaissent aux yeux des auteurs comme étant particulièrement corrélés. Enfin, il apparaît également que l'effort perçu aurait un impact sur la satisfaction (De Pechpeyrou et Nicholson, 2019; Schmutz, Heinz, Métrailler et Opwis, 2009). Au sujet des conséquences, cette satisfaction, lorsque mesurée de cette manière aurait un impact sur la loyauté (Hallowell, 1996; Zhou et al., 2019) et l'intention de consommer (Cronin et Taylor, 1992; Jones et Suh, 2000; Zhou et al., 2019).

La publication de Reichheld en 2003 mettant en avant une toute nouvelle mesure, le NPS, a fait couler beaucoup d'encre : plusieurs études ont tenté d'en évaluer sa réelle validité (surtout prédictive). Certaines publications l'ont fait avec succès. Mecredy, Wright et Feetham ont réalisé en 2018 une étude longitudinale étalée sur cinq ans dans le secteur primaire. Les auteurs ont réussi à supporter les dires de Reichheld et ont observé certaines corrélations (bien que non significatives) entre le NPS et les résultats d'une entreprise. Les auteurs y affirment qu'« *... il n'y a pas assez d'évidence que pour contredire les précédentes conclusions indiquant que le NPS est corrélé positivement aux revenus d'une entreprise.* » (Mecredy, Wright, Feetham, 2018, p4 - traduction libre). D'autres études semblent aller dans le même sens. C'est le cas de celle réalisée par De Haan, Verhoef et Wiesel (2015) qui ont également tenté d'élucider la question à travers plusieurs industries différentes et ont découvert que le NPS avait un impact important sur la rétention, tant au niveau des clients au sein même de chaque entreprise qu'entre les entreprises appartenant à la même industrie et faisait également partie des deux meilleures métriques analysées durant leur étude. Cependant, d'autres n'ont pas réussi à supporter les affirmations de Reichheld. C'est notamment le cas de Kristensen et Eskilden (2014), qui à travers le domaine des assurances, ne sont pas parvenus à mettre en avant la force de cette métrique par rapport aux autres. Au sujet des principaux antécédents, il semblerait

que la satisfaction soit un des éléments permettant de prédire les réponses au NPS (Pollack et Alexandrov, 2013). Les attributs propres à un système en feraient également partie (Chang et Fan, 2013). Enfin, Korneta (2014) a découvert que les éléments propres à une compagnie (image sociale et histoire notamment), la confiance et la qualité du service perçues feraient au même titre partie des principales conséquences du NPS.

Considérant sa relative jeunesse, il n'existe aucune étude à notre connaissance ayant tenté de comprendre et analyser tant les facteurs influençant que ceux permettant de prédire les réponses spécifiquement de l'Affective Slider. Cette échelle mesurant les mêmes construits que le SAM Scale (excepté le niveau de contrôle perçu), il apparaît évident que leurs antécédents soient similaires. Il s'agit ici principalement des attributs propres à un système. Lors d'une étude sur les baladeurs audio, Mahlke et Thüring (2007) ont découvert que les qualités esthétiques et utilitaires tendraient à avoir un certain impact sur les perceptions de plaisir et de l'activation émotionnelle. Les auteurs ont également mis en avant une plus grosse influence de la part des qualités utilitaires. Au sujet des conséquences, il ne semble pas exister de recherche précise sur cette mesure.

Enfin, au sujet du CES, aucune étude ne semble s'être réellement intéressée à ses antécédents et conséquences. Cependant, si l'on se fie aux recherches de Tractinsky, Katz et Ikar (2000), il semblerait que certains éléments propres à la perception des éléments esthétiques et utilitaires auraient aussi un impact sur l'impression de facilité d'utilisation d'une interface. Ces deux construits étant intimement reliés, il apparaît logique que les attributs d' un système influencent les perceptions de l'effort suite à la complétion d'une tâche. Du côté des conséquences, elles sont sensiblement similaires à celles des mesures précédemment citées : un certain degré de satisfaction (Kumar, Singh, Manna, 2013; De Pechpeyrou et Nicholson, 2019) et une loyauté (Dixon, Freeman et Toman, 2010).

Tableau 1.2. Sélection des principaux antécédents et conséquences des quatre mesures utilisées dans notre étude.

Mesure	Construit	Antécédents	Référence	Conséquences	Référence
CSAT	Satisfaction	Qualité du service fourni	Kumar et Mittal, 2015 ; Ishaq, 2011; Cronin et Taylor, 1992; Zhou, Wang, Shi, Zhang, Zhang, et Guo, 2019; De Pechpeyrou et Nicholson, 2019	Intention de consommer	Cronin et Taylor, 1992; Jones et Suh, 2000; Zhou, Wang, Shi, Zhang, Zhang et Guo, 2019
		Confiance	Bharadwaj et Matsuno, 2006	Loyauté	Hallowell, 1996; Zhou, Wang, Shi, Zhang, Zhang et Guo, 2019; De Haan, Verhoef et Wiesel, 2015
		Attributs d'un système	Tractinsky, Katz et Ikar, 2000		
		Effort	De Pechpeyrou et Nicholson, 2019; Schmutz et al., 2009		
NPS	Intention de recommander	Satisfaction	Pollack et Alexandrov, 2013	Intention de consommer	Pollack et Alexandrov, 2013
		Confiance	Korneta, 2014		
		Qualité du service fourni	Korneta, 2014	Loyauté	De Haan, Verhoef et Wiesel, 2015
		Éléments propres à la compagnie	Korneta, 2014		
		Attributs d'un système	Chang et Fan, 2013		
CES	Effort	Attributs d'un système	Tractinsky, Katz et Ikar, 2000	Satisfaction	Kumar, Singh et Manna, 2013; De Pechpeyrou et Nicholson, 2019
				Loyauté	Dixon, Freeman et Toman 2010
Affective Slider	Plaisir / activation	Attributs d'un système	Mahlke et Thüring, 2007	N/A	

1.5 Conclusion

Pour conclure, nous pouvons affirmer que, bien que très utilisées à travers l'industrie, principalement pour leurs avantages tant pratiques (Drolet et Morrisson, 2001) que monétaires (Wanous, Reichers et Hudy, 1997), il n'existe toujours pas à l'heure actuelle de réel consensus sur la recommandation ou non des échelles à un item. La littérature ne semble pas encore être d'accord sur la direction à emprunter lorsqu'il s'agit de mesurer les perceptions. Il s'agit ici d'un débat vieux de plusieurs dizaines d'années : citons à titre d'exemple les travaux de Jacoby et Churchill, respectivement en 1978 et 1979.

Également, il existe à notre connaissance, un deuxième manquement quant aux éléments permettant de prédire les réponses à ce type de mesures. La recherche existe au sujet des construits qu'elles prétendent mesurer, mais peu d'auteurs se sont réellement penchés sur ces quatre mesures décrites plus haut, à savoir : l'Affective Slider, le NPS, le CSAT et le CES. Les comportements découlant des réponses à ces questions ne sont pas en reste, peu d'études ont tenté d'élucider cette problématique.

Nous avons vu qu'il était possible et parfois fiable de mesurer certains construits à l'aide d'une échelle à un item. Cependant, une question simple, mais pourtant fondamentale reste en suspens : « *À quel point ces mesures reflètent-elles réellement ce qu'une personne a vécu lorsqu'elle a interagi avec un système?* ». Il serait réducteur et surtout faux de prétendre que ce lien entre l'expérience vécue (mesurée de manière implicite) et perçue (mesurée de manière explicite) n'a jamais été étudié. La recherche sur le sujet existe (Le Pailleur, Huang, Léger et Sénécal, 2020; Lourties, Léger, Sénécal, Fredette et Chen, 2018; Ortiz de Guinea, Titah et Léger, 2013). Ces dernières études ont observé une certaine relation, à différents degrés, entre les deux types de mesures. Mais aucun consensus ne semble exister au sujet de la représentativité réelle des construits mesurés à l'aide d'une échelle à un item. Ortiz de Guinea, Titah et Léger (2013) suggèrent d'ailleurs qu'utiliser plusieurs méthodes différentes pour évaluer un même construit serait plus intéressant que de n'en utiliser qu'une. C'est à travers ces dernières observations que notre recherche prend ses racines.

Références

André, Nathalie, Nathalie Loye et Louis Laurencelle (2016). « La validité psychométrique : Un regard global sur le concept centenaire, sa genèse, ses avatars », *Mesure et évaluation en éducation*, vol. 37, no 3, p. 125-148.

Baumgartner, Hans et Christian Homburg (1996). « Applications of structural equation modeling in marketing and consumer research: A review », *International Journal of Research in Marketing*, vol. 13, no 2, p. 139-161.

Bendle, Neil T., Paul W. Farris, Phillip E. Pfeifer et David J. Reibstein (2016). *Marketing metrics: The manager's guide to measuring marketing performance, third edition*, 3e éd., Upper Saddle River, USA, Pearson, 439-439 p.

Bergkvist, Lars et John R. Rossiter (2007). « The predictive validity of multiple-item versus single-item measures of the same constructs », *Journal of Marketing Research*, vol. 44, no 2, p. 175-184.

Betella, Alberto et Paul F. M. J. Verschure (2016). « The affective slider: A digital self-assessment scale for the measurement of human emotions », *PLoS ONE*, vol. 11, no 2.

Bharadwaj, Neeraj et Ken Matsuno (2006). « Investigating the antecedents and outcomes of customer firm transaction cost savings in a supply chain relationship », *Journal of Business Research*, vol. 59, no 1, p. 62-72.

Borsboom, Denny, Gideon J. Mellenbergh et Jaap Van Heerden (2003). « The theoretical status of latent variables », *Psychological Review*, vol. 110, no 2, p. 203-219.

Boyle, Gregory J. (1991). « Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? », *Personality and Individual Differences*, vol. 12, no 3, p. 291-294.

Bradley, Margaret M. et Peter J. Lang (1994). « Measuring emotion: The self-assessment manikin and the semantic differential », *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no 1, p. 49-59.

Brooke, John (1996). « Sus-a quick and dirty usability scale », *Usability evaluation in industry*.

Brooke, John (2013). « Sus: A retrospective », *Journal of usability studies*, vol. 8, no 2, p. 29-40.

Chang, En-Chung et Xiaomeng Fan (2013). « More promoters and less detractors: Using generalized ordinal logistic regression to identify drivers of customer loyalty », *International Journal of Marketing Studies*.

Churchill, Gilbert A. (1979). « A paradigm for developing better measures of marketing constructs », *Journal of Marketing Research*, vol. 16, no 1, p. 64-73.

Churchill, Gilbert A., Neil M. Ford et Orville C. Walker (1974). « Measuring the job satisfaction of industrial salesmen », *Journal of Marketing Research*, vol. 11, no 3, p. 254-260.

Churchill, Gilbert A. et Carol Surprenant (1982). « An investigation into the determinants of customer satisfaction », *Journal of Marketing Research*, vol. 19, no 4, p. 491-491.

Cook, David A. et Thomas J. Beckman (2006). « Current concepts in validity and reliability for psychometric instruments: Theory and application », *American Journal of Medicine*, vol. 119, no 2, p. 166.e167-166.e116.

Cook, T. D., D. T. Campbell et A. Day (1979). *Quasi-experimentation: Design & analysis issues for field settings*, 1e éd., Boston, USA, H. Mifflin, 405-405 p.

Cronbach, Lee J. (1951). « Coefficient alpha and the internal structure of tests », *Psychometrika*, vol. 16, no 3, p. 297-334.

Cronin, J. Joseph et Steven A. Taylor (1992). « Measuring service quality: A reexamination and extension », *Journal of Marketing*, vol. 56, no 3, p. 55-55.

de Haan, Evert, Peter C. Verhoef et Thorsten Wiesel (2015). « The predictive ability of different customer feedback metrics for retention », *International Journal of Research in Marketing*, vol. 32, no 2, p. 195-206.

De Pechpeyrou, Pauline et Patrick Nicholson (2019). *Réclamation et satisfaction : L'effort perçu du client rebat les cartes complaining behavior and satisfaction: Customer's effort score shuffle the cards*. Récupéré de http://archives.marketing-trends-congress.com/2019/pages/PDF/26.pdf

DeVellis, Robert F. (2003). *Scale development: Theory and applications*, 2e éd., Thousand Oaks, USA, SAGE Publications, 171-171 p.

DeVon, Holli A., Michelle E. Block, Patricia Moyle-Wright, Diane M. Ernst, Susan J. Hayden, Deborah J. Lazzara*, et al.* (2007). « A psychometric toolbox for testing validity and reliability », *Journal of Nursing Scholarship*, vol. 39, no 2, p. 155-164.

Dew, Denis (2008). « Construct », dans Paul. J. Lavrakas (dir.), *Encyclopedia of survey research methods*, SAGE Publications, p. 133-134.

Dixon, Matthew, Karen Freeman et Nicolas Toman (2010). « Stop trying to delight your customers », *Harvard Business Review*, vol. 88, no 7-8.

Drolet, Aimee L. et Donald G. Morrison (2001). *Do we really need multiple-item measures in service research?*

Gefen, David et Detmar Straub (2005). « A practical guide to factorial validity using pls-graph: Tutorial and annotated example », *Communications of the Association for Information Systems*, vol. 16, p. 91-109.

Hallowell, Roger (1996). « The relationships of customer satisfaction, customer loyalty, and profitability: An empirical study », *International Journal of Service Industry Management*, vol. 7, no 4, p. 27-42.

Hassenzahl, Marc, Michael Burmester et Franz Koller (2003). « Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität », *Mensch & Computer 2003 : Interaktion in Bewegung*, p. 187-196.

Haynes, Stephen N., David C. S. Richard et Edward S. Kubany (1995). « Content validity in psychological assessment: A functional approach to concepts and methods », *Psychological Assessment*, vol. 7, no 3, p. 238-247.

Hogan, Thomas P., Amy Benjamin et Kristen L. Brezinski (2000). « Reliability methods: A note on the frequency of use of various types », *Educational and Psychological Measurement*, vol. 60, no 4, p. 523-531.

Imbault, C., D. Shore et V. Kuperman (2018). « Reliability of the sliding scale for collecting affective responses to words », *Behavior Research Methods*, vol. 50, no 6, p. 2399-2407.

Ishaq, Muhammad Ishtiaq (2011). « A study on relationship between service quality and customer satisfaction: An empirical evidence from pakistan telecommunication industry », *Management Science Letters*, vol. 1, no 4, p. 523-530.

Jacoby, Jacob (1978). « Consumer research: How valid and useful are all our consumer behavior research findings? », *Journal of Marketing*, vol. 42, no 2, p. 87-96.

Jones, Michael A. et Jaebeom Suh (2000). « Transaction-specific satisfaction and overall satisfaction: An empirical analysis », *Journal of Services Marketing*, vol. 14, no 2, p. 147-159.

Kamakura, Wagner A. (2010). « Common methods bias », dans, Chichester, UK, John Wiley & Sons, Ltd.

Kees, Jasmine (2020). *2020 state of customer service report*, 46-46 p.

Kerlinger, Fred N. et Howard B. Lee (2000). *Foundations of behavioral research*, 4e éd., Fort Worth, USA, Harcourt College Publisher, 890-890 p.

Kim, Yong Mi (2009). « Validation of psychometric research instruments: The case of information science », *Journal of the American Society for Information Science and Technology*, vol. 60, no 6, p. 1178-1191.

Korneta, Pawel (2014). « What makes customers willing to recommend a retailer - the study on roots of positive net promoter score index abstract », *Central European Review of Economics & Finance*.

Kristensen, Kai et Jacob Eskildsen (2014). « Is the nps a trustworthy performance measure? », *TQM Journal*, vol. 26, no 2, p. 202-214.

Kumar, Niraj, Ajay Pal Singh et Reshmi Manna (2013). *Analyzing consumer behaviour towards service quality of indianc electronic gadget firms*, 14-14 p.

Kumar, R. et A. Mittal (2015). « Customer satisfaction and service quality perception of technology based banking services: A study on selected public sector banks in india », *Global Journal of Management and Business Research : E Marketing*, vol. 15, no 5, p. 39-45.

Lallemand, C., V. Koenig, G. Gronier et R. Martin (2015). « Création et validation d'une version française du questionnaire attrakdiff pour l'évaluation de l'expérience utilisateur des systèmes interactifs », *Revue Europeenne de Psychologie Appliquee*, vol. 65, no 5, p. 239-252.

Le Pailleur, Félix, Bo Huang, Pierre Majorique Léger et Sylvain Sénécal (2020). « A new approach to measure user experience with voice-controlled intelligent assistants: A pilot study », communication présentée au *HCII 2020*, Copenhagen, Denmark.

Lewis, Bruce R., Gary F. Templeton et Terry Anthony Byrd (2005). « A methodology for construct development in mis research », *European Journal of Information Systems*, vol. 14, no 4, p. 388-400.

Lin, Aleck, Shirley Gregor et Michael Ewing (2008). « Developing a scale to measure the enjoyment of web experiences », *Journal of Interactive Marketing*, vol. 22, no 4, p. 40-57.

Loiacono, Eleanor T., Richard T. Watson et Dale L. Goodhue (2007). « Webqual: An instrument for consumer evaluation of web sites », *International Journal of Electronic Commerce*, vol. 11, no 3, p. 51-87.

Lourties, Sébastien, Pierre Majorique Léger, Sylvain Sénécal, Marc Fredette et Shang Lin Chen (2018). « Testing the convergent validity of continuous self-perceived measurement systems: An exploratory study », communication présentée au *HCII 2018*, Las Vegas, USA.

MacKenzie, Scott B., Philip M. Podsakoff et Nathan P. Podsakoff (2011). *Construct measurement and validation procedures in mis and behavioral research: Integrating new and existing techniques*, vol. 35, 293-334 p.

Mahlke, Sascha et Manfred Thüring (2007). « Studying antecedents of emotional experiences in interactive contexts », communication présentée au *Conference on Human Factors in Computing Systems*, 2007, New York, New York, USA.

Mecredy, Philip, Malcolm J. Wright et Pamela Feetham (2018). « Are promoters valuable customers? An application of the net promoter scale to predict future customer spend », *Australasian Marketing Journal*, vol. 26, no 1, p. 3-9.

Moore, Gary C. et Izak Benbasat (1991). « Development of an instrument to measure the perceptions of adopting an information technology innovation », *Information Systems Research*, vol. 2, no 3, p. 192-222.

Nagy, Mark S. (2002). « Using a single-item approach to measure facet job satisfaction », *Journal of Occupational and Organizational Psychology*, vol. 75, no 1, p. 77-86.

Nunnally, Jum C. et Ira H. Bernstein (1994). *Psychometric theory*, 3e éd., McGraw-Hill, 752-752 p.

Organisation internationale de normalisation et internationale Commission électrotechnique (2016). *Systems and software engineering : Systems and software quality requirements and evaluation (square) : Measurement of system and software product quality = ingénierie des systèmes et du logiciel : Exigences de qualité et évaluation des systèmes et du logiciel (square) : Mesurage de la qualité du produit logiciel et du système*, ISO/IEC, (1st ed.), c. viii, 45 p.

Ortiz de Guinea, Ana, Ryad Titah et Pierre Majorique Léger (2013). « Measure for measure: A two study multi-trait multi-method investigation of construct validity in is research », *Computers in Human Behavior*, vol. 29, no 3, p. 833-844.

Peter, J. Paul (1979). « Reliability: A review of psychometric basics and recent marketing practices », *Journal of Marketing Research*, vol. 16, no 1, p. 6-6.

Peterson, Robert A. et William R. Wilson (1992). « Measuring customer satisfaction: Fact and artifact », *Journal of the Academy of Marketing Science*, vol. 20, no 1, p. 61-71.

Phan, Mikki H., Joseph R. Keebler et Barbara S. Chaparro (2016). « The development and validation of the game user experience satisfaction scale (guess) », *Human Factors*, vol. 58, no 8, p. 1217-1247.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee et Nathan P. Podsakoff (2003). « Common method biases in behavioral research: A critical review of the literature and recommended remedies », *Journal of Applied Psychology*, vol. 88, no 5, p. 879-879.

Pointilist (2019). *State of customer journey management & cx measurement*, 31-31 p.

Pollack, Birgit Leisen et Aliosha Alexandrov (2013). « Nomological validity of the net promoter index question », *Journal of Services Marketing*, vol. 27, no 2, p. 118-129.

Riedl, René, Fred D. Davis et Alan R. Hevner (2014). « Towards a neurois research methodology: Intensifying the discussion on methods, tools, and measurement », *Journal of the Association for Information Systems*, vol. 15, p. 1-35.

Rossiter, John R. (2002). « The c-oar-se procedure for scale development in marketing », *International Journal of Research in Marketing*, vol. 19, no 4, p. 305-335.

Sarstedt, Marko et Petra Wilczynski (2009). « More for less? A comparison of single-item and multi-item », *Die Betriebswirtschaft*, vol. 69, no 2, p. 211-227.

Schmutz, Peter, Silvia Heinz, Yolanda Métrailler et Klaus Opwis (2009). « Cognitive load in ecommerce applications—measurement and effects on user satisfaction », *Advances in Human-Computer Interaction*, vol. 2009, p. 1-9.

Straub, Detmar, Marie-Claude Boudreau et David Gefen (2004). « Validation guidelines for is positivist research », *Communications of the Association for Information Systems*, vol. 13, p. 380-427.

Straub, Detmar W. (1989). « Validating instruments in mis research », *MIS Quarterly: Management Information Systems*, vol. 13, no 2, p. 147-165.

Streiner, David L. (2003). « Starting at the beginning: An introduction to coefficient alpha and internal consistency », *Journal of Personality Assessment*, vol. 80, no 1, p. 99-103.

Thompson, Ed, Don Scheibenreif et Michael Chiu (2020). *How to manage customer experience metrics*, 20-20 p. Récupéré de https://www.gartner.com/document/3979139?ref=lib

Tractinsky, N., A. S. Katz et D. Ikar (2000). « What is beautiful is usable », *Interacting with Computers*, vol. 13, no 2, p. 127-145.

Tse, David K. et Peter C. Wilton (1988). « Models of consumer satisfaction formation: An extension », *Journal of Marketing Research*, vol. 25, no 2, p. 204-212.

Wanous, John P. et Arnon E. Reichers (1996). « Estimating the reliability of a single-item measure », *Psychological Reports*, vol. 78, no 2, p. 631-634.

Wanous, John P., Arnon E. Reichers et Michael J. Hudy (1997). « Overall job satisfaction: How good are single-item measures? », *Journal of Applied Psychology*, vol. 82, no 2, p. 247-252.

Zhou, Ronggang, Xiaorui Wang, Yuhan Shi, Renqian Zhang, Leyuan Zhang et Haiyan Guo (2019). « Measuring e-service quality and its importance to customer satisfaction and loyalty: An empirical study in a telecom setting », *Electronic Commerce Research*.

Zijlstra, F. R. H. et L. Van Doorn (1985). *The construction of a scale to measure subjective effort*, Delft.

Chapitre 2
Quantity over quality: Do single-item scales reflect what users truly experience ?[6]

Abstract

Single-item scales are widely used in the field of user experience to report the emotions. However, they are strongly criticized and discouraged by the scientific community. While they have several practical advantages, single-item scales are mainly criticized for their psychometric weakness. Our research explores to what extent single-item scales reflect what a user has experienced during his or her interaction with technology, overall, but also at specific moments such as first and last impressions. This research also explores the sensitivity of these measures across different contexts of use brought by the presence or absence of certain interface features. We conducted a correlational study with 40 users while interacting with financial institution websites. We used two methods to evaluate the experience: lived and measured implicitly using psychophysiological instruments on one hand and self-perceived and measured using single-item scales, on the other hand. Overall, our results suggest limited correlations and in most cases contradictory between lived and perceived experience. We conclude by highlighting the limits to be taken into account when professionals are led to use single-item scales.

Keywords: Arousal, Cognitive load, Net Promoter Score, Pleasure, Satisfaction, Single-item scales, User experience, Valence.

2.1 Introduction

Microsoft, HSBC, Delta Airlines, Procter & Gamble… All multinationals based in different countries. All from different sectors. One of the elements that connects them: The use of the Net Promoter Score (NPS, Reichheld, 2003). The list of companies using the highly criticized, yet widely used measure created by Reichheld and presented in the

[6] Cet article sera soumis à la revue Computers in Human Behavior Reports lors du dépôt du mémoire.

prestigious Harvard Business Review, to measure their performance is long, according to the website of the company that provides it (Bain & Company, 2020). Single-item scales dominate: out of 700 customer experience professionals surveyed, it appears that the NPS and the CSAT are on top of the most considered metrics when it comes to understanding the users (Pointillist, 2019). A study published by Gartner shows evidence of significant use and especially high popularity of the CSAT (Thompson, Scheibenreif & Chiu, 2020). This notoriety was also highlighted in an article published in the Wall Street Journal (Safdar & Pacheco, 2019). According to them, the NPS was cited more than 150 times on financial conference calls across 500 companies. These two measures are therefore used extensively throughout the industry. But scientists usually criticize these scales for their psychometric weakness (Straub, 1989; Baumgartner & Homburg, 1996). However, we have to recognize their simplicity and the benefits that come with single-item scales: reduced mental workload (Drolet & Morrison, 2001), reduced costs (Bergkvist & Rossiter, 2007). In the end, single-item scales would allow respondents to answer without intellectualizing their responses; single-item scales may actually be closer to what a person experienced when interacting with an interface (Ortiz de Guinea, Titah, & Léger, 2013). However, at the moment, no agreement seems to have been reached between the use or non-use of this type of scale in the user experience (UX) industry.

The first advantage of single-item measures lies in their nature: their format. A certain speed of execution that would have a direct impact on the mental load associated with answering a questionnaire (Straub, Boudreau & Gefen, 2004). Increasing the number of items increases fatigue, boredom, and inattention. Different emotional states lead to certain inappropriate response behaviours (Drolet & Morrison, 2001). Their format also allows them to avoid a kind of contamination of other items. As Nagy (2002) states about job satisfaction: "… *summing up facets not important to an employee's overall satisfaction will lead to misleading conclusions about that employee's overall satisfaction level.* " (Nagy, 2002, p. 78). The third advantage is their speed of development. They are "*quick and easy*" (Sarstedt, Wilczynski, 2009, p. 214). In contrast, multi-items require many steps before they can be used (Moore & Benbasat, 1991; DeVellis, 2003).

Despite these advantages, they have a huge inconvenient: their limited representativeness of a construct. A construct can have several different facets, several dimensions that contribute to defining it (Kerlinger & Lee, 2000). As the number of items in a scale increases, the chances of capturing all the facets and dimensions of a construct increase at the same time (Baumgartner & Homburg, 1996). The second disadvantage is that it would be very difficult to calculate their reliability (Churchill, 1979), since classical theory usually uses Cronbach's alpha (Cronbach, 1951) to calculate it. And this coefficient is based on the correlations between the different items. Finally, the degree of abstraction and complexity in mentally representing a construct is an argument against the use of these scales. The more abstract the construct, the less likely it is to use a single-item scale (Rossiter, 2002; Wanous, Reichers & Hudy, 1997). A first question therefore arises: "To what extent are single-item scales related with the lived user experience when interacting with an interface?" (RQ1).

Obviously, asking several people to give feedback of their experience calls upon their memory. Memory deteriorates over time and is subject to several biases. Biases are common and can come from many sources (Podsakoff, MacKenzie, Lee & Podsakoff, 2003). In our case, we were interested in memory biases and more specifically in the serial position effect, which is the propension to less recall what happened in the middle of an interaction and better recall what happened first and last (Deese, Kaufman, 1957). This tendency to remember the beginning is called the primacy effect, while the tendency to remember the end is called the recency effect (Martin, Carlson & Buskist, 2010). It is possible that the relationship between what we experience and what we say we have experienced may be different at certain moments in our interaction. We therefore raise a second research question (RQ2): "To what extent do the relationship between the lived and the perceived user experience is different at different moments of an interaction?"

Psychometric qualities of a measure concern its validity, reliability, but also its sensitivity. Sensitivity is defined as the "… *property of a measure that describes how well it differentiates values along the continuum inherent of a construct.*" (Riedl, Davis & Hevner, 2014, p. 17). A measure must, therefore, be capable of distinguishing differences, which may emanate from the system used (Lewis, 2002). Several researches have

suggested that the constructs used in our study have a common antecedent: the attributes of the system used (Tractinsky, Katz & Ikar, 2000; Chang & Fan, 2013; Mahlke & Thüring, 2007). These attributes can be diverse and lead to different perceptions, different contexts of use. A third question (RQ3) arises: "To what extent are single-item scales sensitive enough to distinguish different contexts of user experience, such as hedonic or utilitarian contexts?"

Our study is based on this axis: Exploring different relationships. More specifically, we conducted a study of 40 people on various websites of financial institutions. Using psychophysiological measures to capture the automatic and non-conscious reactions of users, we explore the relationships between what participants experienced during their interaction and what they report having experienced using this type of scale. We also tried to find out if there was any influence of system-specific attributes on responses to the questionnaires.

Our paper is organized as follows: We begin by reviewing what the literature says about this type of scale. We then introduce our research hypotheses. Next, we explain our methodology and we finally discuss the results and our interpretations of these results.

2.2 Litterature review & hypothesis development

2.2.1. Construct definition

Our research is based on the analysis of five widely used constructs in the UX evaluation: Pleasure (valence), emotional arousal, cognitive load, satisfaction, and intention to recommend (close to retention).

Valence: Six emotional states are commonly accepted as universal (Ekman & Friesen, 2003): joy, sadness, surprise, fear, anger and disgust. The research usually uses the construct "valence", which is the result of the difference between positive and negative emotional states. Thus, Valence can be defined as: "… *the direction of behavioral activation and the degree of positive (toward) or negative (away from) emotion for a stimulus*" (Seo, Lee, Chung & Park, 2015 – p. 73-74). This construct is thus lived, through our body and manifests itself, in particular through our facial

muscles (Lang, Greenwald, Bradley & Hamm, 1993). Valence is one of the dimensions of the emotions (Herbon, Peter, Markert & Meer, 2005), and thus can be viewed as a unidimensional construct.

Emotional arousal: This corresponds to the strength of the expression of an emotion. This construct is also lived within us and is expressed, in part, by the activation of our sweat glands. It is very strongly linked to valence. It is common in research to associate them in the same circumplex pattern with, on the one hand, level of pleasure and activation on the other hand (calm or excited) (Russel, 1980). The arousal is another dimension of the emotions (Herbon & al., 2005), and seems to be a unidimensional construct. It is common to measure these last constructs (valence and arousal) by using the Affective Slider (Betella & Verschure, 2016). In this scale, respondents are asked to indicate their perceived level of enjoyment and perceived level of arousal by dragging a slider. Responses usually range from 1 to 100.

Cognitive load : Cognitive load represents the cognitive requirement necessary to complete a task (Xie, Wang, Hao, Chen, An, Wang & Liu, 2017). It can be dependable of the system used or to the nature of the task (Brünken, Plass & Leutner, 2003). This is a multidimensional construct (causal and assessment dimensions) and related to mental load, defined as one of his aspects "… *that originates from the interaction between a task and subject characteristics*" (Paas, Tuovinen, Tabbers & Van Gerven, 2003, p. 64). The cognitive load is frequently measured by the Customer Effort Score (CES, Dixon, Freeman & Toman, 2010). Responses vary from 1 (low effort) to 5 (high effort).

Satisfaction: Satisfaction is also an emotional response (Hansemark & Albinson, 2004), unidimensional and resulting from an experience (Wirtz & Lee, 2003). More specifically, this construct is a comparison between the reward and the associated costs (Churchill & Surprenant, 1982). Satisfaction can be defined as "*A summary affective response of varying intensity (...). With a specific point of determination and limited duration (...). Directed toward focal aspects of product acquisition and/or consumption*" (Giese & Cote, 2002, p. 15). Satisfaction is therefore very much linked to

a person's expectations before an interaction. The metric commonly used to assess the satisfaction is the Customer SATisfaction score (CSAT). Here too, responses vary from 1 (unsatisfied at all) to 5 (fully satisfied).

Customer retention: Retention is a concept very strongly linked to client loyalty and consists of a construct with several emotional and behavioural dimensions (Ranaweera & Neel, 2003). It is "… *customer's stated continuation of a business relationship with the firm.*" (Keiningham, Cooil, Aksoy, Andreassen & Winer, 2007 – p. 364). Retention is a matter of a person's investment (Pingitore, Morgan, Rego, Gigliotti & Meyers, 2007), is strongly related to word of mouth (De Haan, Verhoef & Wiesel, 2015) and strongly related to satisfaction too (Razavi, Safari, Shafie & Vandchali, 2012). One of the most used proxy measures to assess customer retention is the NPS (Reichheld, 2003). Based on an 11-point scale (from 0, being a zero probability of recommendation to 10, being a maximal probability of recommendation), respondents are asked to indicate how likely they would recommend a company to a relative. The score obtained is usually a ratio between the number of people who responded positively (scores of 9 and 10, referred to as promoters) and those who responded negatively (from 0 to 7, referred to as detractors), the others being considered passive.

2.2.2. Research on scale validation

Scale validation is an essential point through the psychometric literature. In 1989, Straub asserted a certain weakness in the validation of tools (Straub, 1989). Twenty years later, this weakness is still present, although the situation has improved (Kim, 2009). There are two main ways of looking at measurement validation. These are reliability and validity. Reliability is an issue of stability (Straub, 1989). It is "*The extent to which a measurement is free of measurement error, and therefore yields the same results on repeated measurement of the same construct.*" (Rield, Davis & Hevner, 2014, p. 29). If a researcher decides to conduct several different surveys, the responses collected should behave in the same way and the results should be equivalent (Kerlinger & Lee, 2000; Nunnally & Bernstein, 1994). Reliability is usually assessed by calculating the degree of variability in responses. Probably the most widely used indicator is Cronbach's alpha

(Hogan, Benjamin & Brezinski, 2000). Here, the aim is to calculate inter-item correlations. It varies from 0 to 1 and is considered acceptable when it reaches 0.7 (Nunnally & Bernstein, 1994). The goal is to achieve the highest coefficient; however, a too high coefficient would indicate some redundancy between the different items (Streiner, 2003). Several approaches exist to assess the reliability of a measure (Straub, Boudreau & Gefen, 2004) and probably the best known and the most widely used is internal consistency. This method involves the simple calculation of correlations between items, using Cronbach's alpha (Straub, Boudreau & Gefen, 2004).

If reliability enables to know how stable a measure is, there is no indication that it actually precisely measures a construct. This is the purpose of validity. Validity is the degree to which a scale actually measures what it purports to measure. This notion is very broad, and we generally distinguish 3 types of validity: content validity, predictive validity and construct validity. Content validity seeks to know the extent of a construct, to create the most representative possible measure of the construct of interest (Riedl, Davis & Hevner, 2014). The main technique consists of conducting a literature review and surveying several experts to help define the construct and thus create the scale (Straub, Boudreau & Gefen, 2004). Considering scale validity becomes, this way, very subjective. Indeed, Kerlinger & Lee (2000) pointed out that each item of the measure is judged according to its relevance. This relevance is only assumed and hypothetical. This form of validity seems very difficult and requires good sampling (DeVellis, 2003). The second type of validity is purely quantitative and seeks to determine the predictive power of a measure. Here, the aim is to establish the relationship between measures and constructs by demonstrating a correlation between items (Straub, Boudreau & Geffen, 2004). A correlation that does not necessarily require a causal relationship (DeVellis, 2003). The third type of validity is very important and is called construct validity. The idea here is to know whether the selected measures agree with each other (Straub, Boudreau & Gefen, 2004). Several methods exist. This is the case with convergent validity, and its inverse, divergent validity. These two forms test the extent to which we are able to associate (or differentiate) the construct with other constructs that are considered similar (Kerlinger & Lee, 2000). In other words, we seek to know whether different measures of the same construct converge (or diverge) together (DeVon, Block, Moyle-Wright, Ernst, Hayden,

Lazzara, Savoy & Kostas-Polston, 2007). We can also point out factorial and nomological validity; the former refers to factorial analysis techniques and the latter seeks to test whether relationships exist between a construct and its antecedents and predictors (Lewis, Templeton & Byrd, 2005).

2.2.3. Research on the relation between lived and self-reported experience

Lived experience refers to data collected implicitly (implicit measures), whereas the self-reported experience refers to explicit measures (Riedl, Léger, 2016). Studying this relationship is not new, for example, the works of Cannon (1987) highlighted the weaknesses of James-Lange's theory of emotions. However, no real consensus seems to emerge in contemporary literature. Some studies have tried to find out whether this relationship is sufficient to claim that one way of measuring would be alternative to the other.

Some studies suggest that this relationship is weak. Alpers & Sell (2008) tried to find out the relationship between self-reported fear and physiological arousal, through a study carried out on 10 people. The authors observed only weak concordances between these two measures (Alpers & sell, 2008). In the same field, Ordoñana, González-Javier, Espín-López & Gómez-Amor (2009) tried to understand the relationship between self-reported fear and physiological arousal (skin conductance and heart rate). They neither found strong enough evidence to affirm a strong relationship (Ordoñana & al., 2009).

For others, the opposite appears to be the case. Lang & al. (1993) attempted to identify these relationships by asking several people to respond to a questionnaire after viewing several images. Their responses were confronted with their psychophysiological reactions. The authors found a relationship between self-reported valence and their facial expressions, but also between their self-reported activation and the conductance of their skin (Lang & al., 1993).

Each type of measurement would capture different facets of the same construct, a complementarity between them can be assessed. Tams, Thatcher, Hill, Grover & Ortiz de Guinea (2014) raised the question of the extent to which implicitly measuring an

experience could be considered an alternative to explicitly measuring it. Their field of study was technostress and their results suggest, on the contrary, a complementarity between the two methods (Tams & al., 2014). Other studies seem to go in the same direction (Tomarken, 1995; Li, Walters, Packer, Scott, 2018). Relations between what a person experiences and what he or she perceives to have experienced seems thus really existing (Maia & Furtado, 2019). Although this relationship would not be similar depending on the construct of interest or the type of experience offered to a participant (Le Pailleur, Huang, Léger & Sénécal, 2020). While some constructs would not suffer from single-method bias (use of a single method to measure it), others would. This would be the case for constructs considered less complex than others, using one method would be effective; for others, this effectiveness would be less clear (Ortiz de Guinea, Titah & Léger, 2013).

We are all probably familiar about this theory in which we remember the beginning and the end of a book, a song, a meal, but not the middle. This theory is called the serial position effect (Murdock, 1962) and involves both short- and long-term memory (Martin, Carlson, Buskist, 2010). More specifically, a person's memory (when memorizing words, reading a book, having a conversation, etc.) forms a curve (Murdock, 1962): This person will remember what happened first (primacy effect), his or her memory will deteriorate later, only to improve at the end of the interaction (recency effect). Several studies have looked at this theory and suggested some strength in either the first impression (DiGirolamo & Hintzman, 1997; Li, 2009; Lindgaard, Fernandes, Dudek & Brown, 2006), the last impression (Cockburn, Quinn & Gutwin, 2017; Hassenzahl & Sandweg, 2004; Barnes, 1992; Bergeron, Fallu & Roy, 2008; Hansen & Danaher, 1999; Wang, 2011), or both (Lourties, Léger, Sénécal, Fredette & Chen, 2018; Murphy, Hofacker & Mizerski, 2006). It is thus possible that this relationship may vary at different moments of an interaction. We therefore postulate the following hypothesis:

H1: The relationship between lived experience (emotional and cognitive) and perceived experience (emotional and cognitive) is likely to differ between the overall level, first and last impressions.

2.2.4. Effects of specific contexts

There seems to exist a link between the attributes that constitute a system and the resulting perceptions. Thielsch, Blotenberg & Jaron (2014) took an interest in the subject and, through 3 studies, noted that both aesthetic and informational features have an influence on the way users perceive their experience. This influence depends on the moment of the interaction and the construct of interest (Thielsch, Blotenberg & Jaron, 2014). In our study, we are interested in elements considered as hedonic (aesthetic elements) and features considered as utilitarian (pragmatic). The hedonic attributes correspond to those that cause a certain well-being, as opposed to the utilitarian attributes that allow completing a goal: "… *a product may be perceived as pragmatic because it provides effective and efficient means to manipulate the environment. A product may be perceived as hedonic because it provides stimulation, identification or provokes memories*." (Hassenzahl, 2003, p. 36). Other researches emphasize the importance of aesthetic elements. Aesthetics increase the pleasure resulting from an interaction (Cai & Xu, 2011), and thus have a positive influence on valence (Seo, Lee, Chung & Park, 2015). Research by Thüring & Mahlke (2007) seems to go in this direction, adding emotional activation that would be positively impacted by these aesthetic aspects. According to Tractinsky, Katz & Ikar (2000), the satisfaction a person would have when he or she interacts with an interface would also be impacted by these aesthetic attributes, indirectly. Retention should also be influenced by these elements. Here again, indirectly, both utilitarian and hedonic elements would influence the intention to reuse a website or service (Bilgihan & Bujusic, 2015). One strength of a measure may refer to its ability of distinguishing differences in contexts, its sensitivity. And it is possible that when the experience is lived through different hedonic elements, the resulting perceived experience may be impacted. In our case, and considering these arguments, we posit the following hypothesis:

H2: The relationship between the lived experience (emotional and cognitive) and the perceived experience (emotional and cognitive) is stronger for website hedonic features than utilitarian features.

2.3 Method

2.3.1. Experimental design

In order to answer our research questions and test our hypotheses, we used a correlational design through which the participants lived different experiences on several websites in the same industry. Although they were given several tasks, participants were free to browse the websites assigned to them. Each participant therefore followed a navigation path that was unique to them and thus, no experience was the same as the other. Since each website is different and contains different characteristics, the stimuli contained naturally hedonic (not necessary for navigation) and utilitarian (necessary for navigation). Participants performed three tasks on two different websites (see Procedures)". Figure 2.1 shows the data collection procedure. In order to ensure that each "pair of websites" was correctly randomized, we used a Latin Square (Easterling, 2015). Using this allowed us to ensure that each website was used in the same order.

2.3.2. Industry

According to Statista (2020), more than half (53%) of Canadians used banking websites to carry out the majority of their banking transactions. The data were originally collected by a survey on 4000 people living in Canada between 2012 and 2018 (Statista, 2020). Considering the measures we wanted to use, and the fact that we needed relatively short and balanced between hedonic and utilitarian features online tasks, we decided to use a business's specific domain. The business sector, and therefore the financial industry seemed to be the most adequate solution. We also chose to focus on the most of retail banking institutions in Canada. Our goal was to obtain a maximum of diversity and variance between the different websites. We wanted to explore across the entire industry and avoid having the results associated with a specific financial institution.

2.3.3. Sample

Participants were recruited through our university's research panel. They received a $20 gift card as a compensation. The conditions of participation were that participants had to be of legal age and did not have to meet any of the following conditions: astigmatism, laser-eye correction, neurological or psychiatric diagnosis, epilepsy and skin allergies. We made sure of this before calibrating each tool. Forty people participated in this experiment, 26 women and 14 men (mean age = 25.025 years old). This study was approved by our institution's Ethic's Research Committee.

2.3.4. Procedure

A pretest with 4 participants was conducted in order to ensure the quality of the data on one hand and a good flow of the scenario, on the other hand. Before beginning the experiment, each participant was welcomed by a research assistant. He or she was then asked to fill out the consent form and once it was completed, the assistant placed the sensors on the participant. The experiment could then begin. Each experiment started in the same way, with a baseline. This consisted of calculating the number of white squares in an animation of 1'30 minute long. Using this simple task allowed us to know the psychophysiological state of each participant before being confronted to the stimuli. The three tasks were all related to mortgage and were completed in the same order. They consisted first of finding a way to access a mortgage pre-approval form, second filling out the form, and third, finding a way to contact a mortgage specialist. And each experiment ended with a demographic questionnaire, the removal of the tools and the compensation.

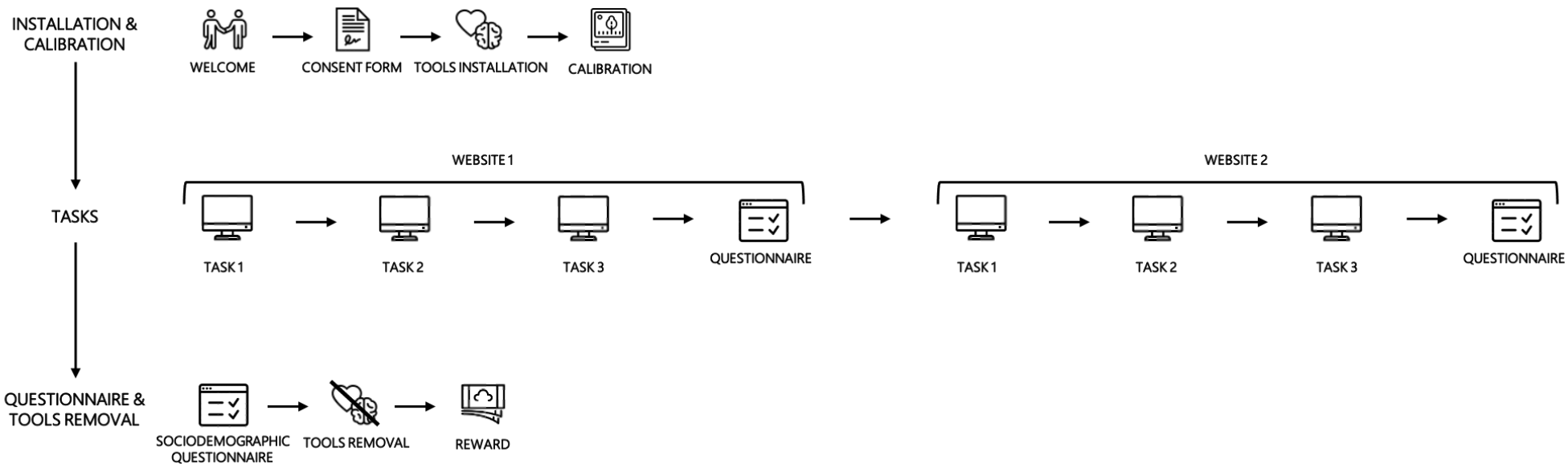


Figure 2.1. Data collection procedure

2.3.5. Measures

Measures of the lived experience. These are implicit measures, referring to the emotional and cognitive experience of the user. Valence was measured using an automatic facial analysis software (FaceReader© version 6.0, Noldus, Wageningen, Netherlands). This software calculates, by analyzing the micromovements of the facial muscles, the intensity of positive emotions from which the negative emotion with the highest intensity is subtracted (Skiendziel, Rösch & Schultheiss, 2019; Loijens, Krips, Grieco, Van Kuilenburg, Den Uyl & Ivan, 2016). This corresponds to the valence. Arousal was measured in two different ways: the measurement of the electrodermal activity of the skin and the measurement of the cardiac activity. The electrodermal activity (EDA) corresponds to the measurement of the skin conductance. EDA is related to the sympathetic nervous system : this is an automatic response to several different situations (Reidl & Léger, 2016). It is common to infer an arousal state by measuring the skin conductance (Dawson, Shell & Fillion, 2007). Different placements are usually accepted (Dawson, Schell & Fillion, 2007): two placements on the index and middle fingers (volar surfaces of the distal or medial phalanges) and one placement on the thenar and hypothenar eminences of the palm of the hand. Since our participants were going to have to use the computer keyboard, we followed the general guidelines in our field by placing the electrodes on the palm of the non-dominant hand. Cardiac activity (ECG) is related both to the sympathetic and parasympathetic nervous systems and it is also possible to infer the arousal by measuring the heart rate variability (Riedl, Léger, 2014). This was measured by placing 3 electrodes : 2 beneath the clavicles and one on the left side of the ribs, following the common guidelines. EDA and ECG were recorded using the Acqknowledge© software, version 4.3 (Biopac, Goleta, USA). The devices used to record these activities were Biopac© MP150, sampled at 500 Hz. The pupil variation is also a response both from the sympathetic and parasympathetic nervous systems and is related to an increased attention (Riedl & Léger, 2016). Researches suggest that the cognitive load can be inferred by measuring the pupil dilatation (Léger, Charland, Sénécal & Cyr, 2017). This was measured using a Tobii© X-60 eye tracker sampled at 60 Hz and recorded using Tobii Studio© software version 3.4.8 (Stockholm, Sweden). All our measures were synchronized every 120 seconds using the Observer XT© software, version 11.5.

Measures of the perceived experience. We used four different self-reported single-item measures : Affective Slider, CES, CSAT and NPS, presented in Table 1. Since none of the participants responded above 8 on the NPS, we did not calculate it the conventional way, but rather considered it as a classic scale. At the end of the experiments, participants were asked to complete a demographic questionnaire in order to better understand their profile. The questions asked were about age, gender, marital status, education, residence and technology use.

Measure	Label	Reference
Affective Slider	How do you feel about your experience? (Affective_1) Please, drag the cursor to represent your level of pleasure (unhappy-happy).  (Affective_2) Please, drag the cursor to represent your level of arousal (calm-excited). 	Betella, Verschure, 2016
Net Promoter Score	(NPS) Based on your experience today, how likely is that you would recommend this institution to a friend or a colleague? Please, give a response ranging from 0 to 10 (0 = extremely unlikely; 10 = extremely likely)	Reichheld, 2003
Customer SATisfaction score	(CSAT) Please, indicate your level of satisfaction with your experience today. Please, give a response ranging from 1 to 5 (1 = very dissatisfied; 5 = very satisfied)	
Customer Effort Score	(CES) How much effort do you think it took to apply for a mortgage pre-approval? Please give a response ranging from 1 to 5 (1 = very low; 5 = very high)	Dixon, Freeman et Toman, 2010

Table 2.1. Psychometric measures. Please note that the questions were asked in French in our study. The images of the Affective Slider measure were adapted from Betella and Verschure (2016).

2.3.6. Data manipulation

In order to prepare our analyses, we had to carry out several transformations. The first one was the codification of the different tasks and the different moments to keep for our analyses by naming them. To measure the overall experience, we decided to consider the series of three tasks a whole, excluding moments when participants were reading the

instructions. In addition, since we were also interested in first and last impression biases, we decided to codify the first and last pages of each website used. The first impression is therefore defined as the first page used and the last impression is the last page used. The second transformation was the creation of areas of interest (AOI) on each page of the first and last impression. The purpose of these AOIs is to better understand which attributes seem to play a role in the perception of an interaction. We defined two types of attributes, following the guidelines provided by Hassenzahl (2003): pragmatic (utilitarian) and hedonic attributes. A pragmatic attribute is purely functional and consists of a way to achieve a very specific goal. It is usually a navigation menu, a text block explaining how to fill in a form or a button. Hedonic attributes have no functional role but have rather an emotional role. They are, for the most part, photographs or text blocks whose intention is to maintain (or improve) the brand image of the website used. In total, nine pages of the first impression (of eight different sites) and 28 pages of the last impression (divided into seven sites) have been coded. Figure 2 illustrates the coding strategy. The psychophysiological data were compiled, triangulated and exported using CubeHX©'s software (Montreal, Canada) (Léger, Courtemanche, Fredette & Sénécal, 2019). Exportation window was defined at 250 ms. Exporting this way allowed us to have four psychophysiological data points per second, per participant.

Figure 2.2. Example of AOIs drawn in a website page used. The colors were randomly displayed by the software.

Statistical analyses were performed using SAS© software, version 9.4 (SAS Institute, Cary, USA). Normality test was calculated for each psychometric variable: Affective_1 (p=0.0781), Affective_2 (p=0.011), CES (p=0.0001), CSAT (p=0.0001), NPS (p=0.0015). They all failed to normality tests and therefore do not seem to follow a normal law. Thus, we decided to calculate Spearman's correlations. The psychophysiological data were aggregated by website level. Our analyses are based on correlations between perceived and lived experience. Since calculating correlations do not allow taking into account the repeated nature of the psychophysiological measures, we decided, to overcome this issue, to take only the first website used by each participant.

2.3.7. Data loss

During data collection, the software used to capture electrodermal activity bugged and forced us to remove two participants of the analyses. Two websites and the last page used for another one had to be dropped since the compatibility between the browser and the software used for the recording did not allow us to make a clear and accurate coding of the AOIs.

2.4 Results

2.4.1. Descriptive statistics

Psychometric variables. As can be seen in **Table 2.2**, we can observe that the perceived pleasure (Affective_1) has a mean of 45.03, a median of 43, a minimal value of 6 and a maximal value of 89. Mean of perceived arousal (Affective_2) is 51.55, median is 55, minimal value of 2 and maximal value of 88. Cognitive load (CES) has a mean of 2.97, a median of 3, a minimal value (low perceived cognitive load) of 1 and a maximal value (high perceived of cognitive load) of 5. The NPS has a mean of 4.84, a median of 5, a minimal value equal to 0 and a maximal value equal to 8. Finally mean and median of satisfaction (CSA) have a value of 3, a minimal value of 1 and a maximal value of 5.

Affective_1		Affective_2		CES		NPS		CSAT	
Mean	45.03	Mean	51.55	Mean	2.97	Mean	4.84	Mean	3
Median	43	Median	55	Median	3	Median	5	Median	3
Min.	6	Min.	2	Min.	1	Min.	0	Min.	1
Max.	89	Max.	88	Max.	5	Max.	8	Max.	5

Table 2.2. Descriptive statistics of the psychometric variables

Psychophysiological variables. Valence has a mean of -0.0009 at the overall level, 0.0005 during the first impression (hedonic: -0.0062; utilitarian: 0.0115) and -0.0108 during the last impression (hedonic: -0.1025; utilitarian: -0.0343). The mean of EDA has been calculated at 1.0626 at the overall level, 1.6019 during the first impression (hedonic: 1.7907; utilitarian: 1.6831) and at 1.0151 during the last impression (hedonic: 1.1145; utilitarian: 1.0634). The mean of the ECG is -0.2583 at the overall level, 0.6656 during the first impression (hedonic: -0.8935; utilitarian: 0.2514) and -0.1601 during the last impression (hedonic: -2.9234; utilitarian: -0.1106). There is an average pupil dilatation of 0.0104 at the overall level, 0.2344 during the first impression (hedonic: 0.2919; utilitarian: 0.2467) and 0.0664 during the last impression (hedonic: 0.0099; utilitarian: 0.0649). Complete descriptive statistics can be seen in **Table 2.3** and **Table 2.4**.

Variable	Moment	N	Mean	Stddev	Min	Max
Valence	O	31	-0.0009	0.1118	-0.3161	0.1874
	F	31	0.0005	0.1285	-0.2403	0.4178
	L	26	-0.0108	0.1294	-0.2219	0.3760
Arousal (EDA)	O	30	1.0626	1.6691	-2.2265	5.8354
	F	30	1.6019	1.5168	-0.4469	5.2683
	L	26	1.0151	2.0424	-3.1960	6.3907
Arousal (ECG)	O	31	-0.2583	3.5424	-5.9039	8.3119
	F	31	0.6656	4.7533	-9.7765	10.2594
	L	27	-0.1601	4.2328	-8.1693	7.9637
Cognitive load	O	31	0.0104	0.1636	-0.2227	0.4410
	F	31	0.2344	0.2249	-0.2245	0.6826
	L	27	0.0664	0.2160	-0.2203	0.5642

Table 2.3. Descriptive statistics of the implicit measures without considering the website attributes. "O" is referring to the overall interaction, "F" and "L" are respectively referring to the first and last impressions.

Variable	Feature	Moment	N	Mean	Stddev	Min	Max
Valence	H	F	22	-0.0062	0.1286	-0.1592	0.4147
		L	5	-0.1025	0.1450	-0.3267	0.0025
	U	F	25	0.0115	0.1291	-0.1624	0.4355
		L	23	-0.0343	0.1245	-0.3306	0.2261
Arousal (EDA)	H	F	21	1.7907	1.6328	-0.5274	5.1305
		L	7	1.1145	1.2596	-0.9899	2.7054
	U	F	24	1.6831	1.7368	-1.3830	6.0048
		L	23	1.0634	2.1570	-2.9983	6.2755
Arousal (ECG)	H	F	22	-0.8935	4.9152	-10.3122	10.0128
		L	7	-2.9234	4.9139	-7.4744	6.9099
	U	F	25	0.2514	6.2714	-15.7212	11.8700
		L	24	-0.1106	4.4706	-7.6268	10.0092
Cognitive load	H	F	22	0.2219	0.2559	-0.2276	0.7082
		L	7	0.0099	0.2077	-0.3425	0.2197
	U	F	25	0.2467	0.2365	-0.2357	0.6558
		L	24	0.0649	0.2460	-0.2904	0.6824

Table 2.4. Descriptive statistics of the implicit measures during the different contexts of use: hedonic ("H") and utilitarian ("U"). "F" and "L" are respectively referring to the first and last impressions.

2.4.2. Relationship between lived and perceived experience

Our first hypothesis was to assume relationships between the lived and the perceived experience, relationships which are different between the overall, the first and the last impression levels. In order to answer our hypothesis as precisely as possible, correlation tests were carried out, following the guidelines provided by Steiger (Steiger, 1990). All the results of our correlations can be found in the appendices (Appendix A and B). The graphs used in this section (**Figures 2.3** and **2.4**) are presented in a visual way in order to better understand the different patterns between our variables.

As suggested by **Figure 2.3**, when we look at the overall experience, we can observe that several perceptual variables seem to be in relationship with its direct corresponding implicit measure. Perceived arousal correlates with electrodermal activity (0.3356; $p = 0.0698$), shown in cell 4. Both ways of measuring cognitive load correlate together (-0.324; $p = 0.0754$), cell 8, but in a negative direction. We can also observe that satisfaction is in a positive correlation with cognitive load (cell. 10, 0.4155; $p = 0.0201$).

Looking at first and last impressions, and starting with the first impression, perceived valence seems to be in negative relationship with lived valence, visible in cell 1 (-0.3426; p = 0.0592). This would indicate that the more the lived valence increases, the more the perceived valence decreases. Perceived arousal seems to be in relationship with both implicit ways of measuring arousal : EDA and ECG, respectively visible in cells. 5 (0.3771; p = 0.04) and 6 (0.3655; p = 0.0432). Two correlations were also observed between satisfaction and valence, in cell 9 (-0.3798; p = 0.0351) and cognitive load, in cell. 11 (0.502; p = 0.004). Finally, we can also mention one last correlation between the intention to recommend and valence, visible in cell 12 (-0.3475; p = 0.0554). No correlation was, however observed for the last impression.

In order to answer our hypothesis as precisely as possible, we wanted to know if our correlations were significantly different. Correlation tests were therefore carried out, following the guidelines provided by Steiger (Steiger, 1990). Two relationships deserve to be analyzed: arousal (implicit and explicit) between overall level and first impression and the correlation containing cognitive load and satisfaction between overall level and first impression. The first relationship is not statistically significant but the second one is (p = 0.0067). We are therefore not able to provide enough support for H1.

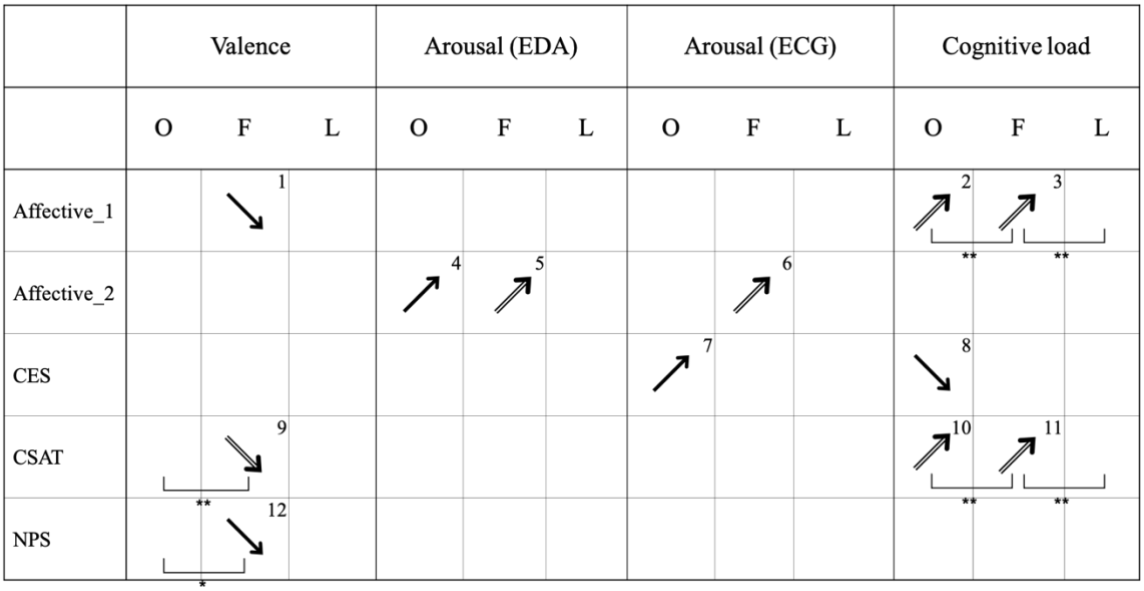


Figure 2.3. Overview of the correlations between our variables (H1). The directions of the arrows represent the direction of the correlations. The thickness of the arrows is representing the level of significance (thin arrows: p<0.10; thick arrows: p<0.05). The hooks are representing the level of significance between the correlations (*p<0.10; **p<0.05; *** p<0.0001). "O" is referring to the overall level, "F" and "L" are respectively referring to the first and last impressions.

2.4.3. Effects of context

Our second hypothesis was assuming stronger relationships in hedonic contexts. In order to know if these relationships were different, correlation tests were also carried out. Attention, however, considering the low number of data points of hedonic attributes during the last impression, precautions must be taken when drawing conclusions.

We observed a direct negative relationship between the two ways of measuring valence during the first impression, show in cell 1 (-0.6084; p = 0.0027) for the hedonic attributes and for the utilitarian attributes, cell 4 (-0.4831; p = 0.0144). The perceived arousal seems to correlate positively with the lived arousal (EDA) for the hedonic attributes during the last impression, shown in cell 8 (0.8289; p = 0.0212); and for the utilitarian attributes during the first impression, cell 11 (0.3449; p = 0.0989). A correlation was also observed for these same attributes, during the same moment between reported arousal and lived arousal, measured with ECG, cell 12 (0.4591; p = 0.0210). A direct negative correlation was also observed for the cognitive load, in utilitarian attributes present during the first impression, cell 14 (-0.41128; p = 0.0411). Satisfaction is in negative relationship with valence at several levels : first impression hedonic (cell 15, -0.5536; p = 0.0075) and utilitarian (cell 18, -0.3993; p = 0.048) attributes, and last impression for hedonic attributes, cell 16 (-0.9747; p = 0.0048). A positive relationship was also observed between satisfaction and cognitive load during the first impression both for hedonic (cell 17, 0.5425; p = 0.0091) and utilitarian (cell 19, 0.4731; p = 0.0169) attributes. Finally, we also observed relationships between the NPS and implicit measures, starting with valence at several levels: first impression both for hedonic (cell 20, -0.6294; p = 0.0017) and utilitarian (cell 23, -0.4125; p = 0.0404) attributes, and at last impression level for hedonic attributes (cell 21, -0.9747; p = 0.0048). Intention to recommend seems also being in relationship with arousal (EDA) during last impression for utilitarian attributes (cell 24, 0.3794; p = 0.0742). This last construct is also in positive relationship with cognitive load at first impression level for the hedonic attributes (cell 22, 0.3973; p = 0.0671) and utilitarian attributes (cell 25, 0.4774; p= 0.0158). In addition, we observed several other correlations between lived and perceived experience from different contexts. These correlations can be seen in **Figure 2.4**.

Concerning differences of correlations, three relationships would seem to be statistically different, each containing the implicit way of measuring valence at the first impression and between hedonic and utilitarian attributes: perceived valence ($p = 0.01250$), satisfaction ($p = 0.00344$) and intention to recommend ($p < 0.001$). Considering that hedonic attributes seem to have a greater correlation coefficient for each perceived measure, this would suggest that there is an effect from the hedonic attributes present during the first impression. However, these correlations are negative. We have thus not enough evidence to provide support for H2.

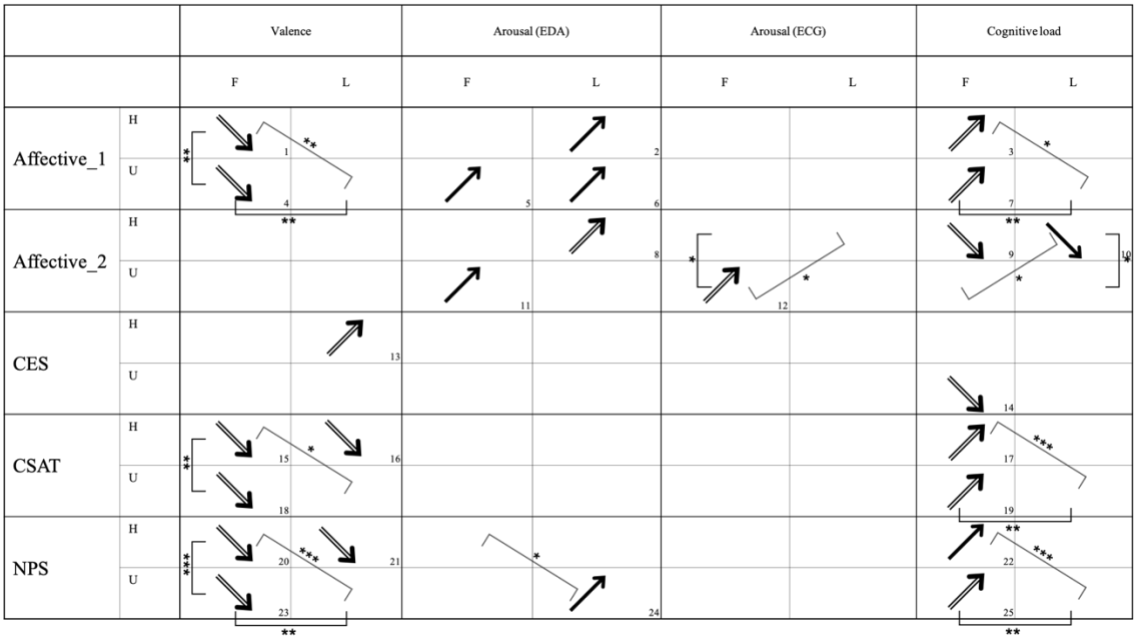


Figure 2.4. Overview of the correlations between our variables considering AOIs (H2). The directions of the arrows represent the direction of the correlations. The thickness of the arrows is representing the level of significance (thin arrows: p<0.10; thick arrows: p<0.05). The hooks are representing the level of significance between the correlations (*p<0.10; **p<0.05; *** p<0.0001). "H" is referring to hedonic and "U" is referring to utilitarian.

2.5 Discussion

The aims of our research were threefold. Our first objective was to know to what extent the single-item measures of perceived experience were related to the implicit ones. The second objective was to find out if, at the first or last impression, this relationship was different. Finally, we wanted to know if the relationships would be different considering the experience lived throughout different contexts (hedonic and utilitarian contexts). In other words, we wanted to explore these relationships.

In answer to RQ1, there appears to exist some kind of relationships between what our participants experienced and what they reported experiencing. At the overall level, we found that the level of perceived arousal and the level of perceived cognitive load appear to be consistent with their direct implicit measures. However, the relationship between both ways of measuring cognitive load is not behaving in a normal way: this correlation is negative. We also found that satisfaction is in relationship with cognitive load in a positive way.

These same relationships differ depending on the moment of the interaction (RQ2). Indeed, we observed several primacy and recency effects. For example, the perceived overall level of pleasure did not correlate with its implicit correspondence: valence. However, when we only consider the first and the last pages used, it correlates. That depends on the psychometric variables used. Overall, we do not seem to observe enough evidence to provide support for H1.

Our study also highlights different relationships when the experience is lived across specific contexts (RQ3). Our second hypothesis was to assume stronger relationships when the experience is lived through hedonic website features. However, we observed that both utilitarian and hedonic contexts seem to correlate with single-item scales. This is depending on the variables of interest. If we count them, we can see that hedonic contexts seem to lightly better correlate with perceived measures. 13 correlations have been observed when taking into account the hedonic features, against 12 for the utilitarian attributes. In this sense, it would appear that the experience lived through

hedonic features better correlate to single-item scales. However, we are not able to provide support for H2, given the small number of correlations.

Further analysis showed us several differences in correlations. They are either between different forms of attributes for the same moment of the interaction, or between different moments of the interaction, or between different attributes, at different moments. Although the vast majority of our correlation differences appear between a significant and a non-significant correlation, we can observe differences between moments of an interaction (cf. **Figure 2.3**) or between different features (cf. **Figure 2.4**).

2.5.1. Correspondences and differences with other studies

We wanted to investigate the relationships between two ways of measuring an interaction. On the one hand, they would be too general to capture the full extent of a construct (Bergkvist, & Rossiter, 2007; Jacoby, 1978). On the other hand, however, having too many items produce the opposite effect: too many elements are captured at the risk of being redundant (Boyle, 1991). In our study, we observed that this kind of scale would be able to capture parts of a construct. However, their behaviour does not allow us to agree or disagree with either point of view.

Our results appear to be consistent with those of other previous studies. Ortiz de Guinea, Titah & Léger (2013) had observed certain correspondences between different ways of measuring their constructs. In particular, they highlighted certain relationships between self-reported arousal and arousal measured by electrodermal activity and heart rate. This relationship exists both considering the overall experience (EDA) and at first impression level (EDA and HR). This last observation seems to converge with the observations of Lourties & al. (2018). They had tried to find out whether the continuous emotional reporting using a single-item scale was subject to primacy and recency effects. They reported finding some influences of these effects: a primacy effect on valence and a recency effect on emotional arousal (Lourties & al., 2018). Our data behaved in part the same way, for emotional arousal. When taking website features into account, we can see that these relationships change, and a last impression effect appears on the relationship between perceived arousal and lived arousal (EDA) when the experience is lived

throughout hedonic features. Regarding valence, there also seems to be a serial position effect in our data, although it is not differentiated between first and last impressions. This difference could come from the way we measured perceptions: at the end of the interaction.

On the other hand, it would seem that our results are in contradiction with other studies. Cai & Xu (2011) highlighted the importance of certain aesthetic aspects in the perception of pleasure. Certainly, these aspects play a certain role. But in our case, it seems that other elements have an impact, sometimes just as important.

Analyzing directions of our correlations allowed us to observe some strange findings: they almost all behave in an opposite way. Almost all of them except those between perceived and lived arousal (EDA and HR). Perceived pleasure seems to negatively correlate with lived pleasure; CES negatively correlates with cognitive load. CSAT seems to behave in the same way: negatively correlated with valence but positively correlated with cognitive load. A similar finding for the NPS: a negative correlation with valence. Le Pailleur & al. (2020) have observed this phenomenon in their research. This direction could be explained by a certain bias from which self-reported measures of emotions suffer: social desirability bias (Tams, Thatcher, Hill, Grover, Ortiz de Guinea, 2014). Another possible explanation lies in the particular context of our research and mainly in the tasks carried out by our participants: these were purely utilitarian tasks on financial institutions. It is possible that our participants expected to perform tasks of some complexity. These directions, however, give us some indications of the weakness of single-item scales. They may not represent an entire construct (Baumgartner & Homburg, 1996). Asking one and only one question when thinking of measuring a construct does not seem to us, therefore, to be valid or reliable in our context. Other factors are very likely to exist when reporting emotions.

2.5.2. Managerial implications

Our study has a main managerial implication: it highlights the limit of single-item scales. Explicit measures allow us to know what users remember about their feelings, how they analyzed their emotions after interacting with an interface. Implicit measures allow us to know what a person experiences throughout the interaction. Our results raise an important point, we are not able to systematically observe what we should expect from them: their behaviour in relation to what people actually experience during their interaction. It seems important to us to distinguish two levels of analysis: using two methods to measure the same construct (for example: lived valence and perceived valence) and using two methods to know to what extent one is able to predict the other. When it comes to measuring the same construct in two different ways, we found that the single-item scales used in our study did not behave logically. This observation seems valid for the level of perceived pleasure as well as for the perceived cognitive load. The only measure that seems to be truly related to what a person experienced during the interaction is perceived arousal. Regarding the prediction of one with the other, we were also unable to obtain sufficient clues to affirm that satisfaction or intention to recommend being strongly enough related to the lived experience. It would appear that these measures do not provide the most complete picture of what happened during an interaction. We are not claiming that either method is better, we are simply saying that they should not replace implicit measures. We believe it is important to emphasize the importance of using implicit measures in order to obtain the most complete data possible, the importance of using two methods rather than one.

2.5.3. Limits and research opportunities

Our research has its limits. The first one comes from the operationalization of our data: we have codified the AOIs following specific criteria (Hassenzahl, 2003). Very few hedonic attributes are present on the last impression pages. These last pages were in large majority pages of confirmation of appointment scheduling. What if the criteria were different? What if there were more hedonic elements? These questions deserve to be raised and studied through further research. It is possible that the hedonic effect, naturally

present on the websites used in our study, is not prominent given the tasks completed by the participants and the financial context of use. It would seem appropriate to replicate this type of study with another type of stimulus, choosing ones that provide more emotional reactions (video games for example). The second limitation comes from our analyses. We wanted to explore these relationships between lived and perceived experience, using single-item scales only. We observed certain absences of relationships. However, we have not used longer scales to compare them that would allow us to affirm that single-item scales should really be avoided. It would be important to continue the research and compare them to longer scales. Finally, the vast majority of our participants belong to a relatively similar demographic profile, especially in terms of age. There is evidence of an age effect with respect to primacy and recency effects (Griffin, John, Adams, Bussell, Saurman & Gavett, 2017). It is therefore important to investigate what these effects would be on people of different age groups.

In conclusion, we can state that the general criticism proposed by several authors regarding single-item scales would not seem to be fully verified. At the very least, in our case and in our context. We were able to observe several relationships, sometimes relatively strong, between what a person experienced during his or her interaction and what he or she reported having experienced. But the directions of our correlations indicate a certain weakness of single-item scales, a lack of representativeness of our constructs of interest. Thus, we have not been able to find any real agreement, any real confirmation between the use or not of this type of scale. They would seem to be sufficient in some cases, but not in others.

Appendices

Appendix A : Correlations between implicit and explicit measures (no website attributes).

	Valence			Arousal (EDA)			Arousal (ECG)			Cognitive load		
	O	F	L	O	F	L	O	F	L	O	F	L
Affective_1	-0.1082	-0.3426 *	-0.2359	0.3048	0.2407	0.3037	-0.1994	-0.1348	-0.1864	0.4253**	0.4705*	0.1760
Affective_2	-0.0636	0.0202	-0.0390	0.3356*	0.3771**	0.2918	0.0832	0.3655**	0.1990	-0.0085	0.0632	0.0911
CES	0.0563	0.1500	0.0756	-0.0968	0.0336	-0.1201	0.3175*	0.2675	0.2218	-0.3240*	-0.2909	-0.2243
CSAT	-0.1004	-0.3798**	-0.3279	0.1973	0.0782	0.1745	-0.1056	-0.1456	-0.0013	0.4155**	0.5020*	0.1922
NPS	-0.0737	-0.3475*	-0.2437	0.2415	0.0407	0.3128	-0.1189	-0.1169	-0.1447	0.2468	0.2411	0.0253

Appendix B: Correlations between implicit and explicit measures, considering website attributes

		Valence		Arousal (EDA)		Arousal (ECG)		Cognitive load	
		F	L	F	L	F	L	F	L
Affective_1	H	-0.6084**	-0.7000	0.3454	0.7208*	-0.0656	0.0000	0.4275**	0.0360
	U	-0.4831**	-0.1723	0.3467*	0.3693*	-0.1633	-0.1794	0.5085**	0.1977
Affective_2	H	0.0232	-0.6669	0.2758	0.8289**	0.1227	-0.0541	-0.0079	-0.6847*
	U	-0.0258	-0.1262	0.3449*	0.3411	0.4591**	0.1860	0.0963	0.0344
CES	H	0.1390	0.8944**	-0.0865	-0.1594	0.0327	-0.0591	-0.3498	0.2561
	U	0.1368	0.2241	-0.1287	-0.2364	0.2499	0.2399	-0.4113**	-0.2901
CSAT	H	-0.5536**	-0.9747**	0.2223	0.4117	0.0192	0.3368	0.5425**	0.0749
	U	-0.3993**	-0.3051	0.1500	0.1543	-0.2267	-0.0516	0.4731**	0.1637
NPS	H	-0.6294**	-0.9747**	0.1208	0.1091	-0.2087	0.1081	0.3973*	0.0360
	U	-0.4125**	-0.0609	0.2014	0.3794*	-0.1309	-0.2065	0.4774**	0.0448

References

Alpers, Georg W. et Roxane Sell (2008). « And yet they correlate: Psychophysiological activation predicts self-report outcomes of exposure therapy in claustrophobia », *Journal of Anxiety Disorders*, vol. 22, no 7, p. 1101-1109.

Bain & Company (2020). *Bain & company - measuring your net promoter score - net promoter system*. Récupéré le 23 novembre 2020 de https://www.https://www.netpromotersystem.com/about/

Barnes, G. Michael (1992). « Digitized speech's serial position effect », communication présentée au *1992 SIGCHI Conference on Human Factors in Computing Systems*, 1992, Monterey, California.

Baumgartner, Hans et Christian Homburg (1996). « Applications of structural equation modeling in marketing and consumer research: A review », *International Journal of Research in Marketing*, vol. 13, no 2, p. 139-161.

Bergeron, Jasmin, J. M. Fallu et Jasmin Roy (2008). « Une comparaison des effets de la première et de la dernière impression dans une rencontre de vente », *Recherche et Applications en Marketing*, vol. 23, no 2, p. 19-36.

Bergkvist, Lars et John R. Rossiter (2007). « The predictive validity of multiple-item versus single-item measures of the same constructs », *Journal of Marketing Research*, vol. 44, no 2, p. 175-184.

Betella, Alberto et Paul F. M. J. Verschure (2016). « The affective slider: A digital self-assessment scale for the measurement of human emotions », *PLoS ONE*, vol. 11, no 2.

Bilgihan, Anil et Milos Bujisic (2015). « The effect of website features in online relationship marketing: A case of online hotel booking », *Electronic Commerce Research and Applications*, vol. 14, no 4, p. 222-232.

Boyle, Gregory J. (1991). « Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? », *Personality and Individual Differences*, vol. 12, no 3, p. 291-294.

Brünken, Roland, Jan L. Plass et Detlev Leutner (2003). « Direct measurement of cognitive load in multimedia learning », *Educational Psychologist*, vol. 38, no 1, p. 53-61.

Cai, Shun et Yunjie Xu (2011). « Designing not just for pleasure: Effects of web site aesthetics on consumer shopping value », *International Journal of Electronic Commerce*, vol. 15, no 4, p. 159-188.

Cannon, W. B. (1987). « The james-lange theory of emotions: A critical examination and an alternative theory. By walter b. Cannon, 1927 », *The American journal of psychology*, vol. 100, no 3-4, p. 567-586.

Chang, En-Chung et Xiaomeng Fan (2013). « More promoters and less detractors: Using generalized ordinal logistic regression to identify drivers of customer loyalty », *International Journal of Marketing Studies*.

Churchill, Gilbert A. (1979). « A paradigm for developing better measures of marketing constructs », *Journal of Marketing Research*, vol. 16, no 1, p. 64-73.

Churchill, Gilbert A. et Carol Surprenant (1982). « An investigation into the determinants of customer satisfaction », *Journal of Marketing Research*, vol. 19, no 4, p. 491-491.

Cockburn, Andy, Philip Quinn et Carl Gutwin (2017). « The effects of interaction sequencing on user experience and preference », *International Journal of Human Computer Studies*, vol. 108, p. 89-104.

Cronbach, Lee J. (1951). « Coefficient alpha and the internal structure of tests », *Psychometrika*, vol. 16, no 3, p. 297-334.

Dawson, Michael E., Anne M. Schell et Diane L. Filion (2007). « The electrodermal system », dans John T. Cacioppo, Louis G. Tassinary et Gary G. Bernston (dir.), 3e éd, Cambridge, Cambridge University Press, p. 159-181.

de Haan, Evert, Peter C. Verhoef et Thorsten Wiesel (2015). « The predictive ability of different customer feedback metrics for retention », *International Journal of Research in Marketing*, vol. 32, no 2, p. 195-206.

Deese, James et Roger A. Kaufman (1957). « Serial effects in recall of unorganized and sequentially organized verbal material », *Journal of Experimental Psychology*, vol. 54, no 3, p. 180-187.

DeVellis, Robert F. (2003). *Scale development: Theory and applications*, 2e éd., Thousand Oaks, USA, SAGE Publications, 171-171 p.

DeVon, Holli A., Michelle E. Block, Patricia Moyle-Wright, Diane M. Ernst, Susan J. Hayden, Deborah J. Lazzara*, et al.* (2007). « A psychometric toolbox for testing validity and reliability », *Journal of Nursing Scholarship*, vol. 39, no 2, p. 155-164.

DiGirolamo, Gregory J. et Douglas L. Hintzman (1997). « First impressions are lasting impressions: A primacy effect in memory for repetitions », *Psychonomic Bulletin and Review*, vol. 4, no 1, p. 121-124.

Dixon, Matthew, Karen Freeman et Nicolas Toman (2010). « Stop trying to delight your customers », *Harvard Business Review*, vol. 88, no 7-8.

Drolet, Aimee L. et Donald G. Morrison (2001). *Do we really need multiple-item measures in service research?*

Easterling, Robert G. (2015). *Fundamentals of statistical experimental design and analysis*, 1e éd., Chichester, UK, Wiley. Récupéré de http://site.ebrary.com/id/11113448

Ekman, P. et W. V. Friesen (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*.

Giese, J. et J. Cote (2000). « Defining consumer satisfaction », *Academy of marketing science review*, vol. 2000, p. 1-1.

Griffin, Jason W., Samantha E. John, Jason W. Adams, Cara A. Bussell, Jessica L. Saurman et Brandon E. Gavett (2017). « The effects of age on the learning and forgetting of primacy, middle, and recency components of a multi-trial word list », *Journal of Clinical and Experimental Neuropsychology*, vol. 39, no 9, p. 900-912.

Hansemark, Ove C. et Marie Albinsson (2004). « Customer satisfaction and retention: The experiences of individual employees », *Managing Service Quality: An International Journal*, vol. 14, no 1, p. 40-57.

Hansen, David E. et Peter J. Danaher (1999). « Inconsistent performance during the service encounter: What's a good start worth? », *Journal of Service Research*, vol. 1, no 3, p. 227-235.

Hassenzahl, Marc (2003). « The thing and i: Understanding the relationship between user and product », dans, Springer, Cham, p. 301-313.

Herbon, Antje, Christian Peter, Lydia Markert et Elke Van Der Meer (2005). « Emotion studies in hci – a new approach », *Proceedings of the 2005 HCI International Conference*, no 1986.

Hogan, Thomas P., Amy Benjamin et Kristen L. Brezinski (2000). « Reliability methods: A note on the frequency of use of various types », *Educational and Psychological Measurement*, vol. 60, no 4, p. 523-531.

Jacoby, Jacob (1978). « Consumer research: How valid and useful are all our consumer behavior research findings? », *Journal of Marketing*, vol. 42, no 2, p. 87-96.

Keiningham, Timothy L., Bruce Cooil, Lerzan Aksoy, Tor W. Andreassen et Jay Weiner (2007). « The value of different customer satisfaction and loyalty metrics in predicting customer retention, recommendation, and share-of-wallet », *Managing Service Quality: An International Journal*, vol. 17, no 4, p. 361-384.

Kerlinger, Fred N. et Howard B. Lee (2000). *Foundations of behavioral research*, 4e éd., Fort Worth, USA, Harcourt College Publisher, 890-890 p.

Kim, Yong Mi (2009). « Validation of psychometric research instruments: The case of information science », *Journal of the American Society for Information Science and Technology*, vol. 60, no 6, p. 1178-1191.

Lang, Peter J., Mark K. Greenwald, Margaret M. Bradley et Alfons O. Hamm (1993). « Looking at pictures: Affective, facial, visceral, and behavioral reactions », *Psychophysiology*, vol. 30, no 3, p. 261-273.

Le Pailleur, Félix, Bo Huang, Pierre Majorique Léger et Sylvain Sénécal (2020). « A new approach to measure user experience with voice-controlled intelligent assistants: A pilot study », communication présentée au *HCII 2020*, Copenhagen, Denmark.

Léger, Pierre Majorique, Patrick Charland, Sylvain Sénécal et Stéphane Cyr (2017). « Predicting properties of cognitive pupillometry in human–computer interaction: A preliminary investigation », *Lecture Notes in Information Systems and Organisation*.

Léger, Pierre Majorique, Francois Courtemanche, Marc Fredette et Sylvain Sénécal (2019). « A cloud-based lab management and analytics software for triangulated human-centered research », dans, vol 29, Springer Heidelberg, p. 93-99.

Lewis, Bruce R., Gary F. Templeton et Terry Anthony Byrd (2005). « A methodology for construct development in mis research », *European Journal of Information Systems*, vol. 14, no 4, p. 388-400.

Lewis, James R. (2002). « Psychometric evaluation of the pssuq using data from five years of usability studies », *International Journal of Human-Computer Interaction*, vol. 14, no 3-4, p. 463-488.

Li, Cong (2009). « Primacy effect or recency effect? A long-term memory test of super bowl commercials », *Journal of Consumer Behaviour*, vol. 9, no 1, p. 32-44.

Li, Shanshi, Gabby Walters, Jan Packer et Noel Scott (2018). « A comparative analysis of self-report and psychophysiological measures of emotion in the context of tourism advertising », *Journal of Travel Research*, vol. 57, no 8, p. 1078-1092.

Lindgaard, Gitte, Gary Fernandes, Cathy Dudek et J. Browñ (2006). « Attention web designers: You have 50 milliseconds to make a good first impression! », *Behaviour and Information Technology*, vol. 25, no 2, p. 115-126.

Loijens, Leanne, Olga Krips, Fabrizio Grieco, Hans van Kuilenburg, Marten den Uyl et Paul Ivan (2016). *Innovative solutions for behavioral research facereader ™ tool for automatic analysis of facial expressions reference manual version 7*. Récupéré de www.noldus.com

Lourties, Sébastien, Pierre Majorique Léger, Sylvain Sénécal, Marc Fredette et Shang Lin Chen (2018). « Testing the convergent validity of continuous self-perceived measurement systems: An exploratory study », communication présentée au *HCII 2018*, Las Vegas, USA.

Maia, Camila Loiola Brito et Elizabeth Sucupira Furtado (2019). « An approach to analyze user's emotion in hci experiments using psychophysiological measures », *IEEE Access*, vol. 7, p. 36471-36480.

Moore, Gary C. et Izak Benbasat (1991). « Development of an instrument to measure the perceptions of adopting an information technology innovation », *Information Systems Research*, vol. 2, no 3, p. 192-222.

Murdock, Bennet B. (1962). « The serial position effect of free recall », *Journal of Experimental Psychology*, vol. 64, no 5, p. 482-488.

Murphy, Jamie, Charles Hofacker et Richard Mizerski (2006). « Primacy and recency effects on clicking behavior », *Journal of Computer-Mediated Communication*, vol. 11, no 2, p. 522-535.

Nagy, Mark S. (2002). « Using a single-item approach to measure facet job satisfaction », *Journal of Occupational and Organizational Psychology*, vol. 75, no 1, p. 77-86.

Nunnally, Jum C. et Ira H. Bernstein (1994). *Psychometric theory*, 3e éd., McGraw-Hill, 752-752 p.

Ordoñana, Juan R., Francisca González-Javier, Laura Espín-López et Jesús Gómez-Amor (2009). « Self-report and psychophysiological responses to fear appeals », *Human Communication Research*, vol. 35, no 2, p. 195-220.

Ortiz de Guinea, Ana, Ryad Titah et Pierre Majorique Léger (2013). « Measure for measure: A two study multi-trait multi-method investigation of construct validity in is research », *Computers in Human Behavior*, vol. 29, no 3, p. 833-844.

Paas, Fred, Juhani E. Tuovinen, Huib Tabbers et Pascal W. M. Van Gerven (2003). « Cognitive load measurement as a means to advance cognitive load theory », *Educational Psychologist*, vol. 38, no 1, p. 63-71.

Pingitore, Gina, Neil A. Morgan, Lopo L. Rego, Adriana Gigliotti et Jay Meyers (2007). « The single-question trap », *Marketing Research*, vol. 19, no 2, p. 8-13.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee et Nathan P. Podsakoff (2003). « Common method biases in behavioral research: A critical review of the literature and recommended remedies », *Journal of Applied Psychology*, vol. 88, no 5, p. 879-879.

Pointilist (2019). *State of customer journey management & cx measurement*, 31-31 p.

Razavi, Seyed Mostafa, Hossein Safari, Hessam Shafie et Hadi Rezaei Vandchali (2012). « How customer satisfaction, corporate image and customer loyalty are related? », *European Journal of Scientific Research*, vol. 78, no 4, p. 588-596.

Riedl, René, Fred D. Davis et Alan R. Hevner (2014). « Towards a neurois research methodology: Intensifying the discussion on methods, tools, and measurement », *Journal of the Association for Information Systems*, vol. 15, p. 1-35.

Riedl, René et Pierre-Majorique Léger (2016). *Fundamentals of neurois*, Berlin, Heidelberg, Springer Berlin Heidelberg, coll. Studies in neuroscience, psychology and behavioral economics.

Rossiter, John R. (2002). « The c-oar-se procedure for scale development in marketing », *International Journal of Research in Marketing*, vol. 19, no 4, p. 305-335.

Russel, James A. (1980). « A circumplex model of affect », *Journal of Personality and Social Psychology*, vol. 39, no 6, p. 1161-1178.

Safdar, By Khadeeja et Inti Pacheco (2019). « The dubious management fad sweeping corporate america », *Wall Street Journal*, p. 1-10.

Sarstedt, Marko et Petra Wilczynski (2009). « More for less? A comparison of single-item and multi-item », *Die Betriebswirtschaft*, vol. 69, no 2, p. 211-227.

Seo, Kwang Kyu, Sangwon Lee, Byung Do Chung et Changsoon Park (2015). « Users' emotional valence, arousal, and engagement based on perceived usability and aesthetics for web sites », *International Journal of Human-Computer Interaction*, vol. 31, no 1, p. 72-87.

Skiendziel, Tanja, Andreas G. Rösch et Oliver C. Schultheiss (2019). « Assessing the convergent validity between the automated emotion recognition software noldus facereader 7 and facial action coding system scoring », *PLoS ONE*, vol. 14, no 10, p. e0223905-e0223905.

Statista (2020). *Ways of conducting bank transactions in canada 2018*. Récupéré le 9 décembre 2020 de https://www-statista-com.proxy2.hec.ca/statistics/709927/canadian-banking-transactions-methods/

Steiger, James H. (1980). « Tests for comparing elements of a correlation matrix », *Psychological Bulletin*.

Straub, Detmar, Marie-Claude Boudreau et David Gefen (2004). « Validation guidelines for is positivist research », *Communications of the Association for Information Systems*, vol. 13, p. 380-427.

Straub, Detmar W. (1989). « Validating instruments in mis research », *MIS Quarterly: Management Information Systems*, vol. 13, no 2, p. 147-165.

Streiner, David L. (2003). « Starting at the beginning: An introduction to coefficient alpha and internal consistency », *Journal of Personality Assessment*, vol. 80, no 1, p. 99-103.

Tams, Stefan, Jason Thatcher, Kevin Hill, Varun Grover et Ana Ortiz de Guinea (2014). « Neurois—alternative or complement to existing methods? Illustrating the holistic effects of neuroscience and self-reported data in the context of technostress research », *Journal of the Association for Information Systems*, vol. 15, p. 723-753.

Thielsch, Meinald T., Iris Blotenberg et Rafael Jaron (2014). « User evaluations of websites: From first impression to recommendation », *Interacting with Computers*, vol. 26, no 1, p. 89-102.

Thompson, Ed, Don Scheibenreif et Michael Chiu (2020). *How to manage customer experience metrics*, 20-20 p. Récupéré de https://www.gartner.com/document/3979139?ref=lib

Thüring, Manfred et Sascha Mahlke (2007). « Usability, aesthetics and emotions in human-technology interaction », *International Journal of Psychology*, vol. 42, no 4, p. 253-264.

Tomarken, Andrew J. (1995). « A psychometric perspective on psychophysiological measures », *Psychological Assessment*, vol. 7, no 3, p. 387-395.

Tractinsky, N., A. S. Katz et D. Ikar (2000). « What is beautiful is usable », *Interacting with Computers*, vol. 13, no 2, p. 127-145.

Wang, Xuehua (2011). « The effect of inconsistent word-of-mouth during the service encounter », *Journal of Services Marketing*, vol. 25, no 4, p. 252-259.

Wanous, John P., Arnon E. Reichers et Michael J. Hudy (1997). « Overall job satisfaction: How good are single-item measures? », *Journal of Applied Psychology*, vol. 82, no 2, p. 247-252.

Wirtz, Jochen et Meng Chung Lee (2003). « An examination of the quality and context-specific applicability of commonly used customer satisfaction measures », *Journal of Service Research*, vol. 5, no 4, p. 345-355.

Xie, Heping, Fuxing Wang, Yanbin Hao, Jiaxue Chen, Jing An, Yuxin Wang*, et al.* (2017). « The more total cognitive load is reduced by cues, the better retention and transfer of multimedia learning: A meta-analysis and two meta-regression analyses », *PLoS ONE*, vol. 12, no 8.

Conclusion

Ce dernier chapitre a pour intention de revenir sur les questions de recherche et présente les principales observations de nos explorations. Des contributions théoriques et managériales y sont ensuite discutées. Il se termine par les différentes limites ainsi que les propositions de recherches futures.

Ce mémoire possédait un objectif global. Le principal but de ce mémoire était d'explorer différentes relations. Nous voulions savoir s'il existait de réelles corrélations entre une expérience qui a été vécue par plusieurs personnes et la même expérience qui a été perçue par ces mêmes personnes et reportée à l'aide d'un format d'échelles en particulier : des échelles à un item. Ce mémoire a permis, jusqu'à un certain point, de mieux comprendre ces relations, mais surtout, de mettre en lumière les limites à prendre en compte lorsque nous utilisons ce format d'échelles.

Une expérience en laboratoire a été réalisée entre les mois de décembre 2019 et janvier 2020 avec 40 participants. Ces derniers devaient réaliser deux séries de trois tâches sur 10 sites internet d'institutions financières canadiennes différentes. Les participants devaient dans un premier temps explorer un site afin de trouver l'espace pour remplir une demande de préautorisation hypothécaire. Ensuite, il leur était demandé de remplir une demande de préautorisation hypothécaire. Enfin, nous leur demandions d'explorer à nouveau le site afin de trouver l'outil pour entrer en contact avec un conseil en hypothèques. L'expérience vécue a été mesurée à l'aide de quatre outils différents : un logiciel d'analyse des expressions faciales (Facereader ®, Noldus, Wageningen, Pays-Bas), un oculomètre (Tobii X-60 ®, Stockholm, Suède), un outil pour mesurer l'activité électrodermale (Biopac ®, Goleta, États-Unis) et enfin, un électrocardiogramme (Biopac ®, Goleta, États-Unis). L'expérience perçue a été mesurée par quatre questionnaires différents : l'Affective Slider, le CES, le CSAT et le NPS.

Rappel des questions de recherche et principaux résultats

Notre étude nous a permis de répondre partiellement à nos questions de recherche. Ces réponses ont été adressées dans notre article. Pour rappel, elles étaient les suivantes :

QR1 : « *Dans quelle mesure l'expérience utilisateur vécue est-elle en relation avec l'expérience perçue?* »

QR2 : « *Dans quelle mesure la relation entre l'expérience utilisateur vécue et l'expérience utilisateur perçue est-elle différente à différents moments d'une interaction? »*

QR3 : « *Dans quelle mesure les échelles à un item sont-elles suffisamment sensibles pour distinguer l'expérience utilisateur à travers différents contextes d'utilisation, tels que des contextes hédoniques ou utilitaires?* »

De manière générale, la réponse à ces questions est mitigée. Les analyses effectuées permettent de mettre en lumière plusieurs corrélations entre différentes variables, mesurées des deux manières. Nous avons observé de faibles liens au niveau de l'expérience globale. Des corrélations qui, lorsque nous ne tenons compte que de la première ou de la dernière impression, se précisent ou apparaissent. C'est le cas pour la relation entre l'activation perçue et l'activation vécue et mesurée par l'activité électrodermale, qui apparaît comme étant une paire de construits équivalents. Au sujet des construits qui ne sont pas équivalents, nous avons observé une relation entre le plaisir perçu et la charge cognitive vécue.

Lorsque nous distinguons les différents types d'attributs présents sur les sites internet utilisés, plusieurs autres liens font leur apparition : par exemple, le plaisir perçu corrèle avec la valence, lors de la première impression tant pour les attributs hédoniques qu'utilitaires. La satisfaction et l'intention de recommander corrèlent également toutes les deux avec la valence à différents niveaux de l'expérience. Certaines relations se précisent : l'activation perçue semblerait corréler avec l'activité électrodermale à la première impression lorsque nous tenons compte uniquement des attributs utilitaires, mais à la dernière impression pour les attributs hédoniques. Ce même construit corrèle avec le

rythme cardiaque uniquement à la première impression pour les éléments utilitaires. Cependant, de manière étonnante, nos corrélations ne se comportent, en très grande majorité, pas comme elles le devraient. Par exemple, le plaisir perçu corrèle négativement avec la valence. Cela indiquerait que plus cette dernière augmente, plus la perception du plaisir diminue. Ces directions étonnantes sont valables pour la très nette majorité de nos variables, à l'exception de l'activation perçue. Les directions de nos corrélations mettent en avant un inconvénient majeur des mesures à un item : il est possible qu'elles ne représentent pas toute la dimensionnalité d'un construit d'intérêt (Baumgartner et Homburg, 1996). Il apparaît comme étant probable que d'autres facteurs puissent entrer en jeu lorsqu'une personne reporte l'émotion qu'elle a vécue à l'aide d'une de ces échelles.

Nous avions, en premier lieu, supposé des différences de corrélations entre nos trois niveaux d'analyse : le niveau global, la première impression et la dernière impression (H1). Ces différences existent, bien qu'elles concernent en majorité des différences entre des relations significatives et des relations non-significatives. En second lieu, notre deuxième hypothèse supposait des différences de corrélations entre différents types d'attributs présents sur les sites internet utilisés, avec un effet plus important pour les attributs considérés comme hédoniques (H2). Le constat est similaire à H1, plusieurs différences existent, certes, mais elles apparaissent également surtout entre des corrélations significatives et non-significatives. Il nous apparaît cependant important de mettre en avant les différences entre la valence et trois mesures perceptuelles : le plaisir perçu, la satisfaction et l'intention de recommander. Il semblerait exister des différences significatives entre les attributs présents lors de la première impression sur les sites internet utilisés. Nos résultats ne nous permettent, de ce fait, que de confirmer partiellement nos hypothèses.

Contributions

Les principales contributions de ce mémoire peuvent être divisées en deux parties : des contributions théoriques et plusieurs implications managériales.

Contributions théoriques

Sur le plan théorique, nos analyses permettent de donner un éclairage sur la théorie émanant de la psychométrie. Ce mémoire explore jusqu'à quel point le fait d'avoir vécu une expérience était proche des perceptions que l'on pouvait en avoir, cela en reportant nos émotions sur un format d'échelles en particulier. D'autres études se sont déjà intéressées à ce phénomène (Ortiz de Guinea, Titah et Léger, 2013 ; Lourties, Léger, Sénécal, Fredette, Chen, 2018). Notre recherche s'inscrit dans leur continuité et permet, elle aussi, d'insister sur l'importance d'utiliser plusieurs méthodes différentes afin de mesurer un construit. Le premier chapitre cherchait à connaître et comprendre la théorie sur laquelle se basaient les quatre échelles utilisées dans notre étude. Nous avions mis en avant plusieurs bénéfices et inconvénients d'utiliser une seule question par construit. Le deuxième chapitre a permis de mettre la théorie en pratique et de trouver des pistes de réponses à nos questions.

Notre recherche permet également de contribuer à la théorie sur les biais cognitifs et plus précisément au sujet de l'effet de position sérielle. Des effets de première impression sont apparus dans nos résultats. Ces derniers appuient les observations faites par Lindgaard, Fernandes, Dudek et Brown (2006). Des effets de dernière impression sont aussi apparus dans notre recherche, appuyant les dires d'Hassenzahl et Sandweg (2004), qui suggéraient un certain effet de récence.

Implications managériales

Ce mémoire possède plusieurs implications managériales. Il met en lumière l'une des principales limites des échelles à un item. Si mesurer l'expérience de manière implicite permet d'estimer ce qu'une personne vit durant une interaction, les mesures explicites donnent l'occasion d'observer ce que cette personne se souvient de son interaction, son opinion. Un point majeur a été soulevé dans notre étude : nous ne sommes

pas en mesure d'observer systématiquement ce que nous devrions nous attendre d'elles. Nous ne pouvons pas examiner la manière dont elles se comportent. Ces mesures, finalement, ne permettraient pas de brosser un portrait le plus complet, le plus précis du vécu d'un utilisateur lors de son interaction avec une interface. Nous ne prétendons cependant pas qu'une méthode est meilleure qu'une autre, chacune possédant ses bénéfices et limites. À notre sens, il nous apparaît important de considérer plusieurs méthodes lorsque nous mesurons une interaction afin d'obtenir des données les plus riches possibles.

Notre recherche permet également d'insister sur l'importance des attributs spécifiques à un site internet. Le premier chapitre montrait que les attributs d'un système faisaient partie des antécédents communs aux mesures que nous avions décidé d'utiliser. Nos résultats suggèrent que, parfois, ces attributs possèderaient un certain impact sur l'expérience vécue et sa relation avec l'expérience reportée à l'aide d'échelles à un item. Bien que cette relation n'apparaisse pas systématiquement.

Limites et pistes de recherches

Plusieurs limites sont présentes dans notre recherche. À commencer par l'opérationnalisation de nos données. Lorsque nous avons créé les différentes aires d'intérêt issues des fixations oculaires des participants, nous nous sommes basés sur les lignes directrices Hassenzahl (2003), qui proposaient la classification d'éléments hédoniques et d'éléments utilitaires. De ces lignes directrices, nous avons été en mesure de calculer le pourcentage d'attributs différents. Étant donné la nature très utilitaire des dernières pages utilisées par nos participants, très peu d'éléments hédoniques sont apparus dans nos données. Les résultats de nos corrélations sur ces éléments précis sont alors à considérer avec précaution. Il est possible que, si nous avions utilisé d'autres critères, ces corrélations puissent être quelque peu différentes. Il est possible que, dans un contexte beaucoup plus hédonique, ces corrélations soient différentes. Il nous apparaîtrait approprié de répliquer ce type d'études avec d'autres types de stimuli.

La deuxième limite provient directement de notre méthodologie. Nous n'avons volontairement pas utilisé d'autres échelles mesurant les mêmes construits afin de les comparer aux échelles à un item. Il serait intéressant pour la recherche future d'utiliser d'autres formats d'échelles et de les comparer à ces mesures afin de savoir si, finalement, elles sont réellement plus éloignées de l'expérience vécue.

Enfin, la dernière limite provient de nos participants. Ils étaient variés mais leur profil sociodémographique est relativement similaire, principalement en termes d'âge. Des effets d'âge existent sur les effets de première et dernière impression (Griffin, John, Adams, Bussell, Saurman et Gavett, 2017). Il serait pertinent de considérer un autre groupe de participants afin de comparer leurs résultats à ceux obtenus.

En conclusion, il nous apparaît important d'insister sur l'intérêt de mieux comprendre à quel point ce type d'échelles représente une manière fiable, valide et suffisamment sensible pour mesurer l'expérience utilisateur. Bien que nos résultats mitigés nous aient permis d'adresser plusieurs pistes de réponses à nos questions de recherche, ces derniers ne nous permettent pas de comprendre précisément ces relations. La recherche future gagnerait à continuer dans cette direction en explorant ces relations à travers d'autres tâches (par exemple, en jouant à un jeu vidéo); mais également en faisant appel à d'autres types de stimuli (par exemple des stimuli purement utilitaires ou au contraire, purement hédoniques). Ces futures explorations permettraient de brosser un portrait encore plus précis des forces et faiblesses des échelles de mesure à un item.

Références

Baumgartner, Hans et Christian Homburg (1996). « Applications of structural equation modeling in marketing and consumer research: A review », *International Journal of Research in Marketing*, vol. 13, no 2, p. 139-161.

Griffin, Jason W., Samantha E. John, Jason W. Adams, Cara A. Bussell, Jessica L. Saurman et Brandon E. Gavett (2017). « The effects of age on the learning and forgetting of primacy, middle, and recency components of a multi-trial word list », *Journal of Clinical and Experimental Neuropsychology*, vol. 39, no 9, p. 900-912.

Hassenzahl, Marc (2003). « The thing and i: Understanding the relationship between user and product », dans, Springer, Cham, p. 301-313.

Hassenzahl, M., & Sandweg, N. (2004). From mental effort to perceived usability: Transforming experiences into summary assessments. *Conference on Human Factors in Computing Systems - Proceedings*, 1283–1286. https://doi.org/10.1145/985921.986044

Lindgaard, Gitte, Gary Fernandes, Cathy Dudek et J. Browñ (2006). « Attention web designers: You have 50 milliseconds to make a good first impression! », *Behaviour and Information Technology*, vol. 25, no 2, p. 115-126.

Lourties, S., Léger, P. M., Sénécal, S., Fredette, M., & Chen, S. L. (2018). Testing the convergent validity of continuous self-perceived measurement systems: An exploratory study. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10923 LNCS*, p. 132–144.

Ortiz de Guinea, Ana, Ryad Titah et Pierre Majorique Léger (2013). « Measure for measure: A two study multi-trait multi-method investigation of construct validity in is research », Computers in Human Behavior, vol. 29, no 3, p. 833-844.

Bibliographie

Alpers, Georg W. et Roxane Sell (2008). « And yet they correlate: Psychophysiological activation predicts self-report outcomes of exposure therapy in claustrophobia », *Journal of Anxiety Disorders*, vol. 22, no 7, p. 1101-1109.

André, Nathalie, Nathalie Loye et Louis Laurencelle (2016). « La validité psychométrique : Un regard global sur le concept centenaire, sa genèse, ses avatars », *Mesure et évaluation en éducation*, vol. 37, no 3, p. 125-148.

Bain & Company (2020). *Bain & company - measuring your net promoter score - net promoter system*. Récupéré le 23 novembre 2020 de https://www.https://www.netpromotersystem.com/about/

Barnes, G. Michael (1992). « Digitized speech's serial position effect », communication présentée au *1992 SIGCHI Conference on Human Factors in Computing Systems*, 1992, Monterey, California.

Baumgartner, Hans et Christian Homburg (1996). « Applications of structural equation modeling in marketing and consumer research: A review », *International Journal of Research in Marketing*, vol. 13, no 2, p. 139-161.

Bendle, Neil T., Paul W. Farris, Phillip E. Pfeifer et David J. Reibstein (2016). *Marketing metrics: The manager's guide to measuring marketing performance, third edition*, 3e éd., Upper Saddle River, USA, Pearson, 439-439 p.

Bergeron, Jasmin, J. M. Fallu et Jasmin Roy (2008). « Une comparaison des effets de la première et de la dernière impression dans une rencontre de vente », *Recherche et Applications en Marketing*, vol. 23, no 2, p. 19-36.

Bergkvist, Lars et John R. Rossiter (2007). « The predictive validity of multiple-item versus single-item measures of the same constructs », *Journal of Marketing Research*, vol. 44, no 2, p. 175-184.

Betella, Alberto et Paul F. M. J. Verschure (2016). « The affective slider: A digital self-assessment scale for the measurement of human emotions », *PLoS ONE*, vol. 11, no 2.

Bharadwaj, Neeraj et Ken Matsuno (2006). « Investigating the antecedents and outcomes of customer firm transaction cost savings in a supply chain relationship », *Journal of Business Research*, vol. 59, no 1, p. 62-72.

Bilgihan, Anil et Milos Bujisic (2015). « The effect of website features in online relationship marketing: A case of online hotel booking », *Electronic Commerce Research and Applications*, vol. 14, no 4, p. 222-232.

Borsboom, Denny, Gideon J. Mellenbergh et Jaap Van Heerden (2003). « The theoretical status of latent variables », *Psychological Review*, vol. 110, no 2, p. 203-219.

Boyle, Gregory J. (1991). « Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? », *Personality and Individual Differences*, vol. 12, no 3, p. 291-294.

Bradley, Margaret M. et Peter J. Lang (1994). « Measuring emotion: The self-assessment manikin and the semantic differential », *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no 1, p. 49-59.

Brooke, John (1996). « Sus-a quick and dirty usability scale », *Usability evaluation in industry*.

Brooke, John (2013). « Sus: A retrospective », *Journal of usability studies*, vol. 8, no 2, p. 29-40.

Brünken, Roland, Jan L. Plass et Detlev Leutner (2003). « Direct measurement of cognitive load in multimedia learning », *Educational Psychologist*, vol. 38, no 1, p. 53-61.

Cai, Shun et Yunjie Xu (2011). « Designing not just for pleasure: Effects of web site aesthetics on consumer shopping value », *International Journal of Electronic Commerce*, vol. 15, no 4, p. 159-188.

Cannon, W. B. (1987). « The james-lange theory of emotions: A critical examination and an alternative theory. By walter b. Cannon, 1927 », *The American journal of psychology*, vol. 100, no 3-4, p. 567-586.

Chang, En-Chung et Xiaomeng Fan (2013). « More promoters and less detractors: Using generalized ordinal logistic regression to identify drivers of customer loyalty », *International Journal of Marketing Studies*.

Churchill, Gilbert A. (1979). « A paradigm for developing better measures of marketing constructs », *Journal of Marketing Research*, vol. 16, no 1, p. 64-73.

Churchill, Gilbert A., Neil M. Ford et Orville C. Walker (1974). « Measuring the job satisfaction of industrial salesmen », *Journal of Marketing Research*, vol. 11, no 3, p. 254-260.

Churchill, Gilbert A. et Carol Surprenant (1982). « An investigation into the determinants of customer satisfaction », *Journal of Marketing Research*, vol. 19, no 4, p. 491-491.

Cockburn, Andy, Philip Quinn et Carl Gutwin (2017). « The effects of interaction sequencing on user experience and preference », *International Journal of Human Computer Studies*, vol. 108, p. 89-104.

Cook, David A. et Thomas J. Beckman (2006). « Current concepts in validity and reliability for psychometric instruments: Theory and application », *American Journal of Medicine*, vol. 119, no 2, p. 166.e167-166.e116.

Cook, T. D., D. T. Campbell et A. Day (1979). *Quasi-experimentation: Design & analysis issues for field settings*, 1e éd., Boston, USA, H. Mifflin, 405-405 p.

Cronbach, Lee J. (1951). « Coefficient alpha and the internal structure of tests », *Psychometrika*, vol. 16, no 3, p. 297-334.

Cronin, J. Joseph et Steven A. Taylor (1992). « Measuring service quality: A reexamination and extension », *Journal of Marketing*, vol. 56, no 3, p. 55-55.

Cyr, Dianne, Milena Head, Hector Larios et Bing Pan (2009). « Exploring human images in website design: A multi-method approach », *MIS Quarterly: Management Information Systems*.

Dawson, Michael E., Anne M. Schell et Diane L. Filion (2007). « The electrodermal system », dans John T. Cacioppo, Louis G. Tassinary et Gary G. Bernston (dir.), 3e éd, Cambridge, Cambridge University Press, p. 159-181.

de Haan, Evert, Peter C. Verhoef et Thorsten Wiesel (2015). « The predictive ability of different customer feedback metrics for retention », *International Journal of Research in Marketing*, vol. 32, no 2, p. 195-206.

De Pechpeyrou, Pauline et Patrick Nicholson (2019). *Réclamation et satisfaction : L'effort perçu du client rebat les cartes complaining behavior and satisfaction: Customer's effort score shuffle the cards*.

Deese, James et Roger A. Kaufman (1957). « Serial effects in recall of unorganized and sequentially organized verbal material », *Journal of Experimental Psychology*, vol. 54, no 3, p. 180-187.

DeVellis, Robert F. (2003). *Scale development: Theory and applications*, 2e éd., Thousand Oaks, USA, SAGE Publications, 171-171 p.

DeVon, Holli A., Michelle E. Block, Patricia Moyle-Wright, Diane M. Ernst, Susan J. Hayden, Deborah J. Lazzara*, et al.* (2007). « A psychometric toolbox for testing validity and reliability », *Journal of Nursing Scholarship*, vol. 39, no 2, p. 155-164.

Dew, Denis (2008). « Construct », dans Paul. J. Lavrakas (dir.), *Encyclopedia of survey research methods*, SAGE Publications, p. 133-134.

DiGirolamo, Gregory J. et Douglas L. Hintzman (1997). « First impressions are lasting impressions: A primacy effect in memory for repetitions », *Psychonomic Bulletin and Review*, vol. 4, no 1, p. 121-124.

Dixon, Matthew, Karen Freeman et Nicolas Toman (2010). « Stop trying to delight your customers », *Harvard Business Review*, vol. 88, no 7-8.

Drolet, Aimee L. et Donald G. Morrison (2001). *Do we really need multiple-item measures in service research?*

Easterling, Robert G. (2015). *Fundamentals of statistical experimental design and analysis*, 1e éd., Chichester, UK, Wiley. Récupéré de http://site.ebrary.com/id/11113448

Ekman, P. et W. V. Friesen (2003). *Unmasking the face: A guide to recognizing emotions from facial clues*.

Ganglbauer, Eva, Johann Schrammel, Stephanie Deutsch et Manfred Tscheligi (2011). « Applying psychophysiological methods for measuring user experience: Possibilities, challenges and feasibility », *Human-Computer Interaction. INTERACT 2011 (Lecture Notes in Computer Science)*.

Gefen, David et Detmar Straub (2005). « A practical guide to factorial validity using pls-graph: Tutorial and annotated example », *Communications of the Association for Information Systems*, vol. 16, p. 91-109.

Giese, J. et J. Cote (2000). « Defining consumer satisfaction », *Academy of marketing science review*, vol. 2000, p. 1-1.

Griffin, Jason W., Samantha E. John, Jason W. Adams, Cara A. Bussell, Jessica L. Saurman et Brandon E. Gavett (2017). « The effects of age on the learning and forgetting of primacy, middle, and recency components of a multi-trial word list », *Journal of Clinical and Experimental Neuropsychology*, vol. 39, no 9, p. 900-912.

Hallowell, Roger (1996). « The relationships of customer satisfaction, customer loyalty, and profitability: An empirical study », *International Journal of Service Industry Management*, vol. 7, no 4, p. 27-42.

Hansemark, Ove C. et Marie Albinsson (2004). « Customer satisfaction and retention: The experiences of individual employees », *Managing Service Quality: An International Journal*, vol. 14, no 1, p. 40-57.

Hansen, David E. et Peter J. Danaher (1999). « Inconsistent performance during the service encounter: What's a good start worth? », *Journal of Service Research*, vol. 1, no 3, p. 227-235.

Hassenzahl, Marc (2003). « The thing and i: Understanding the relationship between user and product », dans, Springer, Cham, p. 301-313.

Hassenzahl, Marc, Michael Burmester et Franz Koller (2003). « Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität », *Mensch & Computer 2003 : Interaktion in Bewegung*, p. 187-196.

Hassenzahl, Marc et Nina Sandweg (2004). « From mental effort to perceived usability: Transforming experiences into summary assessments », communication présentée au *CHI'04*, 2004, Vienna, Austria.

Haynes, Stephen N., David C. S. Richard et Edward S. Kubany (1995). « Content validity in psychological assessment: A functional approach to concepts and methods », *Psychological Assessment*, vol. 7, no 3, p. 238-247.

Herbon, Antje, Christian Peter, Lydia Markert et Elke Van Der Meer (2005). « Emotion studies in hci – a new approach », *Proceedings of the 2005 HCI International Conference*, no 1986.

Hogan, Thomas P., Amy Benjamin et Kristen L. Brezinski (2000). « Reliability methods: A note on the frequency of use of various types », *Educational and Psychological Measurement*, vol. 60, no 4, p. 523-531.

Imbault, C., D. Shore et V. Kuperman (2018). « Reliability of the sliding scale for collecting affective responses to words », *Behavior Research Methods*, vol. 50, no 6, p. 2399-2407.

Ishaq, Muhammad Ishtiaq (2011). « A study on relationship between service quality and customer satisfaction: An empirical evidence from Pakistan telecommunication industry », *Management Science Letters*, vol. 1, no 4, p. 523-530.

Jacoby, Jacob (1978). « Consumer research: How valid and useful are all our consumer behavior research findings? », *Journal of Marketing*, vol. 42, no 2, p. 87-96.

Jones, Michael A. et Jaebeom Suh (2000). « Transaction-specific satisfaction and overall satisfaction: An empirical analysis », *Journal of Services Marketing*, vol. 14, no 2, p. 147-159.

Kamakura, Wagner A. (2010). « Common methods bias », dans, Chichester, UK, John Wiley & Sons, Ltd.

Kees, Jasmine (2020). *2020 state of customer service report*, 46-46 p.

Keiningham, Timothy L., Bruce Cooil, Lerzan Aksoy, Tor W. Andreassen et Jay Weiner (2007). « The value of different customer satisfaction and loyalty metrics in predicting customer retention, recommendation, and share-of-wallet », *Managing Service Quality: An International Journal*, vol. 17, no 4, p. 361-384.

Kerlinger, Fred N. et Howard B. Lee (2000). *Foundations of behavioral research*, 4e éd., Fort Worth, USA, Harcourt College Publisher, 890-890 p.

Kim, Yong Mi (2009). « Validation of psychometric research instruments: The case of information science », *Journal of the American Society for Information Science and Technology*, vol. 60, no 6, p. 1178-1191.

Korneta, Pawel (2014). « What makes customers willing to recommend a retailer - the study on roots of positive net promoter score index abstract », *Central European Review of Economics & Finance*.

Kristensen, Kai et Jacob Eskildsen (2014). « Is the nps a trustworthy performance measure? », *TQM Journal*, vol. 26, no 2, p. 202-214.

Kumar, Niraj, Ajay Pal Singh et Reshmi Manna (2013). *Analyzing consumer behaviour towards service quality of indianc electronic gadget firms*, 14-14 p.

Kumar, R. et A. Mittal (2015). « Customer satisfaction and service quality perception of technology based banking services: A study on selected public sector banks in india », *Global Journal of Management and Business Research : E Marketing*, vol. 15, no 5, p. 39-45.

Lallemand, C., V. Koenig, G. Gronier et R. Martin (2015). « Création et validation d'une version française du questionnaire attrakdiff pour l'évaluation de l'expérience utilisateur des

systèmes interactifs », *Revue Europeenne de Psychologie Appliquee*, vol. 65, no 5, p. 239-252.

Lang, Peter J., Mark K. Greenwald, Margaret M. Bradley et Alfons O. Hamm (1993). « Looking at pictures: Affective, facial, visceral, and behavioral reactions », *Psychophysiology*, vol. 30, no 3, p. 261-273.

Le Pailleur, Félix, Bo Huang, Pierre Majorique Léger et Sylvain Sénécal (2020). « A new approach to measure user experience with voice-controlled intelligent assistants: A pilot study », communication présentée au *HCII 2020*, Copenhagen, Denmark.

Léger, Pierre Majorique, Patrick Charland, Sylvain Sénécal et Stéphane Cyr (2017). « Predicting properties of cognitive pupillometry in human–computer interaction: A preliminary investigation », *Lecture Notes in Information Systems and Organisation*.

Léger, Pierre Majorique, Francois Courtemanche, Marc Fredette et Sylvain Sénécal (2019). « A cloud-based lab management and analytics software for triangulated human-centered research », dans, vol 29, Springer Heidelberg, p. 93-99.

Lewis, Bruce R., Gary F. Templeton et Terry Anthony Byrd (2005). « A methodology for construct development in mis research », *European Journal of Information Systems*, vol. 14, no 4, p. 388-400.

Lewis, James R. (2002). « Psychometric evaluation of the pssuq using data from five years of usability studies », *International Journal of Human-Computer Interaction*, vol. 14, no 3-4, p. 463-488.

Li, Cong (2009). « Primacy effect or recency effect? A long-term memory test of super bowl commercials », *Journal of Consumer Behaviour*, vol. 9, no 1, p. 32-44.

Li, Shanshi, Gabby Walters, Jan Packer et Noel Scott (2018). « A comparative analysis of self-report and psychophysiological measures of emotion in the context of tourism advertising », *Journal of Travel Research*, vol. 57, no 8, p. 1078-1092.

Lin, Aleck, Shirley Gregor et Michael Ewing (2008). « Developing a scale to measure the enjoyment of web experiences », *Journal of Interactive Marketing*, vol. 22, no 4, p. 40-57.

Lindgaard, Gitte, Gary Fernandes, Cathy Dudek et J. Browñ (2006). « Attention web designers: You have 50 milliseconds to make a good first impression! », *Behaviour and Information Technology*, vol. 25, no 2, p. 115-126.

Loiacono, Eleanor T., Richard T. Watson et Dale L. Goodhue (2007). « Webqual: An instrument for consumer evaluation of web sites », *International Journal of Electronic Commerce*, vol. 11, no 3, p. 51-87.

Loijens, Leanne, Olga Krips, Fabrizio Grieco, Hans van Kuilenburg, Marten den Uyl et Paul Ivan (2016). *Innovative solutions for behavioral research facereader ™ tool for automatic analysis of facial expressions reference manual version 7*. Récupéré de www.noldus.com

Lourties, Sébastien, Pierre Majorique Léger, Sylvain Sénécal, Marc Fredette et Shang Lin Chen (2018). « Testing the convergent validity of continuous self-perceived measurement systems: An exploratory study », communication présentée au *HCII 2018*, Las Vegas, USA.

MacKenzie, Scott B., Philip M. Podsakoff et Nathan P. Podsakoff (2011). *Construct measurement and validation procedures in mis and behavioral research: Integrating new and existing techniques*, vol. 35, 293-334 p.

Mahlke, Sascha et Manfred Thüring (2007). « Studying antecedents of emotional experiences in interactive contexts », communication présentée au *Conference on Human Factors in Computing Systems*, 2007, New York, New York, USA.

Maia, Camila Loiola Brito et Elizabeth Sucupira Furtado (2019). « An approach to analyze user's emotion in hci experiments using psychophysiological measures », *IEEE Access*, vol. 7, p. 36471-36480.

Mecredy, Philip, Malcolm J. Wright et Pamela Feetham (2018). « Are promoters valuable customers? An application of the net promoter scale to predict future customer spend », *Australasian Marketing Journal*, vol. 26, no 1, p. 3-9.

Moore, Gary C. et Izak Benbasat (1991). « Development of an instrument to measure the perceptions of adopting an information technology innovation », *Information Systems Research*, vol. 2, no 3, p. 192-222.

Murdock, Bennet B. (1962). « The serial position effect of free recall », *Journal of Experimental Psychology*, vol. 64, no 5, p. 482-488.

Murphy, Jamie, Charles Hofacker et Richard Mizerski (2006). « Primacy and recency effects on clicking behavior », *Journal of Computer-Mediated Communication*, vol. 11, no 2, p. 522-535.

Nagy, Mark S. (2002). « Using a single-item approach to measure facet job satisfaction », *Journal of Occupational and Organizational Psychology*, vol. 75, no 1, p. 77-86.

Nunnally, Jum C. et Ira H. Bernstein (1994). *Psychometric theory*, 3e éd., McGraw-Hill, 752-752 p.

Ordoñana, Juan R., Francisca González-Javier, Laura Espín-López et Jesús Gómez-Amor (2009). « Self-report and psychophysiological responses to fear appeals », *Human Communication Research*, vol. 35, no 2, p. 195-220.

Organisation internationale de normalisation et internationale Commission électrotechnique (2016). *Systems and software engineering : Systems and software quality requirements and evaluation (square) : Measurement of system and software product quality = ingénierie des systèmes et du logiciel : Exigences de qualité et évaluation des systèmes et du logiciel (square) : Mesurage de la qualité du produit logiciel et du système*, ISO/IEC, (1st ed.), c. viii, 45 p.

Ortiz de Guinea, Ana, Ryad Titah et Pierre Majorique Léger (2013). « Measure for measure: A two study multi-trait multi-method investigation of construct validity in is research », *Computers in Human Behavior*, vol. 29, no 3, p. 833-844.

Paas, Fred, Juhani E. Tuovinen, Huib Tabbers et Pascal W. M. Van Gerven (2003). « Cognitive load measurement as a means to advance cognitive load theory », *Educational Psychologist*, vol. 38, no 1, p. 63-71.

Peter, J. Paul (1979). « Reliability: A review of psychometric basics and recent marketing practices », *Journal of Marketing Research*, vol. 16, no 1, p. 6-6.

Peterson, Robert A. et William R. Wilson (1992). « Measuring customer satisfaction: Fact and artifact », *Journal of the Academy of Marketing Science*, vol. 20, no 1, p. 61-71.

Phan, Mikki H., Joseph R. Keebler et Barbara S. Chaparro (2016). « The development and validation of the game user experience satisfaction scale (guess) », *Human Factors*, vol. 58, no 8, p. 1217-1247.

Pingitore, Gina, Neil A. Morgan, Lopo L. Rego, Adriana Gigliotti et Jay Meyers (2007). « The single-question trap », *Marketing Research*, vol. 19, no 2, p. 8-13.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee et Nathan P. Podsakoff (2003). « Common method biases in behavioral research: A critical review of the literature and recommended remedies », *Journal of Applied Psychology*, vol. 88, no 5, p. 879-879.

Pointilist (2019). *State of customer journey management & cx measurement*, 31-31 p.

Pollack, Birgit Leisen et Aliosha Alexandrov (2013). « Nomological validity of the net promoter index question », *Journal of Services Marketing*, vol. 27, no 2, p. 118-129.

Razavi, Seyed Mostafa, Hossein Safari, Hessam Shafie et Hadi Rezaei Vandchali (2012). « How customer satisfaction, corporate image and customer loyalty are related? », *European Journal of Scientific Research*, vol. 78, no 4, p. 588-596.

Reichheld, Frederick F. (2003). « The one number you need to grow », *Harvard Business Review*, vol. 81, p. 46-54+124.

Riedl, René, Fred D. Davis et Alan R. Hevner (2014). « Towards a neurois research methodology: Intensifying the discussion on methods, tools, and measurement », *Journal of the Association for Information Systems*, vol. 15, p. 1-35.

Riedl, René et Pierre-Majorique Léger (2016). *Fundamentals of neurois*, Berlin, Heidelberg, Springer Berlin Heidelberg, coll. Studies in neuroscience, psychology and behavioral economics.

Rossiter, John R. (2002). « The c-oar-se procedure for scale development in marketing », *International Journal of Research in Marketing*, vol. 19, no 4, p. 305-335.

Russel, James A. (1980). « A circumplex model of affect », *Journal of Personality and Social Psychology*, vol. 39, no 6, p. 1161-1178.

Safdar, By Khadeeja et Inti Pacheco (2019). « The dubious management fad sweeping corporate america », *Wall Street Journal*, p. 1-10.

Safdar, By Khadeeja et Inti Pacheco (2019). « The dubious management fad sweeping corporate america », *Wall Street Journal*, p. 1-10.

Sarstedt, Marko et Petra Wilczynski (2009). « More for less? A comparison of single-item and multi-item », *Die Betriebswirtschaft*, vol. 69, no 2, p. 211-227.

Schmutz, Peter, Silvia Heinz, Yolanda Métrailler et Klaus Opwis (2009). « Cognitive load in ecommerce applications—measurement and effects on user satisfaction », *Advances in Human-Computer Interaction*, vol. 2009, p. 1-9.

Seo, Kwang Kyu, Sangwon Lee, Byung Do Chung et Changsoon Park (2015). « Users' emotional valence, arousal, and engagement based on perceived usability and aesthetics for web sites », *International Journal of Human-Computer Interaction*, vol. 31, no 1, p. 72-87.

Skiendziel, Tanja, Andreas G. Rösch et Oliver C. Schultheiss (2019). « Assessing the convergent validity between the automated emotion recognition software noldus facereader 7 and facial action coding system scoring », *PLoS ONE*, vol. 14, no 10, p. e0223905-e0223905.

Statista (2020). *Ways of conducting bank transactions in canada 2018*. Récupéré le 9 décembre 2020 de https://www-statista-com.proxy2.hec.ca/statistics/709927/canadian-banking-transactions-methods/

Steiger, James H. (1980). « Tests for comparing elements of a correlation matrix », *Psychological Bulletin*.

Straub, Detmar, Marie-Claude Boudreau et David Gefen (2004). « Validation guidelines for is positivist research », *Communications of the Association for Information Systems*, vol. 13, p. 380-427.

Straub, Detmar W. (1989). « Validating instruments in mis research », *MIS Quarterly: Management Information Systems*, vol. 13, no 2, p. 147-165.

Streiner, David L. (2003). « Starting at the beginning: An introduction to coefficient alpha and internal consistency », *Journal of Personality Assessment*, vol. 80, no 1, p. 99-103.

Tams, Stefan, Jason Thatcher, Kevin Hill, Varun Grover et Ana Ortiz de Guinea (2014). « Neurois—alternative or complement to existing methods? Illustrating the holistic effects of neuroscience and self-reported data in the context of technostress research », *Journal of the Association for Information Systems*, vol. 15, p. 723-753.

Thielsch, Meinald T., Iris Blotenberg et Rafael Jaron (2014). « User evaluations of websites: From first impression to recommendation », *Interacting with Computers*, vol. 26, no 1, p. 89-102.

Thompson, Ed, Don Scheibenreif et Michael Chiu (2020). *How to manage customer experience metrics*, 20-20 p. Récupéré de https://www.gartner.com/document/3979139?ref=lib

Thüring, Manfred et Sascha Mahlke (2007). « Usability, aesthetics and emotions in human-technology interaction », *International Journal of Psychology*, vol. 42, no 4, p. 253-264.

Tomarken, Andrew J. (1995). « A psychometric perspective on psychophysiological measures », *Psychological Assessment*, vol. 7, no 3, p. 387-395.

Tractinsky, N., A. S. Katz et D. Ikar (2000). « What is beautiful is usable », *Interacting with Computers*, vol. 13, no 2, p. 127-145.

Tse, David K. et Peter C. Wilton (1988). « Models of consumer satisfaction formation: An extension », *Journal of Marketing Research*, vol. 25, no 2, p. 204-212.

Wang, Xuehua (2011). « The effect of inconsistent word-of-mouth during the service encounter », *Journal of Services Marketing*, vol. 25, no 4, p. 252-259.

Wanous, John P. et Arnon E. Reichers (1996). « Estimating the reliability of a single-item measure », *Psychological Reports*, vol. 78, no 2, p. 631-634.

Wanous, John P., Arnon E. Reichers et Michael J. Hudy (1997). « Overall job satisfaction: How good are single-item measures? », *Journal of Applied Psychology*, vol. 82, no 2, p. 247-252.

Wirtz, Jochen et Meng Chung Lee (2003). « An examination of the quality and context-specific applicability of commonly used customer satisfaction measures », *Journal of Service Research*, vol. 5, no 4, p. 345-355.

Xie, Heping, Fuxing Wang, Yanbin Hao, Jiaxue Chen, Jing An, Yuxin Wang*, et al.* (2017). « The more total cognitive load is reduced by cues, the better retention and transfer of multimedia learning: A meta-analysis and two meta-regression analyses », *PLoS ONE*, vol. 12, no 8.

Zhou, Ronggang, Xiaorui Wang, Yuhan Shi, Renqian Zhang, Leyuan Zhang et Haiyan Guo (2019). « Measuring e-service quality and its importance to customer satisfaction and loyalty: An empirical study in a telecom setting », *Electronic Commerce Research*.

Zijlstra, F. R. H. et L. Van Doorn (1985). *The construction of a scale to measure subjective effort*, Delft.