HEC MONTRÉAL

La validité convergente des outils de mesure continu des émotions lors de la navigation sur une interface

Par

Sébastien Lourties

Mémoire par articles présenté en vue de l'obtention du grade de

Maîtrise ès science en gestion

(M.Sc.)

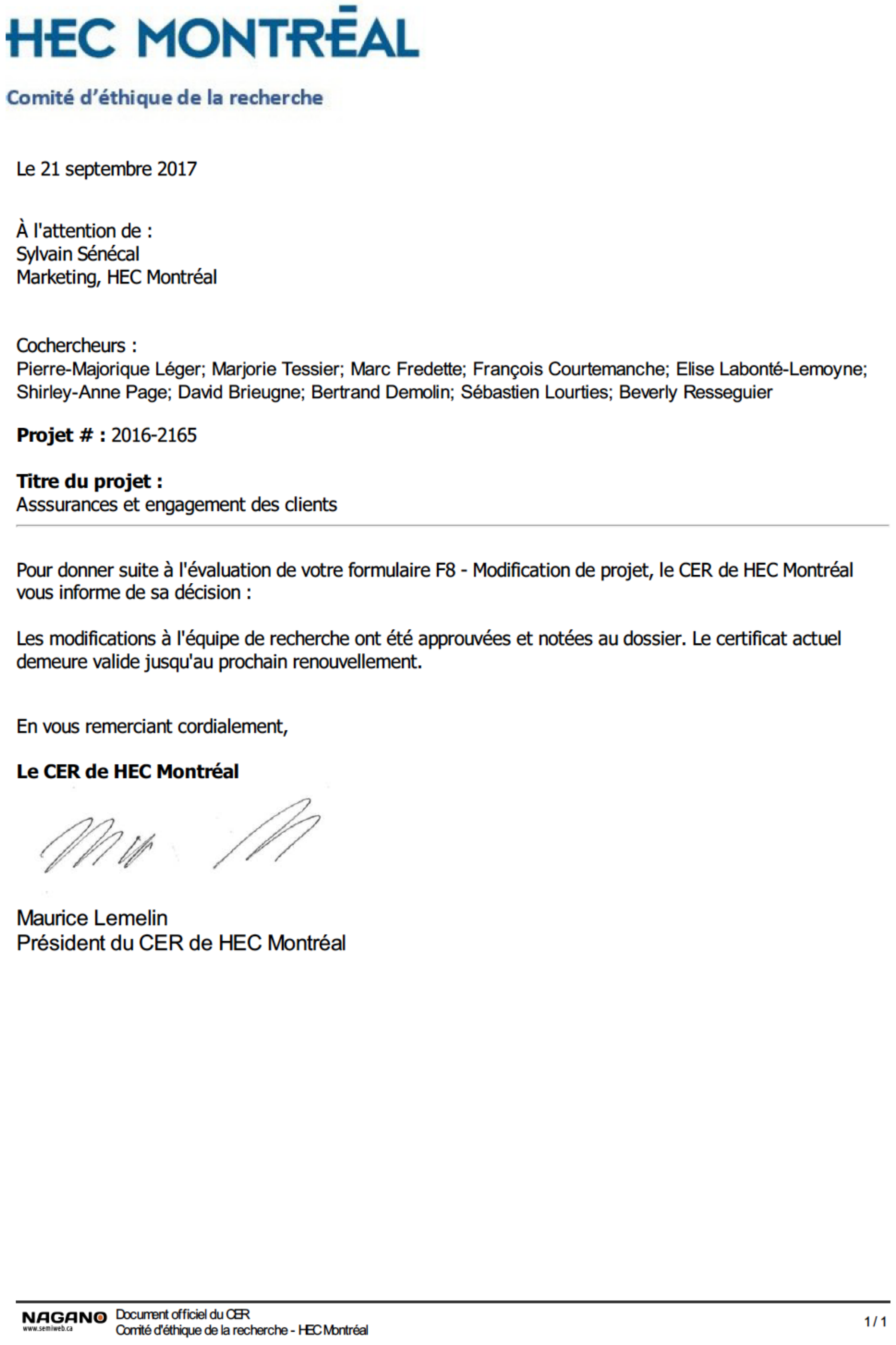
Sous la direction de

Pierre-Majorique Léger, Ph.D. et Sylvain Sénécal, Ph.D.

Sciences de la gestion

Technologies de l'information

Août 2018



HEC MONTRĒAL

Comité d'éthique de la recherche

Le 21 septembre 2017

À l'attention de :
Sylvain Sénécal
Marketing, HEC Montréal

Cochercheurs :
Pierre-Majorique Léger; Marjorie Tessier; Marc Fredette; François Courtemanche; Elise Labonté-Lemoyne; Shirley-Anne Page; David Brieugne; Bertrand Demolin; Sébastien Lourties; Beverly Resseguier

Projet # : 2016-2165

Titre du projet :
Asssurances et engagement des clients

Pour donner suite à l'évaluation de votre formulaire F8 - Modification de projet, le CER de HEC Montréal vous informe de sa décision :

Les modifications à l'équipe de recherche ont été approuvées et notées au dossier. Le certificat actuel demeure valide jusqu'au prochain renouvellement.

En vous remerciant cordialement,

Le CER de HEC Montréal

Maurice Lemelin
Président du CER de HEC Montréal

Sommaire

Ce mémoire par article étudie la validité convergente des outils de report continu des émotions par rapport à deux construits émotionnels : la valence et l'activation. Les outils de report continu des émotions permettent aux utilisateurs de revisionner rétrospectivement leur interaction avec une interface et de l'évaluer en continu. Plus précisément, cette étude présente l'analyse de biais systématiques comme la première impression, la dernière impression ainsi que les pics émotionnels qui affectent l'évaluation rétrospective d'une interface.

Pour étudier la validité convergente de cet outil ainsi que ces biais systématiques, nous avons réalisé deux études en laboratoire avec 24 participants. Chacune de ces études s'est déroulée dans un contexte utilitaire, les participants interagissaient avec l'interface et ils devaient ensuite reporter en continu les émotions qu'ils avaient vécues en visionnant à nouveau l'enregistrement de leur propre interaction. En enregistrant les données psychométriques durant l'interaction, nous pouvions comparer la convergence de la valence et de l'activation à travers deux outils, mais aussi étudier l'effet des biais systématiques.

L'analyse statistique des données suggère premièrement que les outils de report continu des émotions ne convergent pas vers les mêmes valeurs que celles des outils psychophysiologiques. Conscients que les outils rétrospectifs de report des émotions sont sujets à des biais, nous avons analysé plus précisément ces moments. Par ailleurs, il semblerait aussi que les participants avaient beaucoup plus de facilité (précision) à reporter leur valence négative ainsi que leurs moments d'activation positive.

Cette étude contribue à la recherche dans le domaine des instruments de mesure de l'expérience utilisateur, mais aussi sur les biais pouvant influencer la perception des utilisateurs lors d'une expérience. En effet, elle compare deux outils majeurs en expérience utilisateur : les outils psychophysiologiques et rétrospectifs. D'un point de vue pratique, cette étude offre aux concepteurs et développeurs d'interfaces une vue d'ensemble sur les

choix d’outils pour l’évaluation de l’expérience utilisateurs (psychophysiologiques et report continu), mais aussi de la difficulté pour les utilisateurs à correctement rapporter leur expérience a posteriori avec les outils de report continu des émotions.

Table des matières

Liste des figures, tables, scénarios et annexes

Remerciements

Cette année a été faite de grandes réussites, d’essais et d’erreurs. Elle témoigne de nombreux sacrifices, accompagnés de doutes, de quelques nuits blanches, mais surtout de passion, de bonheur et de plaisir.

Mes premiers remerciements iront à mes co-directeurs de mémoire, M. Pierre-Majorique Léger et M. Sylvain Sénécal qui m’ont enseigné une nouvelle manière de structurer ma pensée, qui ont façonné un peu plus mon esprit critique et qui m’ont surtout appris à remettre en question le statu quo. Ce fut un immense privilège d’avoir bénéficié de leur connaissance et de m’avoir donné l’opportunité, il y a maintenant presque un an, de démarrer cette passionnante aventure au sein du Tech3Lab. J’aimerais aussi lever mon chapeau aux assistant(e)s de recherche ainsi que toute l’équipe d’opération, notamment Bertrand, David et Beverly qui font un travail formidable pour permettre à tous les étudiants de donner vie à leurs projets, dont le mien. Merci, j’espère que nos chemins se recroiseront bientôt. Un grand merci aux deux statisticiens que j’ai côtoyés durant mon mémoire pour leur patience et de m’avoir formé sur de nombreux outils, Carl Saint-Pierre et Shang Lin Shen, merci infiniment.

J’aimerais bien évidemment remercier ma famille, qui a plus de 6000 km d’ici, a aussi fait de grands sacrifices tant sur le plan moral que financier pour me voir réussir aujourd’hui. Papa, Maman, merci infiniment ! Je voudrais aussi remercier des collègues devenus de vrais amis : Benjamin Maunier, Christophe Lazure, Tanguy Dargent, merci pour votre soutien moral tout au long du processus. Enfin, merci à mes proches notamment Victor Crest, Colin Gabla, Fred Buffaras et Emmanuelle Rault de m’avoir sortie de mon quotidien pas toujours évident cette année. Finalement, merci à Moncef Ayoun pour ses relectures!

Pour finir, je voudrais remercier le Conseil de Recherches en Sciences Naturelles et en Génie du Canada et la Chaire en Expérience Utilisateur (UX) pour leurs soutiens financiers me permettant de réaliser cette recherche dans les meilleures conditions possibles. Enfin, mon partenaire d’expérience Desjardins Assurances pour l’opportunité d’un projet concret.

Avant-Propos

Ce mémoire a été rédigé par article suite à l'approbation de la direction du programme de la Maîtrise ès sciences en gestion en option Technologie de l'Information.

Les consentements des coauteurs des deux articles ont été obtenus afin de les inclure dans ce mémoire.

Le comité d'éthique de la recherche de HEC Montréal a donné son approbation pour cette expérience en novembre 2017 (# certificat 2016-2165).

Le premier article cherche à comprendre de quelle manière la première et dernière impression peuvent avoir un impact sur la validité convergente de deux construits : la valence et l'activation. Il a été soumis et accepté en février 2017 à *HCI International* et a fait l'objet d'une présentation en juillet 2018 à Las Vegas *(*Lourties et al. 2018).

Le deuxième article tentera de confirmer et d'approfondir les résultats du premier article, en étudiant d'autres biais systématiques, notamment les pics émotionnels. Nous y observons de manière plus précise comment les construits de valence et d'activation sont reporté par les participants, mais aussi l'influence des biais sur ces derniers. Cet article est actuellement en révision finale pour être soumis dans la revue *Computer in Human Behavior* après la complétion du mémoire.

CHAPITRE 1 : Problématique et question de recherche

Mise en contexte et justification de l'étude

L'expérience utilisateur est définie comme la perception d'une personne qui résulte de l'utilisation ou de son anticipation d'un produit, un système ou un service (ISO, 2008). L'expérience utilisateur se déroule en trois temps car elle inclut tout d'abord ce qui est vu, touché et pensé du système ou du produit avant l'expérience. Elle comprend ensuite l'utilité, la facilité d'utilisation et l'impact émotionnel durant l'interaction. Enfin, le souvenir que l'utilisateur conservera de son expérience (Hartson & Pyla, 2012).

La pratique de l'expérience utilisateur est en plein essor au Québec et reste une des priorités des entreprises. En effet, d'après une enquête réalisée par le CEFRIO (2015), sur 534 entreprises interrogées ayant des défis marketing, 21,7% souhaitent améliorer l'expérience d'achat en ligne (deuxième priorité). Parmi les entreprises ayant des défis technologiques (n=230), 57,4% souhaitent faire évoluer leurs technologies (première priorité). Ainsi, les technologies et les interactions avec les humains sont au cœur des préoccupations de nombreuses entreprises, dont celle de notre partenaire d'expérience, une institution financière. Ainsi, d'après une enquête menée par PWC (2017), 81% des PDG d'institutions financières canadiennes estiment que la perturbation apportée par les technologies impactera le comportement du consommateur et donc leurs entreprises d'ici les cinq prochaines années. Ces différentes enquêtes mettent en lumière le fait que les technologies de l'information font désormais partie d'une expérience quotidienne pour les consommateurs (McCarthy & Wright, 2004).

Dans ce mémoire nous nous intéressons à l'expérience vécue pendant et après l'utilisation d'une interface, c'est-à-dire à l'interaction humaine-machine, mais aussi aux outils technologiques permettant d'en faire l'évaluation (pendant et après). En effet, l'interaction humaine-machine a, pendant de nombreuses années, été mesurée post-tâche, de manière rétrospective, grâce à des outils de report d'émotion. Ils peuvent être verbaux (Mehrabian, 1998; Scherer, 2005; Scherer et al, 2013; Waston & Clark, 1988) pour permettre aux utilisateurs de s'exprimer verbalement sur l'expérience ou bien non verbaux (Betella &

Verschure, 2016; Bradley & Lang, 1994; Broekens & Brinkman, 2013; Isbister & al, 2007; Pollak et al, 2011). Par exemple, l'échelle de mesure "Self-Assement Manikin" (Bradley & Lang, 1994) est largement utilisée par la communauté scientifique ainsi que les praticiens pour mesurer les construits de valence et d'activation émotionnelle.

Aujourd'hui, de nouveaux outils de reports des émotions ont vu le jour: les outils de report continu des émotions. Ces outils permettent aux utilisateurs de visionner à nouveau leur propre interaction et d'utiliser un cadran en deux dimensions, valence en abscisse et activation en ordonnée, avec une manette (joystick) pour reporter leur émotion (se positionner sur le cadran). Ces outils sont la plupart du temps en libre accès et permettent aux utilisateurs d'évaluer continuellement leurs émotions (Cowie & al, 2000 ; Girard & Wright, 2017 ; Nagel & al, 2007). Ces outils présentent de nombreux avantages que les outils « classiques » de report des émotions n'ont pas. En effet, la précision et le volume de données généré par ces outils sont importants, car ils ont une fréquence de données de 0,25/s (fréquence la plus haute) et enregistrent en continu tout au long de l'expérience permettant de revisiter l'expérience pour en identifier les moments non optimaux (contrairement aux outils « traditionnels » de report des émotions). De plus, ces instruments permettent d'éviter en partie, la reconstruction d'image (le fait de se remémorer les évènements, pouvant être déformés) que l'utilisateur pourrait faire avec un outil classique et biaiser son évaluation (Christianson & Safer, 1996; Levine, 1997). Enfin, le visionner à nouveau permet d'éviter la perte d'intensité de l'émotion qui a lieu dans la partie post-tâche d'une expérience (Walker & al, 1997). Ainsi, au lieu d'obtenir un score moyen de l'expérience utilisateur il est possible d'avoir des renseignements sur les moments non-optimaux de l'expérience. Ces outils rendent les prises de décisions UX, c'est-à-dire les améliorations de design et d'expérience utilitaire globale, plus certaines.

Avec l'introduction des outils neurophysiologiques ou psychophysiologiques dans le domaine des technologies de l'information (Riedl, et al, 2009), il est désormais possible de capter l'expérience vécue en tout temps par l'utilisateur, et ce, de manière précise, naturelle et non-obstrusive (Ortiz de Guinea, Titah, & Leger, 2014). Il est ainsi possible de mesurer des construits caractéristiques de l'expérience utilisateur comme la valence, l'activation ou la charge cognitive (Dimoka, Pavlou, & Davis, 2011; Ortiz de Guinea & Markus, 2009; Riedl et al., 2009) en utilisant des données psychophysiologiques.

Ainsi, en utilisant ces instruments, nous avons voulu comparer les données générées par les outils de report continu des émotions avec celles générés par les outils psychophysiologiques. Néanmoins, dans toutes rétrospections suite à une expérience vécue, des biais systématiques de première/dernière impression ainsi que les pics émotionnels influencent le résultat de la rétrospection (Kahneman et al, 1993 ; Fredrickson & Kahneman, 1993). La rétrospection et ces biais sont des moments importants, car ils conditionnent la volonté de réutilisation de l'interface (Kahneman 2000; Kahneman et al, 1993). C'est pour cette raison que nous nous sommes intéressés à ces moments précis, pour observer si ces biais étaient présents lors de la rétrospection faite avec les outils de report continu des émotions.

Objectifs de l'étude et questions de recherche

Ce mémoire a pour objectif d'étudier la validité convergente des outils de mesure continue des émotions (à travers la valence et l'arousal), mais aussi les biais systématiques comme la première, dernière impression et pics émotionnels qui viennent impacter la capacité de report émotionnel des utilisateurs. Les outils de report des émotions permettent aux participants de visionner le contenu émotionnel tel qu'ils le perçoivent à travers le temps, permettant d'étudier les épisodes de dynamiques émotionnelles (Girard & Wright, 2017).
Cette étude a été réalisée dans un contexte utilitaire (c'est à dire qui se concentre sur des tâches du quotidien ; par opposition à des tâches destinées à créer du plaisir) et vise à comprendre davantage ces nouveaux outils permettant de mesurer avec précision l'expérience utilisateur vécue, de manière rétrospective.

Un utilisateur, lorsqu'il navigue sur une interface, aura une réaction émotionnelle. Cette dernière lorsqu'elle est rapportée suite à l'expérience n'est pas tout à fait fidèle à ce qu'il s'est réellement déroulé, car l'utilisateur est influencé par certains moments, notamment ceux de très forte ou très faible intensité (Kahneman et al, 1993 ; Fredrickson & Kahneman, 1993). Dans un deuxième article, nous nous intéressons aux pics émotionnels et la manière donc ils influencent les deux construits de valence et d'activation émotionnelle. De fait, nous nous posons les deux questions suivantes :

Question 1 : Est-ce que la valence et l'activation émotionnelles vécues en début et fin d'interaction correspondent à la valence et à l'activation rapportées par les utilisateurs (i.e., première et dernière impression) à l'aide d'outils de report continu ?

Question 2 : Est-ce que la valence et l'activation émotionnelles vécues lors des pics émotionnels correspondent à la valence et à l'activation rapportées par les utilisateurs (i.e., pics émotionnels) à l'aide d'outils de report continu ?

Pour répondre à ces deux questions de recherche, nous avons réalisé deux études en laboratoire avec 24 participants et deux interfaces différentes. L'expérience s'est déroulée en deux temps, la première en octobre 2017 et la deuxième en février 2018, chacune durant 1h30 et approuvée par le comité d'éthique de HEC Montréal (CER). Les participants devaient magasiner sur les interfaces et ensuite visionner à nouveau leur interaction en l'évaluant. Dans chacune des interfaces, il y avait entre 14 et 19 séquences (pages différentes) que nous étudions séparément pour y observer les moments de première/dernière impression et pics émotionnels.

Pour évaluer l'expérience vécue durant l'interaction, nous avons utilisé la valence et l'activation émotionnelle. La valence fut enregistrée par le logiciel Facereader (Noldus, Wageningen, Pays-Bas) grâce aux micro-mouvements du visage et contraste l'état de l'utilisateur entre ses émotions positives et négatives. L'activation émotionnelle fut enregistrée avec le logiciel Acknowledge (Biopac, Goleta, USA) grâce à des senseurs placés sur la paume de la main, captant le niveau d'activité électrodermale (sudation). L'activation émotionnelle permet de contraster le niveau d'excitation de l'utilisateur. Après l'expérience, nous avons enregistré la rétrospection des utilisateurs avec le logiciel DARMA (libre d'accès) permettant d'évaluer la valence et l'activation émotionnelle en même temps.

Contributions potentielles

D'un point de vue théorique, ce mémoire contribue tout d'abord à la littérature sous-jacente à la rétrospection d'un utilisateur, car nous testons les outils de report continu des émotions (Gerard & Wright, 2017 ; Cowie et al, 2000 ; Nagel et al, 2007), mais aussi à la littérature sur les biais systématiques qui impactent l'utilisateur (Kahneman et al, 1993 ; Fredrickson & Kahneman, 1993). Il permet de mieux comprendre les résultats proposés par les outils de report continu des émotions comparativement à ceux proposés par les outils psychophysiologiques (Boucsein, 2012 ; Léger & al, 2012 Ortiz de Guinea & al, 2014). Plus précisément, ce mémoire contribue aux études sur les biais systématiques comme la première/dernière impression, ainsi que les pics émotionnels (Kahneman et al, 1993 ; Fredrickson & Kahneman, 1993), mais aussi sur le caractère positif ou négatif de ces biais (Baumeister & al, 2001).

D'un point de vue pratique, ce mémoire aide les concepteurs d'interfaces tout d'abord dans la sélection d'outils pour évaluer l'expérience utilisateur lors d'une navigation, mais aussi les biais pouvant influencer positivement ou négativement l'expérience dans son ensemble. Étant donné l'intérêt grandissant pour l'expérience utilisateur (CEFRIO, 2015), les praticiens pourront donc anticiper le processus de développement d'interfaces en sélectionnant les bons outils et en se concentrant sur les bons moments.

Structure du mémoire

Ce mémoire est structuré sous la forme de deux articles s'adressant aux mêmes types de lecteurs. Le premier article permet de poser les bases pour le second. Le premier article est exploratoire et permet de comprendre si la première et dernière impression ont un impact sur la validité convergente de deux construits : la valence et l'activation émotionnelle. Dans le second article, nous étudions la même question, mais en nous intéressant aux pics émotionnels vécus par les participants. Cet article permet d'obtenir des résultats plus robustes quant aux conclusions du premier article. Ces deux articles s'adressent à des chercheurs et des professionnels UX ayant la volonté d'acquérir des outils de mesures d'expérience utilisateur, mais aussi à des développeurs d'interfaces désireux d'améliorer l'expérience utilisateur.

Information sur l'article 1

Ce premier article a été présenté à la conférence scientifique à *HCI International 2018* en juillet 2018. Ce premier article porte sur la première phase de recherche menée en octobre 2017 (n=13). Cet article a pour objectif de mener une étude exploratoire sur l'impact de la première et dernière impression sur la validité convergente des construits de valence et d'activation émotionnelle. Les résultats de cette première phase de recherche sont à paraître dans le multivolume *Springer* « HCI in Business, Government and Organizations » dans la rubrique « Neuro/IS » (Lourties et al. 2018). Cet article nous a permis d'orienter notre recherche pour le deuxième article.

Résumé de l'article 1

Cet article permet de comprendre les rôles de la première/dernière impression sur une interface, sur les construits physiologique de valence et d'activation émotionnelles. En d'autres termes, lorsque vous rapportez une expérience, quels construits et quels biais psychologiques sont susceptibles de plus vous influencer dans votre perception de l'expérience? Les données psychophysiologiques telles que la reconnaissance des émotions faciales et l'activité électrodermale ont été enregistrées pour mesurer la valence et l'activation émotionnelle durant l'expérience. Après l'interaction avec l'interface, les construits de valence et d'activation émotionnelle ont été mesuré grâce à un logiciel de report continu des émotions (DARMA). Les résultats suggèrent que les participants reportaient leur valence émotionnelle avec plus de précision à la fin des tâches qu'au début et reportaient leur activation de manière plus précise au début qu'à la fin des tâches. Nous concluons aussi l'effet de première impression à plus d'effet sur l'activation émotionnelle tandis que l'effet de dernière impression impacte davantage la valence émotionnelle.

Information sur l'article 2

Ce deuxième article est destiné à être soumis dans le journal « Computers in Human Behavior » après la soumission du mémoire. Ce journal vise à examiner les interactions entre les humains et les appareils électroniques du point de vue de la psychologie. Il s'adresse tant aux besoins de chercheurs que de praticiens. Cet article a pour objectif d'approfondir la recherche menée dans le premier article à travers une autre source de biais systématique : les pics émotionnels. Nous y étudions l'impact des pics émotionnels sur la validité convergente des construits de valence et d'activation émotionnelle.

Résumé de l'article 2

Dans ce deuxième article, nous observons l'expérience vécue et rapportée par l'utilisateur lors de la navigation sur l'interface. Cette observation s'est faite par le prisme de la théorie des pics émotionnels (Kahneman et al, 1993 ; Fredrickson & Kahneman, 1993) pour expliquer les biais que l'utilisateur subit lorsqu'il rapporte son expérience. Dans un même temps, nous évaluons les outils de report continu des émotions. L'article suggère trois conclusions : (1) les outils de report continu des émotions divergent en termes de résultats par rapport aux outils psychophysiologiques pour l'évaluation de l'expérience utilisateur, (2) les utilisateurs rapportent plus précisément leurs émotions négatives que positives, (3) les utilisateurs rapportent plus précisément leurs moments d'excitation que leurs moments d'ennuis. Pour ce faire, nous avons utilisé des outils psychophysiologiques pour mesurer les émotions de l'utilisateur lors de son interaction avec l'interface (reconnaissance des émotions faciales et activité électrodermale). Cette recherche aide les professionnels en expérience utilisateur dans leurs choix d'outils pour mesurer leurs travaux mais aussi sur la capacité des utilisateurs à reporter leurs émotions.

Afin de mieux comprendre mon apport aux deux articles, le tableau ci-dessous présente ma contribution durant les étapes du processus de recherche. Le pourcentage du travail que j'ai effectué est inclut à chacune des étapes.

Étapes du processus	Contribution
Définition des requis du partenaire de la Chaire UX (Desjardins Assurances)	Traduire les besoins du partenaire en question de recherche scientifique- **50%** - Définir les questions de recherche dans les articles - Les besoins d'affaires ont été recueillis par les chercheurs du Tech3Lab.
Revue de littérature	Effectuer la revue de littérature pour déterminer les construits testés dans le domaine des mesures neurophysiologiques et auto-rapportées – **100%** Définir les outils de mesures utilisés pour tester les construits – **75%** - Une aide du laboratoire fut fournie concernant l'évaluation des outils de mesure neurophysiologiques et auto-rapportés.
Stimuli	N/A : les stimuli ont directement été choisis par le partenaire de la Chaire.
Design expérimental	Concevoir le protocole d'expérimentation – **50%** : La majorité du design a été réalisé entre le partenaire et l'équipe d'opération du lab. J'ai inséré la portion recherche en collaboration avec les deux parties prenantes pour garantir la qualité des données tant pour le partenaire que pour la recherche.
Recrutement	Le partenaire de la Chaire a effectué un recrutement externe, nous devions seulement gérer les compensations.
Prétests et collectes	Responsable des opérations pour la partie de "recherche" du design expérimental – **100%** Participation à l'ensemble des collectes : Support technique et aide aux assistantes pour tout problème avec la salle de collecte – **75%**

Extraction et transformation des données	Extraction et mise en forme des données physiologiques, psychométriques, cognitives et émotionnelles pour permettre l'analyse statistique – **100%**
Analyse de données	Analyse des données psychophysiologiques – **100%** (lecteur d'émotions faciales, senseurs d'activité électrodermale) : Une formation effectuée par le Tech3lab m'a permis de mener de manière autonome cette analyse. Analyses statistiques – **75%** - Aide sur SPSS pour les analyses par le statisticien de la Chaire.
Rédaction	Contribution dans l'écriture des articles – **100%** - Les autres auteurs ont apporté des commentaires constructifs à des fins d'amélioration de la qualité de l'article.

Tableau 1: Contributions et responsabilités dans la rédaction des articles

CHAPITRE 2: Article 1

Testing the Convergent Validity of Continuous Self-Perceived Measurement Systems: An Exploratory Study[1]

Abstract. This paper explores the convergent validity of instruments that can provide higher temporal resolution when measuring user experience: the continuous self-perceived measurement system and psychophysiological measures. Specifically, we explore the extent to which primacy and recency effects may have an impact on the convergent validity of two constructs: valence and arousal. Using a Wilcoxon signed rank test, results suggest that users self-evaluate their valence more accurately at the end of each of the sequences than at the beginning while they evaluate their arousal more accurately at the beginning of each of the sequences. This suggests that the recency effect has more impact on valence and the primacy effect has more impact on arousal. These findings contribute to human-computer interaction research by providing more information about the psychophysiological measures that cause recency and primacy.

Keywords: continuous self-perceived measurement system; physiological measurement; primacy effect, recency effect.

[1] Lourties S., Léger PM., Sénécal S., Fredette M., Chen S.L. (2018) Testing the Convergent Validity of Continuous Self-Perceived Measurement Systems: An Exploratory Study. In: Nah FH., Xiao B. (eds) HCI in Business, Government, and Organizations. HCIBGO 2018. Lecture Notes in Computer Science, vol. 10923. Springer, Cham

Introduction

Research is calling for a multimethod approach in human computer interactions [53], [26], [48]. Multiple measurement approaches offer a richer perspective on user experience context and can enable UX researchers to gain better insight on what the user really experiences during a given task [21], [11].

Amongst the new methods proposed in the literature, new instruments offer better temporal resolution on user experience, i.e. measures that can provide a continuous measure of the experience over time. In contrast, non-continuous measure such as psychometric scales only provide a measure at a precise moment in time. Continuous measures can be very useful to designers as they can help to identify the timing of non-optimal experiences (in other words, pain points in an interactive experience). This article focuses on two of those measures: continuous self-perceived measurement systems (CSP) [12], [41], [33], [22], [23] and continuous psychophysiological measures [18], [32], [25], [2]. CSP are retrospective measurements that "let the observer track the emotional content of a stimulus as they perceive it over time, allowing the emotional dynamics of speech episodes to be examined" [12] (p.1). Psychophysiological measures are "an unobtrusive and implicit way to determine the user's affective or cognitive state on the basis of mind-body relations" [14] (p.1362).

It is of high importance to UX researchers to understand the extent to which these new instruments converge in measuring the same constructs. As those signals can evolve over time, they may not to coevolve in the same manner over time. Thus, the objective of this paper is to explore the convergent validity of two important constructs in UX research: valence and arousal. Specifically, we explore the extent to which primacy and recency effect may have an effect on the convergent validity of these constructs.

To answer this research question, we have conducted laboratory experiment with 13 participants performing a series of utilitarian tasks on an insurance company website for 15 minutes. Our results suggest that users self-evaluate their valence more accurately at the end of each of the sequences than at the beginning and evaluate their arousal more accurately at the beginning of each of the sequence. This could suggest that valence has more impact on recency effect and arousal has more impact on primacy effect.

This paper is organized as follows. First of all, in the literature review, we will introduce emotion in user experience, then we will talk about psychophysiological measures of emotion, then its self-perceived evaluation. After that, we will develop our hypothesis and we will explain our research methodology. Finally, we will present the results and discuss the implications.

Literature review: Measuring UX with high temporal resolution instruments

This article focuses on two types of UX measurements that provide high temporal resolution: neurophysiological measure and CSP measure. High temporal measures offer precision with respect to time during an experience. High temporal resolution measures are characterized by a sampling rate that defines the number of measures per minute.

Continuous self-perceived measures of user experience

Continuous self-perceived measurement systems have been proposed as a novel way to enable a user to dynamically report on their experience. The simplest tools to use are composed of only one dimension of the emotion (e.g. emotional valence). The CARMA software and the Emotion Slider [22], [33] were developed with one dimension in order to facilitate the report of basic emotion (negative vs positive), so participants just have to push up or down to report their affective state. For instance, Girard and Wright [23] propose a measurement system in which users operate a joystick to indicate their reactions to a stimulus on two dimensions (e.g., emotional arousal and valence). Before that, other systems have been proposed to measure in a such way emotion. Feel-Trace [12] and EmuJoy [41] were the first software packages to propose a CSP on two dimensions. Software with two dimensions were only tested in a hedonic context: FeelTrace, Emu-joy and DARMA were tested in the music or commercial ad context that had extreme (negative or positive) arousal and valence. Furthermore, even if the authors suggest the use of the tools for retrospection in a utilitarian context, they were mainly tested directly to evaluate music or commercials.

Although these systems provide richer information than self-reported scales due to their continuous nature, to our knowledge, no research has investigated their validity. As research suggests that any self-reported measure is subject to biases, e.g., retrospective bias, efforts to validate these continuous measurement systems will inform researchers and practitioners using such systems to evaluate user experience.

Then, we could wonder why some researchers have focus their activity on CSP. It comes from the limitations of the traditional self-reported measurement system [3]. However, as Lallemand & Grenier [31] report, the limited number of emotions reported in some tools could make it difficult for the participant to clearly express what they feel. Finally, many tools provide a scale and a score that doesn't explain what really happened during the interaction with the product/interface even if much progress has been made by researchers with the succession of tools put at our disposal. All of the self-reported tools provide the overall grade that the participant experience without explaining positive and negative emotions selected during their interaction [7], [8]. Moreover, this last limitation motivated researchers to explore CSP measurement systems. Also, it should be noted that, depending on the tool we use, important biases such as primacy and recency effects, could occur during the retrospection of the participant.

Continuous neurophysiological measures of user experience.

In the last decade, there has been a wide range of studies that employed tools and theories from neurosciences to inform the design of information systems in a user-computer interaction context [48], [42]. More specifically, measures of valence and arousal have been very informative [34]. Valence and arousal measurements are part of psychophysiological measures that are already defined above as "an unobtrusive and implicit way to determine the user's affective or cognitive state on the basis of mind-body relations" [14] (p.1362). Over the years, psychophysiological measures have received a lot of attention as they provide a large quantity of information to pose an accurate diagnostic on an emotional reaction caused by a stimuli [19]. Researcher such as Plutchik, [44] proposed that there are from two to twenty different emotions. In this article, we will focus our intention on the circumplex model that was created to represent emotions with arousal and valence in two dimensions [49], [45], because it has been reused for the construction of the continuous self-perceived system DARMA [23].

Arousal is the reaction to an emotional stimuli characterized by neurophysiological level of vigilance or a state of attention [29]. The arousal construct is used to contrast states of low arousal (e.g., calm) and high arousal (e.g., surprise) [2] and is a useful construct in the human computer literature. For example, Ravaja & al [46], in a video game context, find that the nature of the opponent (computer, friend, stranger) varied the impact on the arousal of the participant. To measure arousal, electrodermal activity (EDA) has been proposed as an accurate way to collect data using the variation of the electrical properties of the skin in response to the secretion of sweat from the palm of the hands in our case [25], [2].

Valence usually refers to “the ‘positive’ and ‘negative’ character of an emotion and/or of its aspects such as behavior, affect, evaluation, faces, adaptive value, etc.” [9]. The valence construct is used to contrast states of pleasure (e.g., happy) and displeasure (e.g., angry) [2]. For example, in the IT field, this construct was used as a determinant factor to create an intelligent tutoring system (AutoTutor) that measures the frustration of the student learning the content and interacting with it [39]. Detecting this negative valence makes it possible to improve the learning curve of the student. Sas & al [50], conducted an experiment to test most memorable moments and the attachment to the Facebook website and positive valence common denominators for all the participants. One way to measure this valence, is to use facial expression [15] because most individuals express their emotion with micro-movements of the facial muscles. Automatic facial analysis tools such as FaceReader (Noldus, Wageningen, Netherlands) permit to determine with relative precision, valence change in an experience and this, in real time with a high temporal precision (6 times per second) [52].

There are several advantages to taking psychophysiological measures into consideration when studying emotion. First, you can capture the unconscious or automatic reactions of a participant without interrupting the experiment to get the most natural reaction because it is non-intrusive [42]. Then, you can have data on the user reaction in real time which is a real advantage to avoid the common bias of memory [26]. Finally, you can combine different measures with neurophysiological tools to limits the common method bias [34].

In this article, we are measuring the neurophysiological states of the user using valence and arousal dimensions of emotion because we are then able to use these two dimensions that are well accepted by research [3], characterized by a two-dimensional cartesian plane. The arousal dimension goes from calm to excited whereas valence ranges from negative to positive.

Hypothesis development

While remembering an experience could be hard for someone, there exist some biases that consciously or unconsciously make this step harder.

This article aims at exploring the convergent validity of neurophysiological measures and CSP measures. Specifically, we explore to what extent the convergent validity is affected by primacy and recency bias. Primacy effect refers to the increased memory and over-weighted influence of the first moments of an experience [40], [51], [54] and recency effect correspond to the important influence of the last moment of an experience [28]. These two biases have the same root and are part of the larger topic of memory bias. To be concerned by the primacy and the recency effect you have to exercise a retrospection which is the capacity to remember the context of an experience and to explain the intention [10]. Several factors can influence the ability to remember an experience; valence and arousal are important ones.

Fredrickson & Kahneman [20] have suggested that the duration of an experience has little effect on the retrospective evaluation. They call it the “duration neglect”. They conducted an experiment with video clips, but also with physical experiences [28] suggesting the same results. Overall, in a free recall context, primacy and recency effect play a U-shape curve described in the serial position curve of Craik & Lockhart [13]. After the experience, items at the beginning and at the end of a series are better reported than those in the middle.

We posit H1: Primacy and recency effects have the same influence on the accuracy of self-reported evaluation

Fredrickson & Kahneman’s results [28] suggested that adding a positive experience at the end of the task improves the retrospective evaluation. Furthermore, researchers have found that an experience with a strong emotional relevance is more likely to be remembered than a more neutral experience [4], [24]. Valence and arousal contribute in different ways to the formation of memory, but studies have shown that arousal is a much more important factor when it is time to remember an experience [6]. First, the neural processes of memory are different for valence and arousal [30]. When remembering an experience with high arousal, different substances such as glucose are released into the bloodstream [38] and the memory using emotional arousal leads to peripheral and central nervous systems activation and intensively involves the amygdala [37] whereas the prefrontal cortex hippocampal network is used for valence [30].

Moreover, recency and primacy effects are part of the selectivity bias. As Mather & Sutherland [36] explained, we are surrounded by information every day and there's a “battle for a share of our limited attention and memory, with the brain selecting the winners and discarding the losers” (p.114). Indeed, arousal is enhancing the memory for specific details but is removing other collateral details [5]. Selectivity bias is common, our brain cannot recall each information that we have absorbed during the day, this is why recency and primacy effects have been found as one of the answers of this selection. Recency effect has been studied much more than the primacy effect due to the peak-end rule, characterized by the most intense moment at the end of an experience [28]. Different stimuli were used to test the recency effect such as medical pain [47]. Such studies found a strong relation between the intensity of pain recorded during the last 3 minutes of the treatment and the self-perceived evaluation.

We posit H2: Arousal has more influence than valence on the recency effect

Research method

To test our hypothesis, we conducted a laboratory experiment with 13 subjects (7 males and 6 females), between the age of 21 and 48 (mean of 32). We provide a 100$ compensation to each participant upon completion of the experiment. Participants had normal or corrected-to-normal vision and were pre-screened for glasses, laser eye surgery, astigmatism, epilepsy, neurological and psychiatric diagnoses. This project was approved by the Ethics Committee of our institution.

Stimuli and procedure

First, a scenario about an incident or a situation was presented to participants. Then, they had to perform a series of utilitarian tasks on website for 15 minutes in order to prepare a claim about an incident (e.g., provide details of the incident, make an appointment). Finally, they had to view their recorded interaction with the website and use a joystick to indicate their level of self-reported emotional arousal and valence continuously during the recording (Figure 1). Every participant had the same task to perform and the same instruction to use the joystick. Moreover, we trained the participants to be sure that they understood the use of the joystick and the use of the dial. To make sure that the experiment was going well, we conducted two pre-tests to assure the tools were working and recording the right data. Also, we wanted to be sure that the tasks performed as well as the instructions were understood by the participants. Minor changes were then made to finalize the protocol.

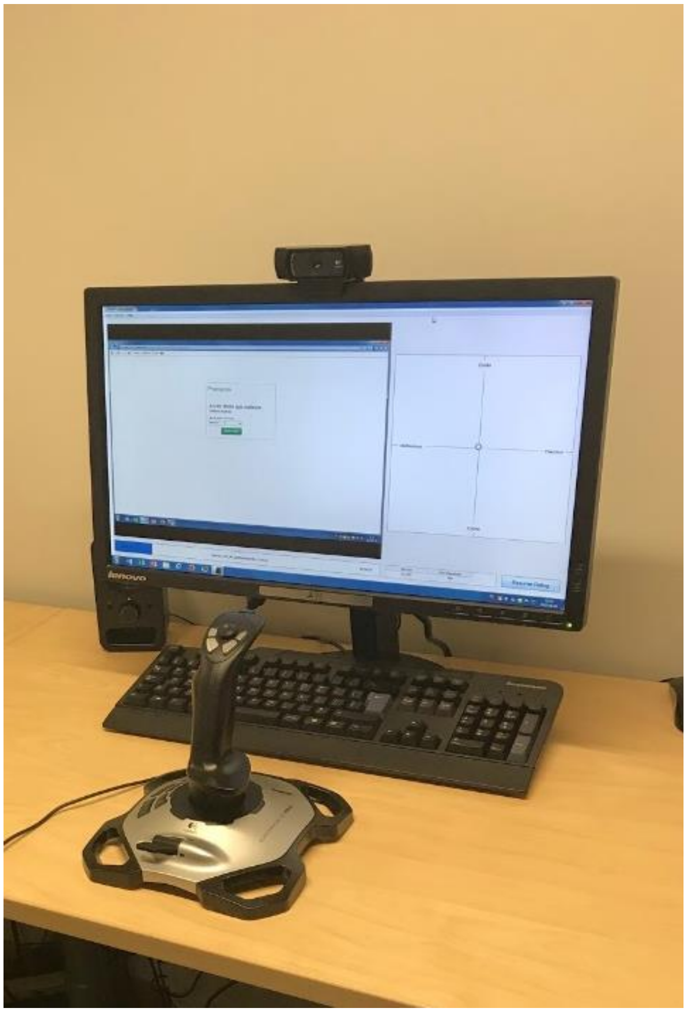


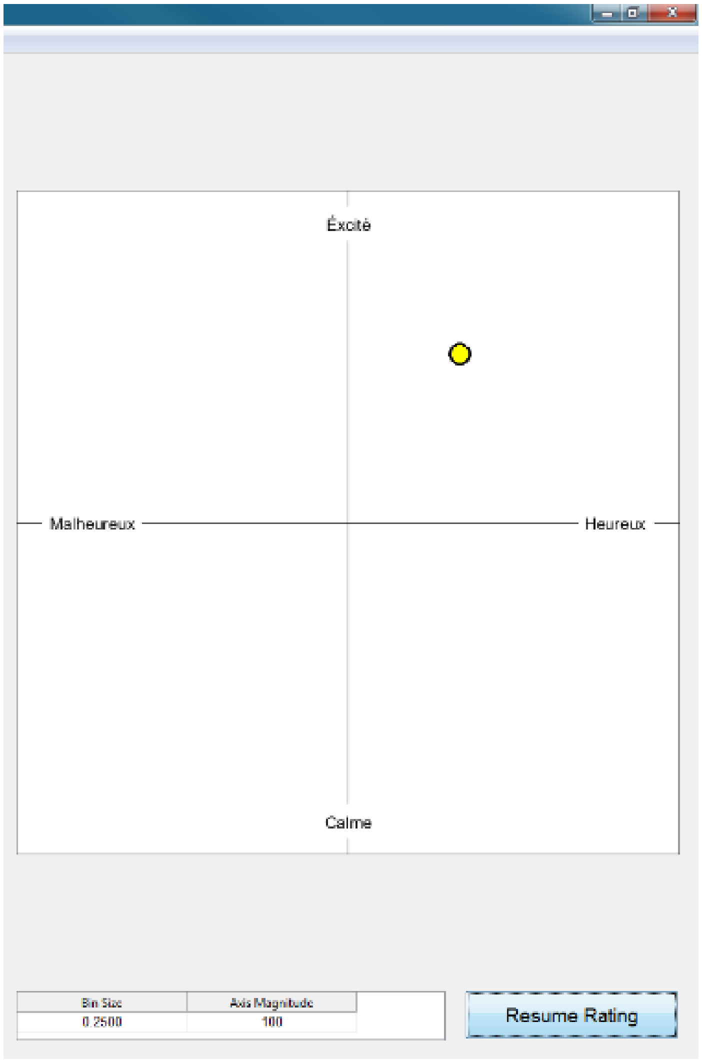
Figure 1: The set up during the self-perceived measure using DARMA software [23]

Instruments and measures

According to Ortiz de Guinea, Titah & Léger [43] physiological measures and self-perceived evaluations interact together. Thus, to test the validity of CSP measurement systems, we assess them with neurophysiological inferences of the same constructs (emotional valence and activation) using physiological activation and automatic facial analysis [42].

We first of all measured emotion with physiological tools. Users' emotional and cognitive states can be measured with physiological signals such as electrodermal activity, heart rate, eye-tracking and facial expressions [42] . This allows researchers and practitioners the possibility of collecting real time information on what the user is experiencing through the interaction. Regarding electrodermal activity (EDA), it has been used to measure physiological arousal [25], [2]. Emotional valence represents the direction of an emotional response negative versus positive [32]. In our case, facial expression and automatic facial analysis provide interesting insights for measuring human emotion [16], [1].

Participants physiological activation (i.e., electrodermal activity) during their interaction with the website was measured with the Acknowledge software mp150 sampled at 500Hz (BIOPAC, Goleta, USA). The FaceReader software (Noldus, Wageningen, Netherlands) was used to measure emotional valence, calculated using the value of the positive emotion minus the strongest of the negative emotions which results in a valence score from -1 to +1 [17]. Media Recorder (Noldus, Wageningen, Netherlands) was used to record participants' interaction with the website. Observer XT (Noldus, Wageningen, Pays-Bas) was used to synchronize the signals of these three recording devices as per guidelines provided by Léger & al [35]. Finally, participants self-reported continuous emotional reactions (arousal and valence) were measured using the DARMA software [23] (p. 1), which "synchronizes media playback and the continuous recording of two-dimensional measurements through the manipulation of a computer joystick to indicate changes in their emotional state" (see figure 2). The output of DARMA is an Excel file in which each line begins with a time code, then gives the valence and activation evaluation coordinates of the joystick at that time. All of the measures and instruments are summarized in Table 1.



Bin Size	Axis Magnitude
0.2500	100

Figure 2: The two-dimension self-report window during media playback [23].

As you can notice, the software was set at 30 Hz, a bin size of 0.25 seconds and an axis magnitude of 100. These measures are summarized in the Table 2.

	Self-reported (retrospective)	Physiological (during interaction)
Valence	Participant’s joystick position on the valence axis at time t (Darma)	Facial emotional valence at time t (Facereader)

Arousal	Participant’s joystick position on the arousal axis at time t (Darma)	Electrodermal activity for time t (Acqknowledge)

Table 2: Measures used for the experiment

Results

To test the convergent validity of CSP measurement, we compared the retrospection performance of the first and last time interval between self-reported and physiological measures.

The validity of CSP emotion, i.e. the degree to which it accurately measures the emotions it was designed to measure, was assessed by recording sessions, broken down into short sequences. Each participant had to report their emotional valence and arousal for 14 short sequences for a total of 182 sequences for the overall sample. Sixty percent of tasks lasted less than 30 seconds. To test the convergent validity of CSP measurement, we compared the retrospection performance of the first and last time interval of each one of the 14 sequences between self-reported and physiological measures. Thus, we calculate the distance between physiological valence and CSP valence in the first and last time interval and the same calculation was done with the distance between physiological arousal and CSP arousal (respectively dis_valence_first versus dis_valence_last and dis_arousal_first versus dis_arousal_last in Table 2). When the movement of the joystick was less than .001, we considered it as static and these data were excluded from the analysis. The 2-tailed p-value from the Wilcoxon signed rank test indicated whether the difference was significant or not.

With an interval of 1 second, we found that participants report more accurately their valence at the end of the task compared to the beginning with a mean difference of .0335 (.155 - .1215) and p-value = .0173. Also, participants report with more precision their arousal at the beginning of the task compared to the end with a mean difference of .0186 (.0377 - .0191) and p-value = .0635. Because sequences were short, the data analysis was performed using one-second time periods (e.g., comparing the first second of self-reported valence vs. physiological valence), but also with two-second periods.

Significant results were found for both time periods, the results for the one and two seconds time periods are reported in Table 3 and Table 4.

Below you can find the name of the variables and their meaning:

- distance _valence_first: the reported valence minus the experienced valence at the first moment of each sequence
- distance _valence_last: the reported valence minus the experienced valence at the last moment of each sequence
- distance _arousal_first: the reported arousal minus the experienced arousal at the first moment of each sequence
- distance _arousal_last: the reported arousal minus the experienced arousal at the last moment of each sequence

Pair of variables (first and last 1s)		Mean at beginning	Mean at end	P-value
distance_valence_first	distance_valence_last	.155	.1215	.0173
distance_arousal_first	distance_arousal_last	.0191	.0377	.0635

Table 3: Results for 1s interval between the beginning and the end of each one of the sequences

Pair of variables (first and last 2s)		Mean at beginning	Mean at end	P-value
distance_valence_first	distance_valence_last	.153	.1235	.0143
distance_arousal_first	distance_arousal_last	.0133	.0216	.0436

Table 4: Result for 2s interval between the beginning and the end of each one of the sequences

Discussion

The objective of this first exploratory study was to test the convergent validity of CSP measurement systems with psychophysiological emotional measures. First of all, results suggest that users self-evaluate their valence more accurately at the end of each of the sequences than at the beginning and more accurately their arousal at the beginning of each sequences. A second study will be conducted in the spring to further validate the results.

This paper confirms that continuous measurement allows for a richer self-perceived evaluation of emotion than traditional methods. Also, we show that there is a link between psychophysiological measurements such as facial expression analysis or electrodermal activity and the self-perceived evaluation of the participant. Regarding the primacy and recency effects our results confirm that they have the same influence, biasing the retrospection of participants [13], [10]. However, participants were more accurate at the beginning and at the end of their interaction, when reporting their emotion. Moreover, arousal has been proposed as the most influential factor regarding the self-evaluation of recency effect [28], [47]. With our experimental design, using continuous self-perceived measurement, we found that that users self-evaluate their valence more accurately at the end of each of the sequences than at the beginning, so our results do not converge with Cahill & McGaugh [6], proposing arousal as more important factor than valence for retrospection in this specific context. This implies that primacy and recency effect have an influence on the way participants reported their emotion. It is important for researchers to have this result in mind when using tools such as DARMA software to explore the self-perceived evaluation because it could lead to the overestimation or underestimation of the results of an experiment.

Our findings can also be useful for the marketing or entertainment industry. With many tools to evaluate subjective emotion, our paper allows a better understanding of all the instruments that could be available for practitioners to conduct (continuous) self-reported measurement. Our results could also be interesting to have in mind when practitioners design their product/service. Indeed, if the valence is best reported at the end of the interaction with the website, you should probably focus on minimizing negative emotion and maximizing positive emotion during this sequence of time, and all the more so knowing that negative information have stronger influence on memory [27]. Regarding the arousal result that is more accurate at the beginning of the session, it depends on your goal, maximizing or minimizing arousal in accordance with the area of activity. Overall, designers could influence the user's memory of an experience when focusing on the beginning and the end of the task. Finally, during user experience testing, designers should first of all pay attention to these two periods of time.

Our experiment faced different limitations. Starting with technical limitations, when participants did not touch the joystick, it would automatically go back by default to the center which represents neutrality in terms of emotions. It is possible that some participants were influenced by this default position.

Regarding the experimental design, this initial study was performed in a specific utilitarian context, future research could use a different context in order to validate the results in a broader range of contexts.

For the results, with our data and tools that we used, we cannot know if participant under or overestimated their emotion when reporting. During our second data collection for this research project, IAPS pictures were used in order to calibrate emotional reaction with extreme stimuli in order to gain knowledge on the under/overestimated part. At this moment, we are still analyzing the results from this second phase.

Conclusion

Emotions and the way we measure them are destined to endless debate [9]. As Cockburn & al [8] explained, an experience is internally composed of several sequences and is influenced by the most intense moment. Such an understanding accords more importance to psychophysiological measures and continuous self-perceived measures, as traditional post-task self-perceived measurements are more subject to bias [31]. In this paper, results suggest that an experience lived by a participant is not exactly the same as it is reported. Different biases may influence this misalignment of the "story". Primacy and recency through the influence of valence and arousal play an important role in the experience that is reported. Researchers in user experience have much to gain through a deeper understanding this topic.

Acknowledgments

We are thankful for the financial support of the Natural Sciences and Engineering Research Council of Canada.

References

1. Bartlett MS, Littlewort G, Lainscsek C, Fasel I, Movellan J: Machine learning methods for fully automatic recognition of facial expressions and facial actions. IEEE International Conference Systems, Man Cybernetics 592–597 (2004).
2. Boucsein W: Electrodermal Activity. Springer US (2012).
3. Bradley M, Lang PJ: Measuring Emotion: The Self-Assessment Semantic Differential Manikin and the Semantic Differential. Journal of Behavioral Therapy and Experimental Psychiatry, 25(1), 49-59 (1994).
4. Bradley MM, Greenwald MK, Petry MC, Lang PJ: Remembering pictures: Pleasure and arousal in memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(2), 379-390 (1992).
5. Burke A, Heuer F, Reisberg D: Remembering emotional events. Memory & Cognition 20(3), 277–290 (1992).
6. Cahill L, McGaugh JL: A Novel Demonstration of Enhance Memory Associated with Emotional Arousal. Consciousness & Cognition 410–421 (1995).
7. Cockburn A, Quinn P, Gutwin C: Examining the Peak-End Effects of Subjective Experience. Chi 1, 357–366 (2015).
8. Cockburn A, Quinn P, Gutwin C: The effects of interaction sequencing on user experience and preference. International Journal of Human-Computer Studies. 108, 89–104 (2017).
9. Colombetti G: Appraising valence. Journal of Consciousness Studies, 12(24), 103–126 (2005).
10. Conway MA: Cognitive Models of Memory. London (1997).
11. Courtemanche F, Léger PM, Dufresne A, Fredette M, Labonté-LeMoyne É, Sénécal S: Physiological heatmaps: a tool for visualizing users' emotional reactions. Multimed Tools Appl 1–28. (2017).
12. Cowie R, Douglas-Cowie E, Savvidou S, Mcmahon E, Sawey M, Schröder M "Feeltrace": An instrument for recording perceived emotion in real time. ISCA ITRW on Speech and Emotion 19–24 (2000).
13. Craik FIM, Lockhart RS: Levels of Processing : A Framework for Memory

Research. Journal of Verbal Learning and Verbal Behavior 11(6), 671–684 (1972).

14. Dirican AC, Göktürk M: Psychophysiological measures of human cognitive states applied in Human Computer Interaction. Procedia Computer Science 3:1361–1367 (2011).
15. Ekman P: Facial expression and emotion. American Psychologist, 48(4), 384-392. (1993).
16. Ekman P, Friesen W: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology 17(2)124–129 (1972).
17. Ekman P, Friesen W V: Unmasking the face. Cambridge, MA (2003).
18. Foster I, Kesselman C: The grid : blueprint for a new computing infrastructure, 2nd ed. Amsterdam (2004).
19. Fredette M, Labonté-LeMoyne É, Léger PM, Courtemanche F, Sénécal S: Research Directions for Methodological Improvement of the Statistical Analysis of Electroencephalography Data Collected in NeuroIS. Information Systems and Neuroscience 201–206 (2015).
20. Fredrickson BL, Kahneman D: Duration neglect in retrospective evaluations of affective episodes. Journal of Personality and Social Psychology, 65(1), 45-55 (1993).
21. Georges V, Courtemanche F, Sénécal S, Léger P-M, Nacke L, Pourchon R: The Adoption of Physiological Measures as an Evaluation Tool in UX. International Conference on HCI in Business, Government, and Organizations 90–98 (2017).
22. Girard JM: CARMA: Software for continuous affect rating and media annotation. Journal of open research software 2:1–11 (2017).
23. Girard JM, Aidan AG: DARMA: Software for dual axis rating and media annotation. Behavior research methods 1–8 (2017).
24. Hamann S: Cognitive and neural mechanisms of emotional memory. Trends in cognitive sciences 5(9) 394–400 (2001).
25. Hassenzahl M, Tractinsky N: User experience - A research agenda. Behaviour & information technology 25(2) 91–97 (2006).
26. Ortiz de Guinea A, Webster J: An Investigation of Information Systems Use Patterns: Technological Events as Triggers, the Effect of Time, and Consequences

for Performance. Mis Quarterly 37(4), 1165–1188 (2013).

27. Ito TA, Larsen JT, Smith NK, Cacioppo JT: Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. Journal of Personality and Social Psychology 75(4), 887–900 (1998).

28. Kahneman D, Fredrickson BL, Schreiber C a., Redelmeier D: When More Pain Is Preferred to Less: Adding a Better End. Psychological Science 4(6), 401-405 (1993).

29. Kandel ER, Schwartz JH, Jessell TM: Principles of Neural Science, 4th ed. McGraw-Hill, Health Professions Division, New York (2000).

30. Kensinger EA, Corkin S: Two routes to emotional memory: Distinct neural processes for valence and arousal. Proceeding of the National Academy of Sciences of the United States of America 101(9), 3310-3315 (2004).

31. Lallemand C, Gronier G: Méthodes de design UX: 30 méthodes fondamentales pour concevoir et évaluer les systèmes interactifs. Eyrolles (2015).

32. Lane RD, Chua PM-L, Dolan RD: Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. Neuropsychologia 37(9), 989–997 (1999).

33. Laurans G, Desmet PMA, Hekkert P: The emotion slider: A self-report device for the continuous measurement of emotion. In: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops pp 1–6 (2009).

34. Léger PM, Davis FD, Cronan TP, Perret J: Neurophysiological correlates of cognitive absorption in an enactive training context. Computers in Human Behavior 34:273–283 (2012).

35. Léger PM, Titah R, Sénecal S, Fredette M, Courtemanche F, Labonte-Lemoyne Él, De Guinea AO: Precision is in the eye of the beholder: Application of eye fixation-related potentials to information systems research. Journal of the Association for Information Systems, suppl. Special Issue on Methods, Tools, and Measurement 15(1), 651–678 (2014).

36. Mather M, Sutherland MR: Arousal-biased competition in perception and memory. Perspectives on Psychological Science 6(2), 114–133 (2011).

37. McGaugh JL: The Amygdala Modulates the Consolidation of Memories of

Emotionally Arousing Experiences. Annual Review of Neuroscience 27:1–28 (2004).

38. McGaugh JL, Cahill L, Roozendaal B: Involvement of the amygdala in memory storage: interaction with other brain systems. Proceeding of the National Academy of Sciences of the United States of America 93(24), 13508–13514 (1996).

39. Mello SKD, Craig SD, Gholson B, Franklin S, Picard R, Graesser AC: Integrating Affect Sensors in an Intelligent Tutoring System. Integrating affect sensors in an intelligent tutoring system. In: Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International conference on Intelligent User Interfaces. ACM Press, New York, pp. 7–13 (2005).

40. Murdoch BB, Lissner E, Marvin C: The serial position effect of free recall Journal of experimental psychology 64(5), 482-488 (1962).

41. Nagel F, Kopiez R, Grewe O, Altenmuller E: EMuJoy : Software for continuous measurement. Nagel F, Kopiez R, Grewe O, Altenmuller E: EMuJoy : Software for continuous measurement. Behavior Research Methods 39(2), 283-290 (2007).

42. Ortiz de Guinea A, Titah R, Leger P-M: Explicit and Implicit Antecedents of Users' Information Systems Behavioral Beliefs : A Neuropsychological Investigation. Journal of Management Information Systems 30(4), 179–210 (2014).

43. Ortiz De Guinea A, Titah R, Léger PM: Measure for Measure: A two study multi-trait multi-method investigation of construct validity in IS research. Computers in Human Behavior 29(3), 833–844 (2013).

44. Plutchik R: A general psychoevolutionary theory of emotion. In: Emotion: Theory, research, and experience, New York: Academic Press, pp 189–217 (1980).

45. Posner J, Russell JA, Peterson BS: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology 17(3), 715–734 (2008).

46. Ravaja N, Saari T, Turpeinen M, Laarni J, Salminen M, Kivikangas M: Spatial Presence and Emotions during Video Game Playing: Does It Matter with Whom You Play? Presence Teleoperators Virtual Environ 15(4), 381–392. (2006).

47. Redelmeier DA, Kahneman D: Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. Pain

66(1), 3–8 (1996).

48. Riedl R, Léger P-M: Fundamentals of NeuroIS. In: Springer (ed) Studies in Neuroscience, Psychology and Behavioral Economics, p.1-115, Springer, Berlin (2016).

49. Russel JA: A Circumplex Model of Affect. Journal of Personality and Social Psychology, 39(6), 1161-1178 (1980).

50. Sas C, Dix A, Hart J, Ronghui S: Emotional Experience on Facebook Site. CHI '09 Extended Abstracts on Human Factors in Computing Systems 4345–4350 (2009).

51. Shteingart H, Neiman T, Loewenstein Y: The role of first impression in operant learning. Journal of Experimental Psychology: General, 142(2), 476-488 (2013).

52. Uyl MJ Den, Kuilenburg H Van: The FaceReader : Online facial expression recognition. Proceedings of measuring behavior 589–590 (2005).

53. Vermeeren APOS, Law EL, Roto V: User Experience Evaluation Methods : Current State and Development Needs. Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries
pp. 521-530 (2010).

54. Zauberman G, Diehl K, Ariely D: Hedonic versus informational evaluations: Task dependent preferences for sequences of outcomes. Journal of Behavioral Decision Making 19(3), 191–211 (2006).

CHAPITRE 3: Article 2

Continuous self-perceived measurement systems: An investigation of the convergent validity of emotional retrospection[2]

Abstract

Context: New instruments to assess users' emotions such as CSPMS (Continuous Self-Perceived Measurement Systems) were developed by researchers to gain insight on the emotional retrospection process with high temporal resolution.

Objective: The primary objective of this research is to test the convergent validity of continuous self-perceived measurement systems to assess user emotions in user testing contexts.

Method: We collected participants' (n=24) emotional reactions while they were performing utilitarian online tasks using psychophysiological measures. Then, we asked participants to report their post-task emotional reactions with a continuous self-reporting system while watching a recording of their online session.

Results: Our results suggest that: 1) arousal emotional peaks (high) are better reported than valleys (low), 2) negative emotional valence is better reported than positive valence; 3) and that in general, data from the continuous self-perceived measurement system was not correlated with users' psychophysiological data.

Conclusion: These findings challenge the validity of continuous self-reported measurement systems in user testing contexts and allow UX professionals to better chose their tool to assess experience.

[2] Sébastien Lourties, Pierre-Majorique Léger, Sylvain Sénécal, and Marc Fredette

HEC Montréal 3000, Chemin de la Côte-Sainte-Catherine, Montréal, Québec, Canada, H3T 2A7

Highlights:

- Users report their arousal peaks with more accuracy than their arousal valleys.
- Users report their valence valleys with more accuracy than their valence peaks.
- There is no correlation between actual emotional peaks and valleys (arousal and valence) and those reported by users with continuous retrospective self-reported measurement tools.

Keywords: Continuous self-perceived measurement system; psychophysiological measurement; instrument validation; systematic bias; peaks; valleys

Introduction

Self-reported measurement tools assess the nature of the subjective experience and provide a global retrospective evaluation (Bradley & Lang, 1994; Mehrabian, 1998; Isbister et al, 2007; Scherer et al, 2013; Betella & Verschure, 2016). Many self-report tools consist of a scale that provides a score, which is quick and easy to interpret. This score, while representative of the user's experience, may not accurately represent the reality of users experience and behavior over time during interactions with the product and/or interface (Ortiz De Guinea, Titah, Léger, 2014) In fact, such methods do not capture the moment-to-moment positive and negative emotions experienced during a specific instance of the interaction (Cockburn et al, 2015; 2017). Finally, these methods are subject to biases such as the primacy, recency, and peak effects (Kahneman et al, 1993). These reasons have motivated researchers such as Gerard & Wright (2017) to explore a novel way to assess emotions with the continuous self-perceived measurement systems (CSPMS).

CSPMS are retrospective measurements but could also be used concurrently (presenting similar characteristics) with physiological measures, as they "let the observer track the emotional content of a stimulus as they perceive it over time" (Cowie et al, 2000, p.1). One such example includes the work of Gerard & Wright (2017), who've applied CSPMS measurements to track and examine the emotional dynamics of speech episodes. In other words, CSPMS allow users to retrospectively assess their emotions with a high temporal resolution and identify with precision the non-optimal moments of the experience. Thus, CSPMS appear to offer the same overall benefits as neurophysiological and physiological tools. In addition, CSPMS allow the recording of continuous data with high temporal precision, without interruption.

However, to our knowledge, no research has been conducted to empirically validate CSPMS by testing the convergent validity of the systems with users' actual emotional responses. Straub (1989) has highlighted the crucial importance of investigating the validity of scientific instruments in HCI and IS research, and to our knowledge, no empirical investigation has been conducted to validate CSPMS. Validation is an important component in any field of research, as it provides more credibility to scientific results and allows a validation of measures (MacKenzie et al, 2011). With valid tools, practitioners and entire industries may make strategic decisions with increased confidence.

Different methods are available to measure actual emotional responses (Scherer 2005). Psychophysiological tools can be used to measure emotional changes during an interaction. These measures are "an unobtrusive and implicit way to determine the user's affective or cognitive state on the basis of mind-body relations" (Dirican & Göktürk, 2011, p.1362). They allow researchers to capture unconscious or automatic emotions, for example, by using the micro-movements of the facial muscles (Ortiz de Guinea, Titah, & Leger, 2014). Psychophysiological tools also allow continuous measurement, permitting the understanding every moment of the interaction, and providing insightful user experience information. However, such tools require advanced knowledge, a specific environmental setting, and some are expensive (Riedl & Léger, 2016)

In this paper, we examined the validity of CSPMS in user experience testing. More specifically, we ask: Do systematic biases such as peak and peak-end effects have an influence on the retrospective evaluation? Thus, to assess CSPMS, we focus on emotional peaks that globally refer to the most intense moments of an experience (Kahneman et al, 1993).

To answer our research question, we conducted our experiment in a laboratory setting. Twenty-four participants performed a series of utilitarian tasks on an insurance company's website for 15 minutes. Continuous self-perceived measures were tested with two different interfaces and compared with different psychophysiological measures (facial micro-movements and skin electrical conductance). Our results are twofold. First, arousal peaks are better reported than arousal valleys, while negative valence is better self-reported than positive valence. Second, the assessment of emotions was less accurate when self-reported by participants than when measured with psychophysiological tools. With this paper, we contribute to the field of IS and HCI by investigating the validity of CSPMS, which were shown to be subject to biases during retrospection.

This paper is organized as follows. First, the peak and peak end effects are defined, and the self-perceived and psychophysiological continuous measurement systems are presented. Second, we explained the methodology, the experiment as well as the results of the two studies. Finally, we provided a discussion of the results and implication of our research for the HCI field.

Literature review and hypothesis development

Peak effect and peak-end rule

Within the literature on retrospection, two major concepts are highlighted. First, the “peak effect”, which refers to the most intense moment of an experience (Kahneman et al, 1993). Second, the “peak-end rule”, refers to the terminating and most intense moment (Kahneman et al, 1993). Together, these concepts explain why most individuals evaluate any experience based on "snapshots" (short moments) representing the most intense and final moments. Fredrickson and Kahneman (p. 46, 1993) summarize that “[...] most moments of an episode are assigned zero weight in the evaluation and a few select “snapshots” receive larger weights”. In other words, human emotion-related memory and retrospective evaluations of emotion in past experiences are selective.

In our daily lives, retrospective evaluations are commonly accepted. For example, an individual’s response to ‘how was your day?’, which requires a retrospective evaluation of the utility of past experiences, is deemed valid although the coherence between the response and the truly lived experience cannot be established. Kahneman et al. (1993) highlight two implicated concepts in these retrospective evaluations. First, the operation of memory, and second, the act of evaluation. These concepts could explain the lack of confidence for a provided retrospective answer.

First of all, the memory. Because of the declining ability to recall emotions over time, the contextual details surrounding an emotional event are forgotten (Eich & Schooler, 2000). In order to make a retrospective evaluation, participants use different methods to reconstruct their memories. They can use episodic information, which is specific to a particular event from the past, or semantic information, that consists of certain generalizations (i.e., beliefs). The retention interval, which is the time difference between when experience is lived and reported, moderates the influence of the peaks and peak-end rule (Kahneman et al., 1993). The peak-end rule becomes a stronger predictor of retrospective evaluation as the retention interval is minimized (Fredrickson & Kahneman, 1993; Redelmeier & Kahneman, 1996; Ariely, 1998; Schreiber & Kahneman, 2000). Recently, Gen et al. (2013) argued that the peak–end rule is a valid predictor for the way people retrospectively evaluate experiences only if it is done in a short retention interval. However, the total duration of the experiment has a minimal influence on the retrospective evaluation, which is called ''duration neglect'' (Kahneman et al., 1993, 1997). In sum, even if the experiment took place lasts weeks (e.g., a vacation), the peak effect represents a strong predictor of retrospective evaluations (Geng et al., 2013).

Second, the evaluation, which is almost completely based on the memory, shows that the moments at which emotions were the most extreme are good predictors of retrospective hedonic evaluations. For example, Fredrickson & Kahneman (1993) conducted an experiment where students evaluated video clips that represent pleasant and unpleasant emotions. The global evaluation of the pleasant and unpleasant clips confirmed Fredrickson & Kahneman's theory, because it was predicted by the average of the peak affect and the final experienced affect (R = .78 for pleasant and R = .69 for unpleasant). Also, in a utilitarian context, retrospective evaluations of payment sequences were influenced by the peak effect (Langer et al., 2005). Moreover, the retrospective evaluation has been demonstrated to influence the prospective choice of repeating or avoiding an experience (Kahneman 2000; Kahneman et al, 1993). Over time, different studies have reached the same conclusion (Redelmeier & Kahneman, 1996; Ariely, 1998; Schreiber & Kahneman, 2000; 2005; Do et al, 2008). Finally, in some studies, the peak and peak-end rule do not influence the retrospection due to the longer retention interval mentioned above (Kemp, Burt, & Furneaux, 2008; Seta et al, 2008; Miron-Shatz, 2009; Liersch & McKenzie, 2009). Thus, the human capacity to recall emotion fades over time, impacting the peak and peak-end effect (Walker et al, 1997). The more you wait to recall an emotion, the less the peak and peak-end effect will impact the retrospective evaluation.

Continuous self-perceived measures of user experience

Traditionally, to conduct a self-perceived evaluation, tools providing global evaluation of an experience are most often characterized by a score (e.g. 1988; Bradley & Lang, 1994; Mehrabian, 1998; Isbister et al, 2007; Scherer et al, 2013; Betella & Verschure, 2016). For example, Betella & Verschure, (2016) developed the Affective Slider (AS) which is a "digital self-reporting tool composed of two slider controls for the quick assessment of pleasure and arousal" (p.1).

In this way, continuous self-perceived measurement systems (CSPMS) have been proposed as a novel way to enable a user to dynamically report on their experience after the task (Cowie et al., 2000). CSPMS "let the observer track the emotional content of a stimulus as they perceive it over time, allowing the emotional dynamics of speech episodes to be examined" (Cowie et al, 2000, p.1). In other words, these measurement systems allow a retrospection of every moment of the interaction after the task. After the experiment, participants watch their own interaction with the task and continuously evaluate their performance. Thus, CSPMS not only improve the way retrospection could be done, CSPMS also compete with psychophysiological and body measurement tools to be compared with continuous retrospective data.

Some CSPMS were only composed of one dimension of the emotion (e.g. emotional valence). For example, CARMA software and the Emotion Slider have one dimension to facilitate the report of basic emotion (negative vs positive), so pushing up or down on a dial was the only thing participant had to do to report their emotion (Girard, 2017; Laurans et al, 2009). Later, other systems were proposed to measure emotion in two dimensions (e.g., emotional arousal and valence): Feel-Trace and EmuJoy (Cowie et al, 2000; Nagel et al, 2007). These were the first software packages to propose a CSPMS on two dimensions. DARMA, the most recent CSPMS, working with a joystick, was released last year (Girard & Wright, 2017). To our knowledge, CSPMS with two dimensions have only been used in hedonic contexts: FeelTrace, Emu-joy and DARMA have been tested in the music or commercial ad context with extreme (negative or positive) arousal and valence. Furthermore, authors (Girard & Wright, 2017) suggest that these CSPMS could be used to rate emotion in any context and used for retrospection. CSPMS allow new possibilities in the self-report field because they provide much more data over time. For example, DARMA has been used to study interpersonal behavior (Girard & Wright, 2017). Six undergraduate participants watched 74 romantic couples arguing, and then re-watched the interaction while providing continuous a rating of the level of communion and agency of the discussion.

Although, CSPMS provide richer information than self-reported scales due to their continuous nature, to our knowledge, no research has investigated CSPMS validity in retrospection. Indeed, in our daily life, our attention is divided on large quantities information so there's a "battle for a share of our limited attention and memory, with the brain selecting the winners and discarding the losers" (Mather & Sutherland, 2011, p.114). Necessarily, researchers suggest that self-report measures are hardly perfect because remembering could be a conscious or unconscious barrier to reality (Kahneman et al, 1993). For example, the emotional peak and the peak-end rule identified in the first section have been shown to influence the retrospective hedonic evaluation. Also, the primacy effect, which refers to the increased memory and over-weighted influence of the first moment of an experience (Murdoch et al, 1962; Shteingart et al, 2013; Zauberman et al, 2006), and recency effect, which correspond to the important influence of the last moment of an experience (Kahneman et al, 1993), were identified as obstacles for accurate emotional self-report. Memory is described as a U-shape curve with items of a series better reported at the beginning and at the end of an experience, with middle items being forgotten (Craik & Lockhart, 1972). Overall, as mentioned, there is a potential bias because of two factors: an operation of memory and an act of evaluation (Kahneman et al, 1993). So, efforts to validate these continuous measurement systems will inform researchers and practitioners on the potential benefits and limitations of using such systems to evaluate user experience.

Continuous psychophysiological measures of user experience

Because of individuals' difficulty to assess, discern or describe their emotions, many researchers have tried to develop instruments to understand users' emotions. Russell (1980), proposed a two-dimensional (2-D) model that allowed users to retrospectively rate their experience: the circumplex model of emotions. Currently, this model has been adopted by researchers in the fields of psychology, information technology and user experience (UX) (Watson et al., 1999 ; Thayer, 1989). The circumplex model of emotions (Russell, 1980; Posner, Russell & Peterson, 2005, Figure 3) was also developed to describe behavioral reactions to stimuli because "each emotion can be understood as a linear combination of these two dimensions, or as varying degrees of both valence and arousal" (p.1). In this way, this 2-D structure allows a user to simultaneously express emotional valence and arousal. Specifically, the valence and activation constructs have been used to measure psychophysiological reactions, and were defined as "an unobtrusive and implicit way to determine the user's affective or cognitive state on the basis of mind-body relations" (Dirican & Göktürk, 2011, p.1362). Today, psychophysiological measures, such as valence and arousal, enable the observation of decision making in UX research (Fredette et al, 2015). In this way, these measures have gained popularity in different fields, notably, information systems (Léger et al., 2012; Ortiz de Guinea et al., 2014).

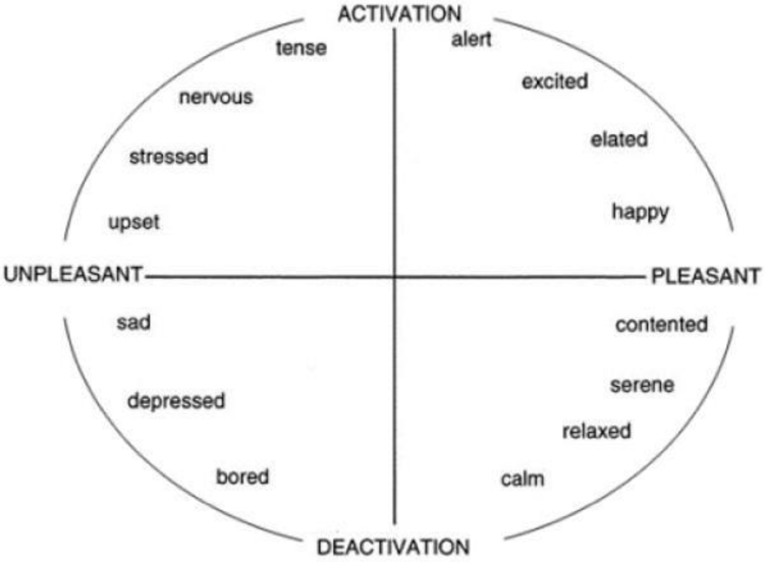


Figure 3 : The circumplex model of emotions with the horizontal axis representing the valence and the vertical axis representing the activation (Posner, Russell & Peterson, 2005).

Valence is “the ‘positive’ and ‘negative’ character of an emotion” and is sourced in behavior, affect, evaluation, faces, adaptive value, etc. (Colombetti et al, 2005, p.103). With the valence construct it is possible to contrast states of pleasure (e.g., happy) and displeasure (e.g., unhappy), which are the direction of the emotional response (Boucsein, 2012; Lane et al, 1999). For example, this construct has been identified as a determining factor to measuring the frustration of students in a learning context with an intelligent tutoring system (Mello et al, 2005). The results of this study show that detecting negative valence earlier could improve students’ learning curve. Also, in e-commerce, positive valence has been used to predict higher intention to purchase on the website (Sheng & Joginapelly, 2012).

Arousal is the reaction to an emotional stimulus characterized by a psychophysiological level of vigilance or a state of attention and represents the intensity of the emotion (Kandel et al, 2000). The arousal construct is used to contrast states of low arousal (e.g., calm) and high arousal (e.g., excitement) (Boucsein, 2012). In fact, arousal has been a largely used construct to study users’ e-commerce experience, capacity to engage with the interface or even their willingness to re-purchase (Menona & Kahn, 2002).

Psychophysiological measures, when studying emotions, present different advantages. First of all, capturing unconscious or automatic reactions of a participant without interrupting the experiment permits the recording of natural reactions to the stimuli and are non-obtrusive (Ortiz de Guinea et al, 2014). Further, this method allows the recording of continuous data on the user reaction in real time. The instant capture of data is an important advantage to avoid the common bias of memory (Ortiz de Guinea & Webster, 2013). In this article, valence and arousal permit to measure the psychophysiological states of the user and also allowed us to compare these two measures with DARMA software.

Hypothesis development

The peak and peak-end rule (Kahneman et al, 1993) refer to the most emotionally intense (positive or negative) moments of an experiment. Graduation is an intense, unforgettable moment that Pillemer (2000) studied. Graduating students remember the moment they went on stage to receive their diploma for many reasons (pride, elevation, insight, connection) that were associated with an arousing moment. This means that their electrodermal activity and heart beats were rising. When performing the cold-water test, another intense experience (Kahneman et al, 1993), participants who had to place their hands in cold water (14 degrees Celcius) remembered the most intense, or most arousing, moment of the experiment. Fredrickson & Kahneman, (1993) ran an experiment combining both aversive and pleasant video clips watched by participant. The results showed that the global retrospective evaluation was predicted by the most intense moment of each of the video clips, whatever the experiment duration. Based on this literature (Fredrickson & Kahneman, 1993; Kahneman et al, 1993; Pillemer, 2000), we propose that users will better report their arousal high points (peaks) than their low points (valleys) when using a retrospective self-report system. Finally, we expect that arousal peaks collected with psychophysiological measures will be more positively correlated to self-reported arousal peaks measured with a continuous retrospective self-report system than arousal valleys.

H_1: With a concurrent self-report system, users report with more accuracy their arousal peaks than their arousal valleys

Some of the first researchers to investigate the difference in memory accuracy of negative and positive emotions were Thomas & Diener (1990). They demonstrated that negative emotions were better recalled than positive emotions when participants were asked to report their daily mood. Thus, participants were underestimating the frequency of positive moments, suggesting that negative moments were well reported. Moreover, when recalling past events, negative emotions seem to be easier to report. Finkenauer & Rime, (1998) asked participants to recall emotional events that they shared with other or that they keep secret, welcoming both positive or negative events. Overall, there were much more negative events reported than positive ones. Negative emotions remain more important in participants' mind, which confirms that "Bad Is Stronger Than Good" (Baumeister et al, 2001). Baumeister et al. (2001) showed that there is no exception in our daily life: negative emotions outweighs positive emotions in a recall and evaluation context. Thus, they suggest that emotions, learning processes, mental processes and many more aspects of life were always more influenced by negative emotions than by positive emotions. Based on prior research (Baumeister et al, 2001; Finkenauer & Rime, 1998; Thomas & Diener, 1990), we propose that users will better report their valence low points (valleys) than their high points (peaks) when using a retrospective self-report system. In this way, due to the strong evidence of the power of negative emotion, we expect that valence valleys collected with psychophysiological measures will be more positively correlated to self-reported valence valleys measured with a continuous retrospective self-report system than valence peaks. Thus, we posit:

H_2: With a concurrent self-report system, users report with more accuracy their valence valleys than their valence peaks.

Both CSPMS and psychophysiological measures were used at the same time in music research investigating the capacity of music to generally induce emotions (Grewe et al, 2007; 2009). Participants were rating their emotions regarding the music with a CSPMS while psychophysiological data was recorded at the same time. Even if the purpose of this experiment was not to test the validity of CSPMS, researchers found low correlation between CSPMS and psychophysiological data (arousal in this case) when the two methods were concurrent. To our knowledge, no research has been conducted to validate the use of CSPMS with psychophysiological measures using retrospection.

As CSPMS have been built to measure many constructs (e.g., valence, arousal, cognitive load, engagement), CSPMS were developed to measure emotional state in the same way as psychophysiological measurements systems do. CSPMS have been also developed with a lot of common characteristics being a complement or an alternative to continuous psychophysiological measure that could be expensive. Thereby, CSPMS can record retrospective emotional state in real time with a relatively high temporal precision (4 times per second for example). Moreover, the retrospective evaluation is affected by the capacity to recall details (Eich & Schooler, 2000), which is no longer an issue with CSPSM that allow to re-watch the exact same experience that the participant just lived. Also, CSPSM system permit to record emotional state without interruption and is almost non-obtrusive because the user is kept in an immersive state while rating at the same time. So, CSPMS is far away from classics self-perceived measure that only were a global evaluation of the experiment and seem to be close to the characteristics of psychophysiological measures (high temporal resolution instrument). Moreover, in a context of retrospection, memory would impact the way participants report their emotions. As researcher found, experiencing emotional variation will have important impact on your retrospection (Fredrickson & Kahneman, 1993) and make other "neutral" emotion insignificant. In this way, the emotional peaks that participants lived during the experiment should be well report with the CSPMS as it provides the exact experience lived, without reconstructions (Christianson & Safer, 1996; Levine, 1997) and emotional fade (Walker et al, 1997). In this way, we hypothesize that user will be accurate at reporting their emotional peaks and valleys (arousal and valence) when using a retrospective self-report system. Thus, we expect that peak and valleys (valence and arousal) recorded with psychophysiological measure will be positively correlated with peak and valleys (valence and arousal) measure with CSPMS.

H$_3$: There is a positive relationship between actual emotional peaks and valleys (arousal and valence) and those reported by users with continuous retrospective self-reported measurement tools.

Study 1

To answer our research question, we conducted two studies. The first one is a task of claim in an insurance context (n=13) and the second one in a purchase of insurance (n=11).

In the first study, the objective was to test our hypotheses to provide an initial overview of CSPMS performance and the influence of systematic bias. The first laboratory study was conducted with 13 participants (6 females). They were between the age of 21 and 48 with a mean of 32. Participant were pre-screened for vision issues such as wearing glasses, laser eye surgery, epilepsy, astigmatism, and psychiatric or neurological diagnoses that could influence measurement. The research was approved by the Ethics Committee of our institution and each participant received a compensation of 100$. Two pretests were performed before starting the data collection to be sure that our instruments and the protocol were aligned with the objective and minor adjustment was made. Participants had to perform a series of utilitarian tasks on an insurance company website for fifteen minutes and they proceed to the retrospective evaluation.

Procedure

After signing a consent form we placed sensors in the palm of the participant hands to measure the arousal and make sure all the instruments were recording correctly. Then, participants had to simulate an insurance claim on a given website. In order to enhance the real conditions of a claim, we recruited participants that had been both involved with a small accident over the past two years and were in possession of a valid driver's license. We displayed a video of an accident before the experiment begun to set all of the participant with the same information to report on the claim[3]. Then they completed the declaration with their personal information and with the one we provided (e.g. the video, the place of accident, the vehicle types etc.). When the claim was done, participants filed the Webqual questionnaire (Eleanor et al, 2007) to assess their perception of the website (planned to compare the interface of Study 2). Directly after that, they had to view and rate their interaction with the website with the DARMA software and a joystick to avoid effect of reconstruction (Christianson & Safer, 1996; Levine, 1997). In Figure 4, you will find the overall experimental protocol of this study. Moreover, DARMA software permit to show the exact interaction that the participant just lived, reducing the fact that emotions could fade over time (Walker et al, 1997). In this way, they express continuously their emotional valence and arousal moving the joystick in the 2-dimensional map. We provided the same instruction to the participant regarding the use of DARMA and the joystick. A three-minute simulation was conducted with the participant to make sure they understood the mechanism

[3] Link to the car accident video: https://www.youtube.com/watch?v=wzGe-ZCvTrM

and the different location of the dial corresponding to their emotions. The experiment was composed of 14 short sequences (log in on the website was a sequence, book an appointment was an another one, etc.) for a total of 182 sequences for the overall sample (13 participants x 14 sequences).

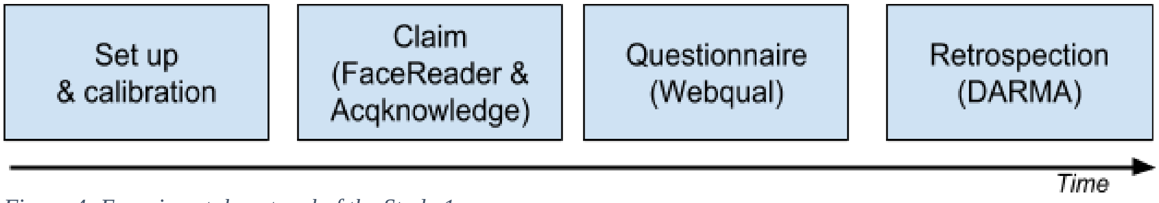


Figure 4: Experimental protocol of the Study 1.

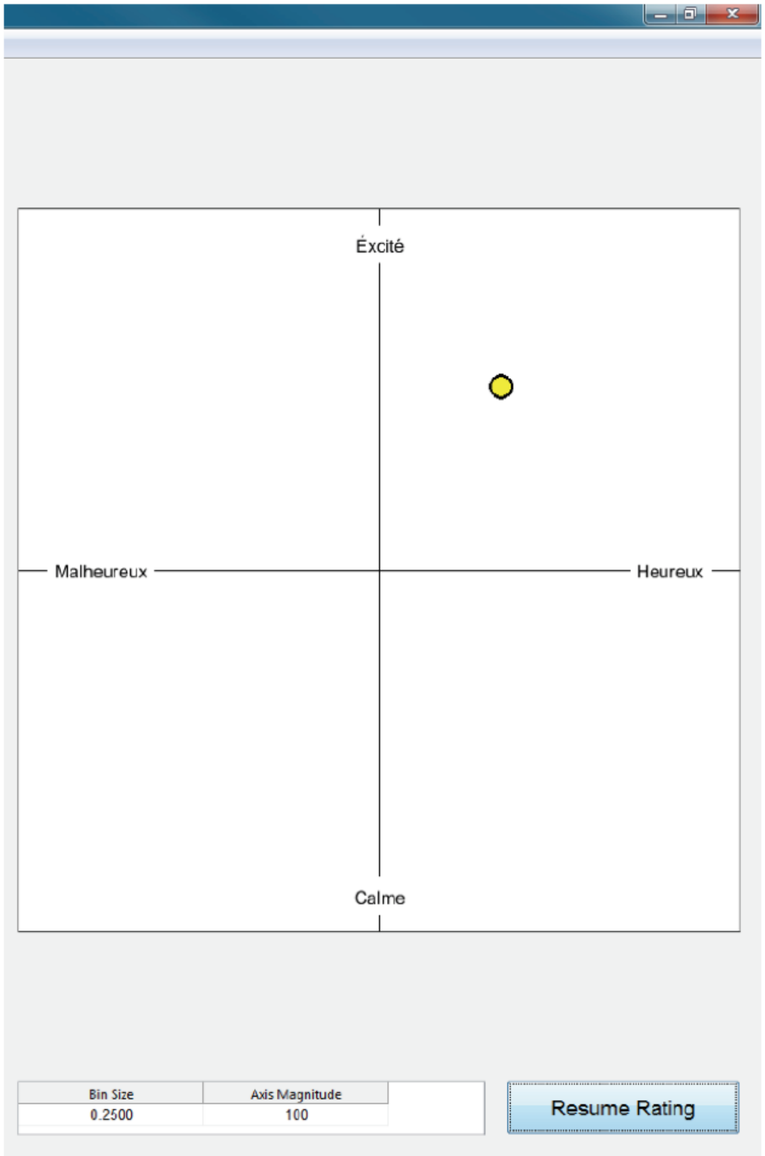
Measures and instruments

In our research, we focused on what Kahneman (2000) coined the moment-utility which is the "sign and intensity of affective/hedonic experience at a given moment in time" and may be inferred from self-reports or psychophysiological measures. Thus, it represented the peak of emotion lived during an experiment.

In order to measure arousal, we used electrical properties of the skin, in response to the secretion of sweat from the palm of the hands (near eccrine glands). Thus, electrodermal activity (EDA) has been proposed as an accurate way to collect arousal data (Hassenzahl & Tractinsky, 2006; Boucsein, 2012; Dawson et al, 2007). In Study 1, we used the electrodermal activity that Acknowledge software could collect to MP150 sampled at 500 Hz (BIOPAC, Goleta, USA).

For the valence, we used facial expression and the micro-movements of the facial muscles because they were instantaneous and unconscious (Ekman, 1993). Indeed, we used FaceReader software (Noldus, Wageningen, Netherlands), that calculated the value of the positive emotion minus the strongest of the negative emotion giving a valence score between -1 to +1 (Ekman & Friesen, 2003) with a temporal precision of 30 inferences per second.

To record the interaction with the website which serve to analyze the valence, we used Media Recorder (Noldus, Wageningen, Netherlands) at a frequency of 30 frames per second and a resolution of 800x600. Then to synchronize all the signals between Acknowledge, FaceReader and Media Recorder we used Observer XT (Noldus, Wageningen, Pays-Bas) as recommended by Léger et al (2014). Finally, to determine the retrospective continuous self-reported emotional reaction of valence and arousal, we used the DARMA software (Girard and Wright, 2017) which “synchronizes media playback and the continuous recording of two-dimensional measurements through the manipulation of a computer joystick to indicate changes in their emotional state”, (p.1). We have chosen DARMA because, it was the latest version of CSPMS built by researcher, open-source and potentially customizable (see Figure 5). Moreover, the output was easy to work with, an Excel file providing the continuous time and the score of arousal and valence between -100 and 100 that the participant rated with the joystick. All these measures to compute emotional state were summarized in Table 5. We have chosen the same joystick as Girard and Wright, (2017) recommended in their paper to let participant reported their emotions: a Logitech Extreme 3D Pro.



Bin Size	Axis Magnitude
0.2500	100

Figure 5: Screenshot of the two-dimension self-report window during media playback (Girard and Wright, 2017)

To compute the psychophysiological valence and arousal, we used Cube software (Léger et al 2018). Also, we compute the psychophysiological data at 25Hz in order to compare them with CSPMS data also sample at 25Hz.

	Psychophysiological (concurrent)	Self-reported (retrospective)

Valence	Instantaneous emotional valence at time t (FaceReader, Noldus, Wageningen, Netherlands)	Using the joystick position on the valence axis at time t (Darma, Girard and Wright, 2017)
Arousal	Electrodermal activity at time t (Acqknowledge, BIOPAC, Goleta, USA)	Using the joystick position on the arousal axis at time t (Darma, Girard and Wright, 2017)

Table 5: Measures used to determine emotional changes during the experiment and during the retrospection

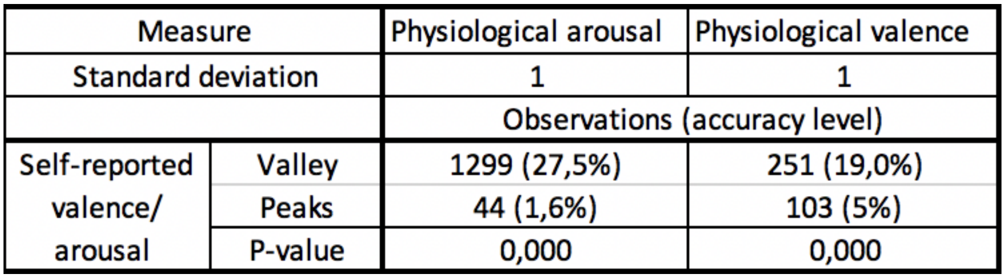
Results

For the overall experiment, we had a total of 12472 data point (observations) at 30 hertz for the emotional valence and 24335 data point at 500 hertz for the arousal. These numbers respectively represent the total of observation recorded by the instruments (FaceReader and Acknowledge), for all the participants and for each construct. The quantitative difference between these two measures is due to the capacity of the instrument to record data from the participant. For example, if the participant turned their face, FaceReader (Noldus, Wageningen, Netherlands) could not record the facial micro movement. In our case, the participant had to check documentation several times such as insurance number, driving license, and vehicle information, in order to enter the information on the insurance website. A Z-score standardization was applied to the reported and psychophysiological values, using the general mean of the sample. In order to identify peaks, valleys, and neutral emotional responses, we segmented the data into three groups. We made a segmentation of our variable in two different ways using a standard deviation (SD) of +/-1 or +/-2 in order to assess the capacity to retrospectively report intense emotional moments. Data between a standard deviation (SD) of -1 and +1 or -2 and +2 was considered as neutral. We defined three (3) different emotional areas such as: valleys, neutral and peaks. With this segmentation, we constructed cross tables to observe the distribution between the reported and psychophysiological emotions for every variable, every tool and every SD, giving us a complete view of the emotional variation (Table 7).

In order answer our hypothesis, we used an adjusted chi-square test because we only consider results that were peaks or valleys (avoiding neutral). Also, observed proportion were not equal with reported proportion, we had to readjust the test with one variable.

Arousal and Valence (H1 and H2)

With (H1), we hypothesize that users will be better at reporting their arousal high points (peaks) than their low points (valleys) when using a retrospective self-report system whereas with (H2), we proposed that users will be better at reporting their valence low points (valleys) than their high points (peaks). In this analysis, the objective was to determine by construct (arousal and valence), which of the valleys or the peaks were better reported. To do so, we use a dummy variable (wrong input = 0 and right input = 1) with a cross table for every construct and standard deviation. Then, we conducted an adjusted chi-square test of independence to provide p-value to our results (Table 6).



Measure		Physiological arousal	Physiological valence
Standard deviation		1	1
		Observations (accuracy level)	
Self-reported valence/ arousal	Valley	1299 (27,5%)	251 (19,0%)
	Peaks	44 (1,6%)	103 (5%)
	P-value	0,000	0,000

Table 6: Comparison of accuracy level between valleys and peaks for arousal and valence construct of Study 1[4]

For the arousal construct, valleys were better reported than peaks with a SD of +/- 1 (27.5% vas 1.6%, p < 0.000). **Thus, H1 is not supported.**

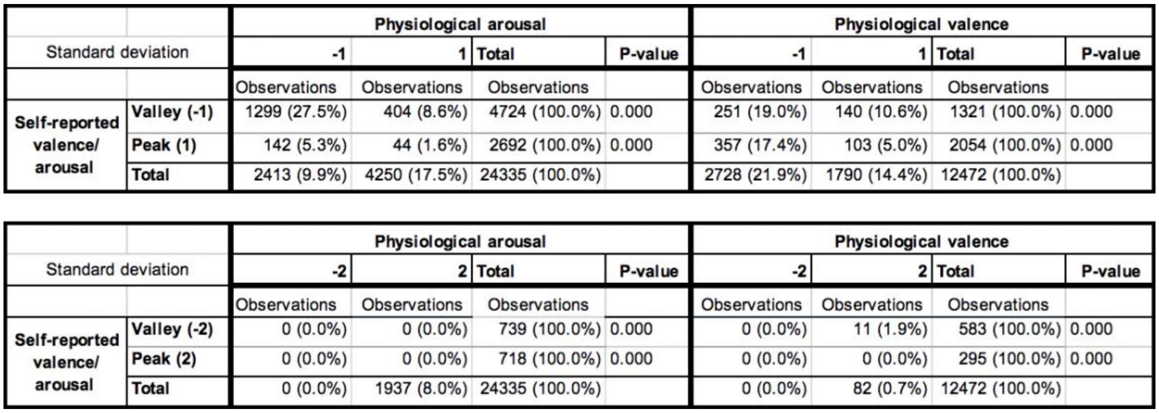
For the valence construct, with a SD of +/- 1, valleys were better reported than peaks with an accuracy level of 19% against 5% for peaks (p < 0.000). **Thus, H2 is supported.**

[4] No comparison could be done for the arousal construct with a SD of +/- 2 because any participant experiments such an emotion, so no data was available to make a comparison between valley or peaks.
No comparison could be done for the valence construct with a SD of +/- 2, because any participant experiments such an emotion, so no data was available to make a comparison between valley or peaks.

Continuous Self-Report Accuracy (H3)

With (H3), we expected that users will be accurate at reporting their emotional peaks and valleys (arousal and valence) when using a retrospective self-report system. Regarding arousal results consolidated in the Table 7, with a SD of +/- 1 there were 4724 reported value of valleys. So, participants were reviewing their interaction with the website and they self-reported 4724 time in the valleys zone. However, they were accurate only 1299 times with what they actually lived (based on psychophysiological arousal). In other words, there were only 1299 data points that were in common between the reported and the real experiment (based on psychophysiological arousal). The accuracy level was the highest of the experiment for the valleys (27.5%, $p < 0.001$). Peaks were not correctly reported with an accuracy level of only 1.6% ($p < 0.001$). With a SD of +/-2 participant totally missed their emotional report. With an accuracy level of 0% for the peaks (718 data points) and valleys (739 data points), they could not well self-report their most intense moments ($p < 0.001$). Notice that there was no psychophysiological data point recorded for valleys, but participant still reported 739 times in this zone.

Overall, for this study, arousal peaks were underestimated, the total of reported peaks, however the SD were always lower ($p < 0.001$) than the data point measured in real time during the experiment (2692 vs 4250 and 718 vs 1937). Regarding valleys, they were always overestimated no matter the SD (4724 vs 2413 and 739 vs 0, $p < 0.001$).



Standard deviation		Physiological arousal				Physiological valence			
		-1	1	Total	P-value	-1	1	Total	P-value
		Observations	Observations	Observations		Observations	Observations	Observations	
Self-reported valence/ arousal	Valley (-1)	1299 (27.5%)	404 (8.6%)	4724 (100.0%)	0.000	251 (19.0%)	140 (10.6%)	1321 (100.0%)	0.000
	Peak (1)	142 (5.3%)	44 (1.6%)	2692 (100.0%)	0.000	357 (17.4%)	103 (5.0%)	2054 (100.0%)	0.000
	Total	2413 (9.9%)	4250 (17.5%)	24335 (100.0%)		2728 (21.9%)	1790 (14.4%)	12472 (100.0%)	

Standard deviation		Physiological arousal				Physiological valence			
		-2	2	Total	P-value	-2	2	Total	P-value
		Observations	Observations	Observations		Observations	Observations	Observations	
Self-reported valence/ arousal	Valley (-2)	0 (0.0%)	0 (0.0%)	739 (100.0%)	0.000	0 (0.0%)	11 (1.9%)	583 (100.0%)	0.000
	Peak (2)	0 (0.0%)	0 (0.0%)	718 (100.0%)	0.000	0 (0.0%)	0 (0.0%)	295 (100.0%)	0.000
	Total	0 (0.0%)	1937 (8.0%)	24335 (100.0%)		0 (0.0%)	82 (0.7%)	12472 (100.0%)	

Table 7: Accuracy results between continuous self-reported and continuous psychophysiological measure of arousal and valence constructs for study 1[5]

Regarding the valence, with a SD of +/- 1 the valleys was better reported than the peaks ($p < 0.001$). It means that participants were more accurate in reporting negative emotion than positive. With a level of accuracy of 19% for the valleys and 5% for the peaks we concluded that participants were not able to report correctly their emotions. Also, with a SD of +/-1, valleys reported were underestimated (1321 data points vs 2728 truly experimented) whereas peaks were overestimated (2054 data points vs 1790 truly experimented), $p < 0.001$).

With a SD of +/-2, poor level of accuracy was found. Even considering the 5% most extreme emotions, participants were not able to correctly report these moments. For the valleys there was no psychophysiological data point recorded but participants still reported 583 times in this zone. There were 295 observations for the peaks with 0 correctly reported. The conclusion is that participants totally missed their emotional report. Surprisingly, in this case, valleys were overestimated because 583 data points were reported whereas no participant had lived such emotion during the experiment. Peaks followed the same trend with an overestimation of the reported valence (295 data points vs 82 truly experimented).

All the results were found with a p-value at 0.000. In our case, the p-value indicated that the cross results between the reported arousal/valence of valleys/peaks and the

[5] The negative SD (-1/-2) is associated to valley and the positive SD (+1/+2) to peaks.

arousal/valence by FaceReader/Acknowledge of valleys/peaks were different from the other results. **Thus, H3 is not supported.**

Discussion

Overall, we found that arousal and valence emotional state were not accurately reported with CSPMS (H3). It was easier for participant to report peaks of emotional arousal (H1) and valleys of emotional valence (H2). The results were not in line with Girard & Wright's (2017) findings of CSPSM being a valid tool in a retrospective user testing situation. The results of the present study even show contradictory expectation of emotions report.

However, different factors could explain the fact that participants could not accurately report their emotions. First of all, the joystick itself may have restrained the participants' capacity to emotionally express themselves. Proper manipulation of the joystick requires maintained hand pressure, without which it automatically goes back to the middle, which represented emotional neutrality (inherent to their tool). We did not have a baseline to control user's psychophysiological level and psychophysiological sensitivity differences. This is a limitation to our experimental design because all participants do not have the same psychophysiological characteristics such as intensity in facial expression or electrodermal activation. These differences in the human body could affect our results; in other words, one participant's psychophysiological characteristics could skew findings. Based on these findings, we performed a second study in the same domain and with a baseline.

Study 2

The experiment was the same as Study 1 and approved by our Ethics Committee. We conducted our experiment over a period of one week for a total of thirteen participants (two rejected due to artefact malfunction). So, eleven valid participants (7 females) were computed between the age of 22 and 62 with a mean at 34. Participant were pre-screen with the same restriction and compensation (100$). The objective of this second study was to further examine the effect of the peak and peak-end effect identified in the retrospective evaluation of the first study and the effect of the baseline. Thus, we tested the same hypothesis (**H1**, **H2**, and **H3**) to obtain more robustness in our findings.

Procedure

First, participants were exposed to twelve (12) images from the IAPS database. There were three (3) images from each extreme emotions (low valence/high arousal; low valence/low arousal; high valence/high arousal; high valence/low arousal). Pictures appear randomly in order to avoid emotional contamination from the other stimuli.

As recommended by Lang et al (2008), pictures were presented one by one during a period of 5 seconds and displayed using Qualtrics to participants.

Subsequently, during 10 to 15 minutes, participant completed utilitarian tasks on the insurance website, such as logging in to the client homepage, providing information about their driving license and choosing the best offer that suits their insurance needs. Finally, they viewed their interaction with the website and self-reported, using the joystick, their emotional arousal and valence. During the retrospection, participants were provided with the same instruction in both Study 1 and Study 2.

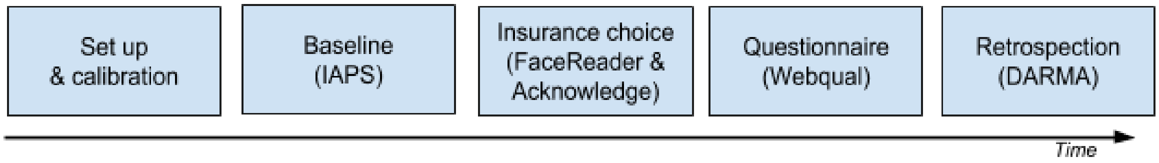


Figure 6: Experimental protocol of the Study 2.

Measures and instruments

As with Study 1, valence and arousal emotional state were measured with the same psychophysiological and self-perceived instrument. Psychophysiological measure was recorded during the series of utilitarian task and self-perceived measure right after that with DARMA (Girard & Wright, 2017).

The only modification to the experimental protocol was the baseline introduced at the beginning of the experiment (see Figure 6). As an emotional reaction to a stimulus did not represent the same direction and intensity for every participant, the baseline permit to consider user’s psychophysiological level (μ) and psychophysiological sensitivity (σ) differences. In other words, due to differences in the human body only one participant could influence with his psychophysiological characteristics, all the findings. In this way, we preserved individual’s psychophysiological characteristics to calculate the emotional interpersonal variation that was missing in Study 1 (Picard et al, 2001). Thus, we had to correct this variation with the subject baseline (Van den Broek et al, 2010). To do so, we used the following z-score equation:

$$\mathrm{W'i} = \frac{W_i - \mu}{\sigma}$$

The z-score equation was chosen because it makes a difference between user differences for their baseline psychophysiological levels and sensitivity. To create this baseline, we selected standardized pictures of the International Affective Picture System (IAPS) (Lang et al 2008) using the psychophysiological data of the average 5 seconds shown to the participants.

The IAPS pictures (Lang et al 2008) used to create the stimuli were: 1670, 2501, 5870, 8185, 8190, 8370, 8485, 9209, 9291, 9330, 9620, 9622. For the emotional valence and arousal, the maximum value could go to 3 and the minimum to -3. In the appendix B, you can see the emotional valence and arousal associated with each image. Literature shows that some of the pictures from the database were outdated so we conducted a pre-screening inter-judge agreement with five researcher assistants to avoid emotional anachronisms and misinterpretation. We tried to choose picture that had similar valence and arousal for each extreme emotions in order to be more confident about the emotional profile of the participants. In the figure 7, you can see the repartition of the picture we choose to set our baseline.

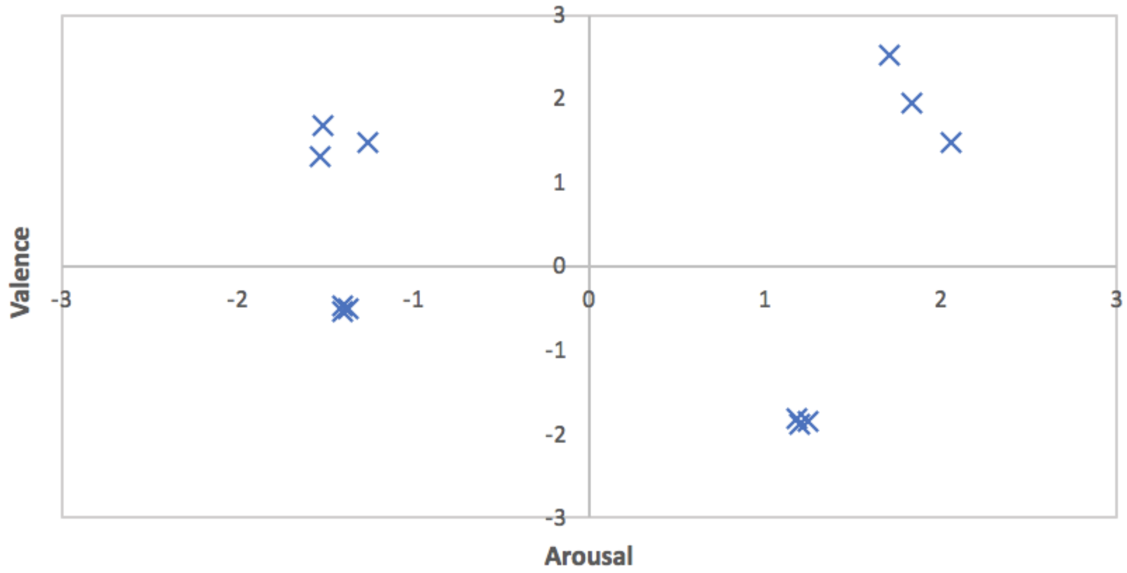


Figure 7: Repartition of the IAPS pictures chosen to create the baseline of the 2nd study.

Results

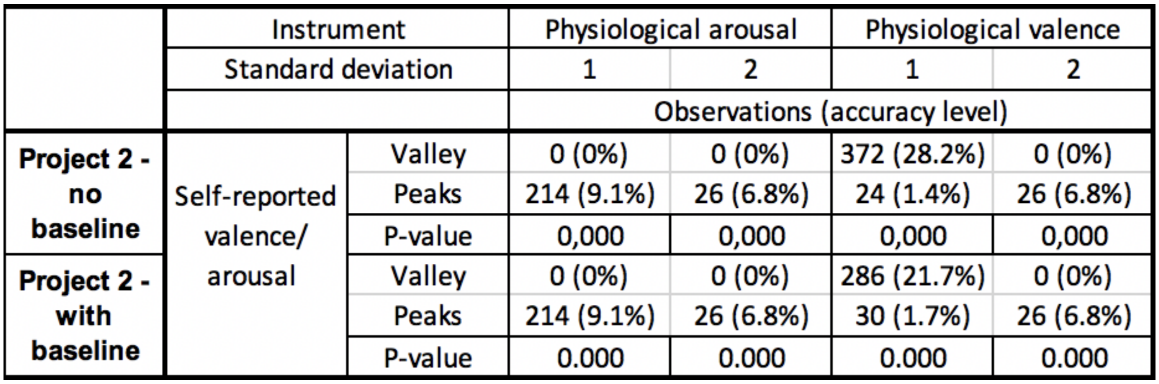
We used the same methodology analysis as Study 1 to compute our results of the Study 2. Participants perform a series of 16 to 19 sequences (every new webpage such as the logging, the page for driving information etc.) for a total of 187 sequences for the overall sample. Sixty percent of sequences lasted less than 30 seconds. However, even if Study 1 and 2 were conducted in a utilitarian context in order to generalize our findings, we used two different websites. We conducted a Webqual analysis that suggest two important differences between the interface. The Webqual results indicate significative difference for 2 dimensions: Informational Fit-to-Task and Tailored Information both at p < .05. In fact, the dimension Informational Fit-to-Task was significantly better perceived in the first study (mean of 6,41 for Study 1 vs 5,64 for Study 2) because it was representing an urgent need (car accident). Tailored Information was also significantly better perceived in Study 1 (5,87 vs 5,33) because the website was responsive to the participant's information (e.g. location of the accident, which side of the car is affected etc.). In the appendix A, you may find the complete results of every dimension.

Arousal and Valence (H1 and H2)

With (H1), we proposed that users will be better at reporting their arousal high points (peaks) than their low points (valleys) when using a retrospective self-report system whereas with (H2), we proposed that users will be better at reporting their valence low points (valleys) than their high points (peaks). In this analysis, the objective was to confirm the findings of the study 1, to know which of the valleys or the peaks was better reported for the two constructs. To do so, we made a create a binary variable (wrong and good) with a cross table for every construct and standard deviation. We only reported the good percentages to simplify the comprehension. Then, we conducted an adjusted chi-square test of independence to provide p-value to our results (Table 8).

For the arousal construct, peaks were always better reported than valleys no matter the SD. The accuracy level was 9.1% for the peaks versus 0% for valleys at +/- 1 SD and 6.8% vs 0% respectively for +/- 2 SD. Moreover, these results were converging to the same conclusion with and without baseline (p < 0 001). **Thus, H1 is supported.**

For the valence construct, with a SD of +/- 1, valleys are better reported than peaks with an accuracy level of 28.2% against 1.4% for peaks (p < 0.001). Same results were found for the analysis without baseline (21.7% for valleys vs 1.7% for peaks). With a SD of +/- 2 surprisingly, participant better reported peaks than valleys, in the analysis with and without baseline. However, the proportion is not the same (6.8% of accuracy level for both analysis) which still make negative valence as an important factor of reported predictor.
Thus, H2 is supported.



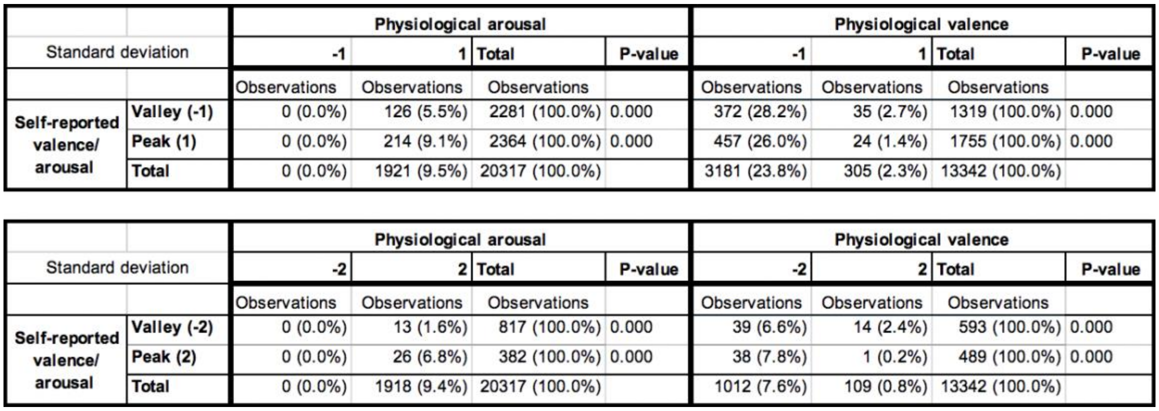
	Instrument		Physiological arousal		Physiological valence	
	Standard deviation		1	2	1	2
			Observations (accuracy level)			
Project 2 - no baseline	Self-reported valence/ arousal	Valley	0 (0%)	0 (0%)	372 (28.2%)	0 (0%)
		Peaks	214 (9.1%)	26 (6.8%)	24 (1.4%)	26 (6.8%)
		P-value	0,000	0,000	0,000	0,000
Project 2 - with baseline		Valley	0 (0%)	0 (0%)	286 (21.7%)	0 (0%)
		Peaks	214 (9.1%)	26 (6.8%)	30 (1.7%)	26 (6.8%)
		P-value	0.000	0.000	0.000	0.000

Table 8: Comparison of accuracy level between valleys and peaks for every construct of study 2.

Continuous Self-Report Accuracy (H3)

With (H3), we expected that user will be accurate at reporting their emotional peaks and valleys (arousal and valence) when using a retrospective self-report system. Regarding arousal results consolidated in the Table 9, with a SD of +/- 1 there were 2281 reported value of valleys. So, participants were reviewing their interaction with the website and they self-reported 2281 time in the valleys zone. However, no psychophysiological data point was actually recorded. In other words, participant never lived such emotion but still reported in the valleys zone, leading to an accuracy level of 0% ($p < 0.001$). Peaks had an accuracy level of 9.1%, which is low. With a SD of +/-2 participant did not perform their emotional report. With a level of accuracy of 6.8% for the peaks (382 data points) and 0% for the valleys (817 data points), they couldn't well recall there most intense moments. Notice that there was no psychophysiological data point recorded for valleys, but participant still reported 827 times in this zone.

For this study, arousal peaks were overestimated with a SD of + 1 (2364 vs 1921) and underestimated with a SD of + 2 (382 vs 1918). Regarding valleys, they were always overestimated no matter the SD (2281 vs 0 and 817 vs 0).



		Physiological arousal				Physiological valence			
Standard deviation		-1	1	Total	P-value	-1	1	Total	P-value
		Observations	Observations	Observations		Observations	Observations	Observations	
Self-reported valence/ arousal	Valley (-1)	0 (0.0%)	126 (5.5%)	2281 (100.0%)	0.000	372 (28.2%)	35 (2.7%)	1319 (100.0%)	0.000
	Peak (1)	0 (0.0%)	214 (9.1%)	2364 (100.0%)	0.000	457 (26.0%)	24 (1.4%)	1755 (100.0%)	0.000
	Total	0 (0.0%)	1921 (9.5%)	20317 (100.0%)		3181 (23.8%)	305 (2.3%)	13342 (100.0%)	

		Physiological arousal				Physiological valence			
Standard deviation		-2	2	Total	P-value	-2	2	Total	P-value
		Observations	Observations	Observations		Observations	Observations	Observations	
Self-reported valence/ arousal	Valley (-2)	0 (0.0%)	13 (1.6%)	817 (100.0%)	0.000	39 (6.6%)	14 (2.4%)	593 (100.0%)	0.000
	Peak (2)	0 (0.0%)	26 (6.8%)	382 (100.0%)	0.000	38 (7.8%)	1 (0.2%)	489 (100.0%)	0.000
	Total	0 (0.0%)	1918 (9.4%)	20317 (100.0%)		1012 (7.6%)	109 (0.8%)	13342 (100.0%)	

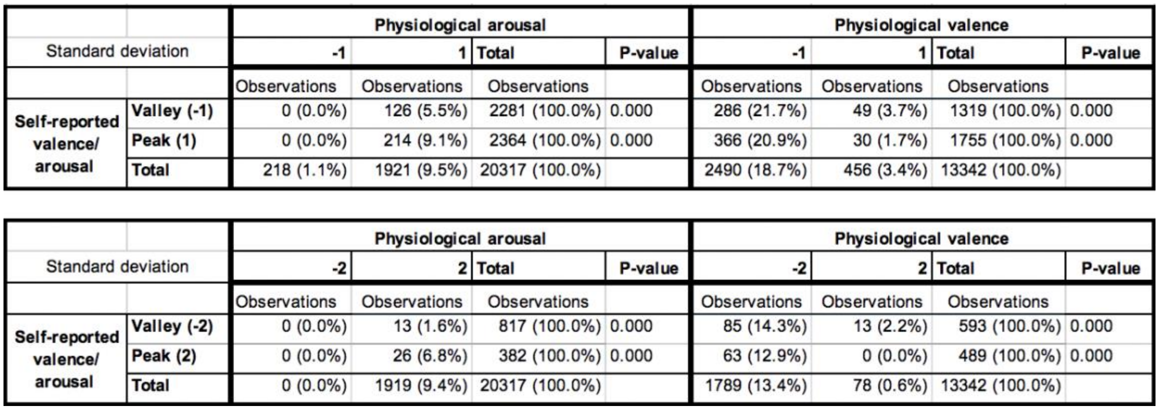
Table 9: Accuracy results between continuous self-reported and continuous psychophysiological measure of arousal and valence constructs for study 2, without baseline.

For valence results, with a SD of +/- 1 the valleys were, again in this study, better reported than the peaks (p < 0.001). With a level of accuracy of 28.2% for the valleys and 1.4% for the peaks we concluded that participants were not able to correctly report their emotions. Also, with a SD of +/-1, valleys reported were underestimated (1319 data points vs 3181 truly experimented) whereas peaks were overestimated (1755 data points vs 305 truly experimented). With a SD of +/-2, poor level of accuracy was found again. With a 6.6% of accuracy over 593 observations for the valleys and 0.2% of accuracy over 489 observations for the peaks, participant totally missed their emotional report. Valleys were underestimated because 593 data points were reported whereas 1012 data points were lived during the experiment. We observed an overestimation of the reported peak valence (489 data points vs 109 truly experimented).

We run the same analysis adding the baseline as recommended by Picard et al, (2001), presented in Table 10. As an emotional reaction to a stimulus had not the same positive or negative effect, we calculated the emotional interpersonal variation to better compare the data between subjects (Picard et al, 2001). With a psychophysiological baseline, our results were aligned with the results without baseline (Table 10). In other word, the baseline had no influence on the global conclusion of the results. For the arousal, peaks were better reported. Also, valleys were overestimated, and peaks underestimated no matter the SD. For the valence valleys were better reported. Also, valleys were underestimated, and peaks overestimated no matter the SD. Overall, participant emotional report was not correlated with psychophysiological data, as found in study 1. Thus, results confirmed our finding without baseline in Study 1 and 2.

All the results were found with a p-value < 0.001 using an adjusted chi-square test of independence. In our case, the p-value indicated that the cross results between the reported arousal/valence of valleys/peaks and the arousal/valence by FaceReader/Acknowledge of valleys/peaks were different from the other results.

Thus, again H3 is not supported.



		Physiological arousal				Physiological valence			
Standard deviation		-1	1	Total	P-value	-1	1	Total	P-value
		Observations	Observations	Observations		Observations	Observations	Observations	
Self-reported valence/ arousal	Valley (-1)	0 (0.0%)	126 (5.5%)	2281 (100.0%)	0.000	286 (21.7%)	49 (3.7%)	1319 (100.0%)	0.000
	Peak (1)	0 (0.0%)	214 (9.1%)	2364 (100.0%)	0.000	366 (20.9%)	30 (1.7%)	1755 (100.0%)	0.000
	Total	218 (1.1%)	1921 (9.5%)	20317 (100.0%)		2490 (18.7%)	456 (3.4%)	13342 (100.0%)	

		Physiological arousal				Physiological valence			
Standard deviation		-2	2	Total	P-value	-2	2	Total	P-value
		Observations	Observations	Observations		Observations	Observations	Observations	
Self-reported valence/ arousal	Valley (-2)	0 (0.0%)	13 (1.6%)	817 (100.0%)	0.000	85 (14.3%)	13 (2.2%)	593 (100.0%)	0.000
	Peak (2)	0 (0.0%)	26 (6.8%)	382 (100.0%)	0.000	63 (12.9%)	0 (0.0%)	489 (100.0%)	0.000
	Total	0 (0.0%)	1919 (9.4%)	20317 (100.0%)		1789 (13.4%)	78 (0.6%)	13342 (100.0%)	

Table 10: Accuracy results between continuous self-reported and continuous psychophysiological measure of arousal and valence constructs for study 2, with baseline.

Discussion

Overall, we found the similar results as Study 1. First, the arousal and valence of emotional states were not correctly reported (H3). Even with the addition of a baseline and focusing on the most extreme emotion (SD of 2), participants could not accurately report their emotional peaks and valleys, which is contradictory with the large literature on retrospective evaluation of peaks (Kahneman et al, 1993; Fredrickson & Kahneman 1993).

The results suggest that there is an emotional gap between the experience recorded with psychophysiological instruments and the retrospection with CSPSM from Girard & Wright (2017). However, in this paper, we do not attempt to explain if psychophysiological measures are better than CSPMS (each instrument of measure has his strengths and weaknesses). We simply describe the gap between the emotion lived during the experiment and the emotion that is reported during the retrospection. In this study, we also use a baseline to control users' level of psychophysiological activity and psychophysiological sensitivity differences. Further, participants more accurately reported arousal peaks than valleys, giving support to our (H1) hypothesis. Finally, participants, as Study 1 suggests, were more accurately reporting negative than positive valence, supporting (H2).

General discussion

The importance of instrument validation has already been demonstrated in the past, notably as a means to increasing researcher's credibility (Straub, 1989; MacKenzie et al, 2011). The objective of this paper was to examine the validity of CSPMS in user testing.

For both studies, skin electrical conductance was recorded with Acknowledge (Biopac, Goleta, USA) to measure arousal, and valence was measured by facial micro-movement recorded using FaceReader (Noldus, Wageningen, Netherlands). We measured these constructs with two different interfaces (see Webqual results) and we compared results with and without a psychophysiological baseline. All of these factors have added to the robustness of our findings.

In sum, results demonstrate that peaks of arousal were better reported than valleys in all of the valid cases (when the comparison was possible) except for the SD +1 of study 1 (H1). Further, participants are globally better at reporting negative than positive emotions in study 1 and in study 2 (H2). Finally, no correlation was observed between peaks and valleys recorded with psychophysiological data, and with continuous self-perceived measurement system data (H3). Indeed, the self-reported accuracy level was most of the time under 10% (19/24 cases) and in some cases at 0% (5/24 cases). Despite high expectations generated with these new HCI tools, results show that they should be used with caution for measuring valence and arousal in utilitarian contexts.

Our research contributes in several ways to the UX and HCI scientific community. The aim of this paper was not to compare the accuracy of CSPMS to the accuracy of psychophysiological measures, but to evaluate the user's accuracy when reporting emotions. However, the gap that we found between lived emotions during the experiment and the emotions that were reported during the retrospection could represent a lack of support for the use of CSPMS in a retrospective context, or a lack in the user's capacity to report emotions. However, these results are contradictory to the literature on retrospective evaluation of peaks, which shows that users were able to accurately report their emotions (Kahneman et al, 1993; Fredrickson & Kahneman 1993). Past research has demonstrated that ''methods are not [necessarily] equally valid measures of the underlying construct'' (Spector, 2006, p. 227). Also, it has been suggested that the use of psychophysiological methods to measure emotions, specifically the two common constructs of arousal and valence, limits some bias and allows deeper insight into the user experience (Ortiz de Guinea et al, 2014; Ortiz de Guinea & Webster, 2013; Léger et al, 2012).

Second, this paper contributes methodologically to the retrospective assessment of emotions by using the circumplex model of emotions (Russell, 1980 ; Posner, Russell & Peterson, 2005) and the joystick technique provided by Girard & Wright (2017), in a utilitarian context. This method has never previously been tested to our knowledge.

Third, this work has demonstrated that using CSPMS in a utilitarian context is possible. However, bias that participants developed while using it are significant. In fact, these instruments suggest that negative emotions and peaks of arousal are better reported, so using this instrument without validation could lead to inaccurate conclusions about a product or an interface design. In this way, we support Straub (1989) and MacKenzie et al, (2011) who highlight the importance of instrument validation. This study has tested CSPMS in a utilitarian context, and future work must be done in other contexts to further validate our findings.

Finally, our research has proposed a novel way to assess CSPMS, based on the peak and peak-end rule literature. Indeed, we know that emotions fade over time (Walker et al, 1997), so it seems difficult to imagine a user perfectly reporting his emotions at every moment of the interaction. However, using emotional peaks or important moments (primacy and recency), we capture the interaction that most strongly influenced the overall retrospection. In contrast, classic self-reported tools such as SAM scales (Bradley & Lang 1994) could not be subject to such assessment because they provided a global score without any insight about specific moments. However, CSPMS could represent such moments and may be synchronized with psychophysiological data, leading to a more precise validation. In this way, high temporal resolution such as CSPMS may be compared with psychophysiological instruments due to the possibility of synchronization and/or be compared retrospectively.

For the professionals in UX and HCI, our results present various implications. First, CSPMS are described as having the same characteristics as the psychophysiological measurement. However, they do not have the same results in a retrospective context. Using CSPMS in order to make strategic decisions in our context would have led to misleading conclusions. CSPMS should be used with caution in a utilitarian context, even if they are a cheaper way to conduct UX testing. Moreover, it is interesting to know that if professionals still use this instrument, it is more probable that participants will point out events with negative valence with high arousal.

To continue in this direction, negative valence was shown to be easier to self-report than positive valence. In other words, negative moments last longer for participant. If CSPMS could put forward the theory that "bad is stronger than good" (Baumeister et al, 2001), customers may do the same on free recall. Moreover, retrospective evaluations have been demonstrated to influence the prospective choice of repeating or avoiding an experience (Kahneman 2000; Kahneman et al, 1993), so if negative emotions are easier to self-report, it could wrongly influence future purchases.

Our research features different limitations. First, no research has investigated the validity of CSPMS and proposed a methodology to assess them. Future research could conduct a multi-trait multi-method (MTMM) matrix to further validate our results (Ortiz de Guinea et al, 2013). The experimental design should be remade to integrate self-reported instruments such as SAM scales (Bradley & Lang 1994). In this way, a comparison of the psychophysiological, self-report and continuous self-perception of the valence and arousal constructs could be conducted. We tested CSPMS with two similar interfaces, which give us a strong result for the context, but the generalization of our study is just partial. Also, given the context, one could argue that there were no important peaks or valleys, so we could have asked participants to recall different extreme moments and find out if there was a correlation with the CSPMS. Finally, our sample was sufficient but could be larger in order to gain in statistical power.

CSPMS has opened the door to more research on retrospective evaluation and retrospective bias. With high temporal continuous data on the retrospective evaluation, many improvements could be done in the field of neuroscience, for example. We hope that researchers will investigate CSPMS specificities in other fields of research, such as in psychology, while looking at the impact of memory in human behavior.

We are thankful for the financial support of the Natural Sciences and Engineering Research Council of Canada.

References

Ariely, D. (1998) “Combining experiences over time: The effects of duration, intensity changes, and on-line measurements on retrospective pain evaluations” Journal of Behavioral Decision Making, 11(1), 19–45.

Baumeister, R.F., Bratslavsky, E., Finkenauer, C. & Vosh, K.D. (2001) “Bad Is Stronger Than Good” Review of General Psychology. 5(4), 323-370

Betella, A. & Verschure, P.F.M.J. (2016) “The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions.” PLoS ONE 11 (2): e0148037. doi:10.1371/journal.pone.0148037

Boucsein W (2012) “Electrodermal Activity” Springer US.

Bradley, M.M., & Lang, P.J. (1994) “Measuring emotion: The Self-Assessment Manikin and the Semantic Differential.” Journal of Behavioral Therapy and Experimental Psychiatry, 25(1), 49-59.

Christianson, S.A. & Safer, M.A. (1996) “Emotional events and emotions in autobiographical memories” In: D.C., Rubin (ed). Remembering our past. New York: Cambridge University Press, 218–243.

Cockburn A, Quinn P, Gutwin C (2015) “Examining the Peak-End Effects of Subjective Experience” Chi 1, 357–366.

Cockburn, A., Quinn, P., Gutwin, C. (2017) “The effects of interaction sequencing on user experience and preference”. International Journal of Human-Computer Studies. 108, 89–104.

Colombetti, G. (2005) “Appraising valence” Journal of Consciousness Studies, 12(24), 103–126.

Cowie, R., Douglas-Cowie, E., Savvidou, S., Mcmahon, E., Sawey, M., Schröder, M. (2000) “Feeltrace”: An instrument for recording perceived emotion in real time”ISCA ITRW on Speech and Emotion, 19–24.

Craik, F.I.M., Lockhart, R. (1972) “A Framework for Memory Research” Journal of Verbal Learning and Verbal Behavior, 11(6), 671–684.

Dawson, M.E, Schell, A.M., Filion, D.L. (2007) “The Electrodermal system.” In: Cacioppo JT, Tassinary LG, Bernston GG (eds) Handbook of psychophysiology, Third edn. Cambride University Press, New York, 159–181

Overbeeke & P.C. Wright (Eds.), “Funology: from Usability to Enjoyment” 111-123, Dordrecht: Kluwer Academic Publishers.

Dirican, A.C., Göktürk, M. (2011) “Psychophysiological measures of human cognitive states applied in Human Computer Interaction. Procedia Computer Science 3, 1361–1367

Do, A. M., Rupert, A. V., & Wolford, G. (2008) “Evaluations of pleasurable experiences: The Peak-End rule” Psychonomic Bulletin & Review, 15(1), 96–98.

Eich, E., & Schooler, J. W. (2000) “Cognition/emotion interactions” In E. Eich, J. F. Kihlstrom, G. H. Bower, J. P. Forgas, & P. M. Niedenthal (Eds.), Cognition and emotion, 3–29, New York: Oxford University Press.

Ekman, P. (1993) “Facial expression and emotion.” American Psychologist, 48(4), 384-392.

Ekman, P., Friesen, W.V. (2003) “Unmasking the face” Cambridge, MA.

Finkenauer, C., & Rime, B. (1998) “Socially shared emotional experiences vs. emotional experiences kept secret: Differential characteristics and consequences” Journal of Social and Clinical Psychology, 17 (3), 295-318.

Fredette, M. Labonté-LeMoyne, É. Léger, P.M., Courtemanche, F., Sénécal, S. (2015) “Research Directions for Methodological Improvement of the Statistical Analysis of Electroencephalography Data Collected in NeuroIS. Information Systems and Neuroscience, 201–206.

Fredrickson, B. L., & Kahneman, D. (1993) “Duration neglect in retrospective evaluation of affective episodes” Journal of Personality and Social Psychology, 65(1), 45–55.

Geng X, Chen Z, Lam W, Zheng Q, (2013) “Hedonic evaluation over short and long retention intervals: The mechanism of the peak-end rule” Journal of Behavioral Decision Making, 26(3), 225-236.

Girard, J.M. (2017) “CARMA: Software for continuous affect rating and media annotation” Journal of open research software, 2, 1–11.

Girard J.M. & Wright A.G. (2017) “DARMA: Software for dual axis rating and media annotation.” Behavior research methods, 1–8.

Grewe, O., Nagel, F., Kopiez, R., Altenmüller, E. (2007) “Emotions over time: Synchronicity and development of subjective, physiological, and facial affective reactions to music” Emotion, 7(4), 774-788.

Grewe, O., Kopiez, R., Altenmüller, E. (2009) “The chill parameter: Goose Bumps and shivers as promising measures in emotion research” Music Perception; Berkeley, 27(1), 61-74.

Hassenzahl, M., Tractinsky, N. (2006) “User experience - A research agenda. Behaviour & information technology 25(2) 91–97.

Isbister, K., Höök, K., Laaksolahti, J., & Sharp, D. (2007) “The Sensual Evaluation Instrument: Developing a Trans-cultural Self-Report Measure of Affect.” International Journal of Human-Computer Studies, 65(4), 315-328.

Kahneman, D., Fredrickson, D. L., Schreiber, C. A., & Redelmeier, D. A. (1993) “When more pain is preferred to less: Adding a better end” Psychological Science, 4, 401– 405. doi:10.1111/j.1467-9280.1993.tb00589.x

Kahneman, D., Wakker, P. P., & Sarin, R. (1997) “Back to Bentham? Explorations of experience’s utility” Quarterly Journal of Economics, 112(2), 375–405.

Kahneman, D. (2000) “Evaluation by moments: Past and future” In D. Kahneman & A. Tversky (Eds.), Choices, values and frames. New York, NY: Cambridge University Press and the Russell Sage Foundation, 693– 708

Kandel, E.R., Schwartz, J.H., Jessell, T.M. (2000) “Principles of Neural Science” 4th ed. McGraw-Hill, Health Professions Division, New York.

Kemp, S., Burt, C. D. B., & Furneaux, L. (2008) “A test of the Peak-End rule with extended autobiographical events” Memory and Cognition, 36(1), 132–138.

Lane, R.D., Chua, P.M., Dolan R.J. (1999) “Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures.” Neuropsychologia 37(9), 989-997.

Langer, T., Sarin, R., Weber, M., (2005) “The retrospective evaluation of payment sequences: duration neglect and peak-and-end effects.” Journal of Economic Behavior & Organization, 58(1), 157–175.

Laurans, G., Desmet, P.M.A., Hekkert, P. (2009) “The emotion slider: A self-report device for the continuous measurement of emotion” Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 1–6.

Léger, P.M., Davis, F.D., Cronan, T.P., Perret, J. (2012) “Neurophysiological correlates of cognitive absorption in an enactive training context. Computers in Human Behavior, 34, 273–283.

Léger, P.M., Titah, R., Sénecal, S., Fredette, M., Courtemanche, F., Labonte-Lemoyne, É., Ortiz De Guinea, A.O. (2014) “Precision is in the eye of the beholder: Application of eye fixation-related potentials to information systems research” Journal of the Association for Information Systems, suppl. Special Issue on Methods, Tools, and Measurement, 15(1), 651–678.

Léger, P.M., Courtemanche, F., Fredette, M., Sénecal, S. (2018) “A Cloud-Based Lab Management and Analytics Software for Triangulated Human-Centered Research” Proceedings, NeuroIS Retreat 2018.

Levine, L.J. (1997) “Reconstructing memory for emotions” Journal of Experimental Psychology: General, 126(2), 165–177.

Liersch, M. J., & McKenzie, C. R. M. (2009) “Duration neglect by numbers—And its elimination by graphs” Organizational Behavior and Human Decision Processes, 108(2), 303–314.

Loiacono, E.T., Watson, R.T. & Goodhue, D.L. (2007) “WebQual: An Instrument for Consumer Evaluation of Web Sites” International Journal of Electronic Commerce, 11(3), 51-87.

Mather, M., Sutherland, M.R. (2011) “Arousal-biased competition in perception and memory” Perspectives on Psychological Science 6(2), 114–133.

Mehrabian, A. (1998) “Manual for a comprehensive system of measures of emotional states: The PAD Model.”

Mello, S.K.D., Craig, S.D., Gholson, B., Franklin, S., Picard, R., Graesser, A.C. (2005) “Integrating Affect Sensors in an Intelligent Tutoring System. Integrating affect sensors in an intelligent tutoring system “In: Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International conference on Intelligent User Interfaces. ACM Press, New York, 7–13.

Miron-Shatz, T. (2009) “Evaluating multiepisode events: Boundary conditions for the Peak-End rule” Emotion, 9(2), 206–213.

Murdoch, B.B., Lissner, E., Marvin, C. (1962) “The serial position effect of free recall” Journal of experimental psychology 64(5), 482-488.

Nagel, F., Kopiez, R., Grewe, O., Altenmuller, E. (2007) “EMuJoy: Software for continuous measurement.” Behavior Research Methods 39(2), 283-290.

Ortiz de Guinea, A., Titah, R., Leger, P.M. (2014) “Explicit and Implicit Antecedents of Users’ Information Systems Behavioral Beliefs: A Neuropsychological Investigation. Journal of Management Information Systems 30(4), 179–210.

Ortiz de Guinea, A., Webster, J. (2013) “An Investigation of Information Systems Use Patterns: Technological Events as Triggers, the Effect of Time, and Consequences for Performance.” Mis Quarterly 37(4), 1165–1188.

Ortiz de Guinea, A., Titah, R., Leger, P.M. (2013) “Measure for Measure: A two study multi-trait multi-method investigation of construct validity in IS research” Computers in Human Behavior 29(3), 833–844.

Picard, R.W., Vyzas, E., Healey, J. (2001) “Toward machine emotional intelligence: analysis of affective physiological state” IEEE Transactions on Pattern Analysis & Machine Intelligence 23(10), 1175-1191

Pillemer, D.B. (2000) “Momentous Events, Vivid Memories: How Unforgettable Moments Helps Us Understand the Meaning of Our Lives” Cambridge, MA: Harvard University Press.

Posner, J., Russell, J.A., Peterson, B.S. (2005) “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology” Development and Psychopathology 17(3), 715–734.

Redelmeier, D. A., & Kahneman, D. (1996) “Patients’ memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures” Pain, 66(1), 3–8.

Riedl, R. & Léger, P.M. (2016) “Fundamentals of NeuroIS: Information Systems and the Brain” Springer, Berlin, Heidelberg

Russell, J.A. (1980) “A Circumplex Model of Affect.” Journal of Personality and Social Psychology, 39(6), 1161-1178.

Schreiber, C. A., & Kahneman, D. (2000) “Determinants of the remembered utility of aversive sounds” Journal of Experimental Psychology. General, 129(1), 27–42.

Scherer, K.R. (2005) “What are emotions? And how can they be measured” Soc. Sci. Inf., 44(4), 695–729.

Scherer, K.R., Shuman, V., Fontaine, J.R.J., & Soriano, C. (2013) “The GRID meets the Wheel: Assessing emotional feeling via self-report” In J.R.J. Fontaine, K.R. Scherer & C. Sheng, H. and Joginapelly, T. (2012) “Effects of Web Atmospheric Cues on Users' Emotional Responses in E-Commerce” AIS Transactions on Human-Computer Interaction, 4(1), 1-24.

Soriano (Eds.). Components of emotional meaning: A sourcebook, 281-298, Oxford: Oxford University Press.

Seta, J. J., Haire, A., & Seta C. E. (2008) “Averaging and summation: Positive and choice as a function of the number and affective intensity of life events” Journal of Experimental Social Psychology, 44(2), 173–186.

Shteingart, H., Neiman, T., Loewenstein, Y. (2013) “The role of first impression in operant learning.” Journal of Experimental Psychology: General, 142(2), 476-488.

Spector, P. E. (2006) “Method variance in organizational research: Truth or urban legend?” Organizational Research Methods, 9(2), 221–232.

Straub, D. W. (1989) “Validating instruments in MIS research” MIS Quarterly, 13(2), 147–169.

Thayer, R.E. (1989) “The origin of everyday moods: Managing energy, tension and stress” New York: Oxford University Press.

Thomas, D.L., & Diener, E. (1990) “Memory accuracy in the recall of emotions” Journal of Personality and Social Psychology, 59(2), 291-297.

Van den Broek, E., Van der Zwaag, M.D., Healey, J.A., Janssen, J.H., Westerink, J.H.D.M. (2010) “Prerequisites for Affective Signal Processing (ASP)” Part IV. Paper presented at the 1st International Workshop on Bioinspired Human-Machine Interfaces and Healthcare Applications - B-Interface 2010, Valencia, Spain

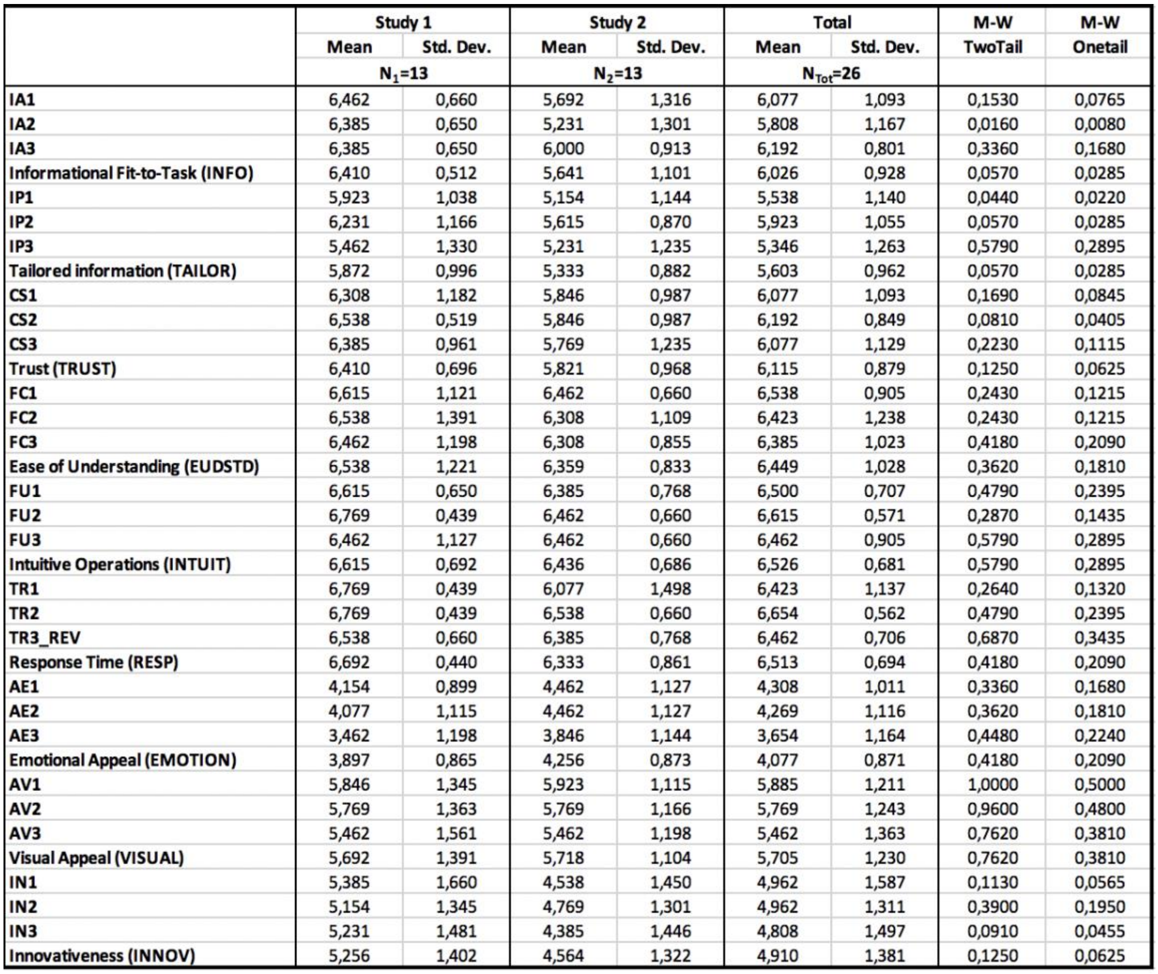
Walker, W.R., Vogl, R.G., & Thompson, C.P. (1997) “Autobiographical memory: Unpleasantness fades faster than pleasantness over time” Applied Cognitive Psychology, 11(5), 399–413

Watson, D., Wiese, D., Vaidya, J., Tellegen, A., (1999) “The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence” Journal of Personality and Social Psychology, 76, 820–838.

Zauberman, G., Diehl, K., Ariely, D. (2006) “Hedonic versus informational evaluations: Task dependent preferences for sequences of outcomes.” Journal of Behavioral Decision Making 19(3), 191–211.

Appendix A: Webqual results

The Webqual form was measuring nine (9) dimension of the perception of the website quality. The scale was from 0 to 7 for every item.



	Study 1		Study 2		Total		M-W	M-W
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	TwoTail	Onetail
	N_1=13		N_2=13		N_{Tot}=26			
IA1	6,462	0,660	5,692	1,316	6,077	1,093	0,1530	0,0765
IA2	6,385	0,650	5,231	1,301	5,808	1,167	0,0160	0,0080
IA3	6,385	0,650	6,000	0,913	6,192	0,801	0,3360	0,1680
Informational Fit-to-Task (INFO)	6,410	0,512	5,641	1,101	6,026	0,928	0,0570	0,0285
IP1	5,923	1,038	5,154	1,144	5,538	1,140	0,0440	0,0220
IP2	6,231	1,166	5,615	0,870	5,923	1,055	0,0570	0,0285
IP3	5,462	1,330	5,231	1,235	5,346	1,263	0,5790	0,2895
Tailored information (TAILOR)	5,872	0,996	5,333	0,882	5,603	0,962	0,0570	0,0285
CS1	6,308	1,182	5,846	0,987	6,077	1,093	0,1690	0,0845
CS2	6,538	0,519	5,846	0,987	6,192	0,849	0,0810	0,0405
CS3	6,385	0,961	5,769	1,235	6,077	1,129	0,2230	0,1115
Trust (TRUST)	6,410	0,696	5,821	0,968	6,115	0,879	0,1250	0,0625
FC1	6,615	1,121	6,462	0,660	6,538	0,905	0,2430	0,1215
FC2	6,538	1,391	6,308	1,109	6,423	1,238	0,2430	0,1215
FC3	6,462	1,198	6,308	0,855	6,385	1,023	0,4180	0,2090
Ease of Understanding (EUDSTD)	6,538	1,221	6,359	0,833	6,449	1,028	0,3620	0,1810
FU1	6,615	0,650	6,385	0,768	6,500	0,707	0,4790	0,2395
FU2	6,769	0,439	6,462	0,660	6,615	0,571	0,2870	0,1435
FU3	6,462	1,127	6,462	0,660	6,462	0,905	0,5790	0,2895
Intuitive Operations (INTUIT)	6,615	0,692	6,436	0,686	6,526	0,681	0,5790	0,2895
TR1	6,769	0,439	6,077	1,498	6,423	1,137	0,2640	0,1320
TR2	6,769	0,439	6,538	0,660	6,654	0,562	0,4790	0,2395
TR3_REV	6,538	0,660	6,385	0,768	6,462	0,706	0,6870	0,3435
Response Time (RESP)	6,692	0,440	6,333	0,861	6,513	0,694	0,4180	0,2090
AE1	4,154	0,899	4,462	1,127	4,308	1,011	0,3360	0,1680
AE2	4,077	1,115	4,462	1,127	4,269	1,116	0,3620	0,1810
AE3	3,462	1,198	3,846	1,144	3,654	1,164	0,4480	0,2240
Emotional Appeal (EMOTION)	3,897	0,865	4,256	0,873	4,077	0,871	0,4180	0,2090
AV1	5,846	1,345	5,923	1,115	5,885	1,211	1,0000	0,5000
AV2	5,769	1,363	5,769	1,166	5,769	1,243	0,9600	0,4800
AV3	5,462	1,561	5,462	1,198	5,462	1,363	0,7620	0,3810
Visual Appeal (VISUAL)	5,692	1,391	5,718	1,104	5,705	1,230	0,7620	0,3810
IN1	5,385	1,660	4,538	1,450	4,962	1,587	0,1130	0,0565
IN2	5,154	1,345	4,769	1,301	4,962	1,311	0,3900	0,1950
IN3	5,231	1,481	4,385	1,446	4,808	1,497	0,0910	0,0455
Innovativeness (INNOV)	5,256	1,402	4,564	1,322	4,910	1,381	0,1250	0,0625

Table 11: Results of the Webqual questionnaire for studies 1 and 2.

Appendix B: List of IAPS selected and their emotional valence and arousal

Category	# IAPS	Valence	Arousal
V+/A+	8190	2.03550654	1.51743231
	8370	1.81669601	1.98569008
	8185	1.68408358	2.5475994
V-/A+	8485	-1.5251375	1.70473542
	9620	-1.5450293	1.34053493
	9622	-1.2798044	1.49662085
V-/A-	9290	-1.4256	-0.4388
	9330	-1.419	-0.4908
	9291	-1.3925	-0.4596
V+/A-	1670	1.1801	-1.8436
	2501	1.2332	-1.8019
	5870	1.1602	-1.7915

Table 12: Repartition by category of the IAPS picture selected for the experiment and the emotional valence and arousal associated from the IAPS database (V+/A+: high valence, high arousal; V- / A+: low valence, high arousal; V- / A-: low valence, low arousal)

CHAPITRE 4 : Conclusion

Ce mémoire a pour objectif de comprendre l'usage des outils de report continu des émotions dans le but d'évaluer une expérience utilisateur. Nous souhaitions savoir si les outils psychophysiologiques et de report continu des émotions convergent vers les mêmes conclusions. Étant donné l'écart entre ces deux mesures, nous voulions observer l'impact des biais systématiques tels que la première/dernière impression et les pics émotionnels. Ainsi nous avons pu observer l'impact de ces biais sur la manière et la précision dont les utilisateurs reportaient leurs émotions. Ces résultats sont intéressants, car ils permettent d'avoir un aperçu des outils de report continu des émotions, mais aussi de l'important dont les concepteurs d'interfaces doivent traiter ces biais.

Nous avons réalisé cette expérience dans un contexte de laboratoire et sur des tâches de type utilitaire. Nous avons collecté des données psychophysiologiques (reconnaissance des émotions faciales et activité électrodermale) et des données rétrospectives évaluant les émotions post-tache de manière continue (en revisualisant sa propre interaction). Nous avons un échantillon total de 24 participants sur les deux collectes et avec deux interfaces différentes.

Nous profiterons de ce chapitre pour rappeler les questions de recherche, les hypothèses et la méthodologie nous ayant permis de conduire cette expérience. Nous récapitulerons les résultats et la contribution des deux articles. Enfin, nous exposerons les limites globales du mémoire et les recommandations pour de futures recherches.

Rappel des questions de recherche et de la méthodologie

Pour comprendre la validité des outils de report continu des émotions et l'impact des biais systématiques, nous avons mené deux expériences. Nous avons donc rédigé deux articles pour répondre aux questions de recherches ci-dessous :

Question 1 : Est-ce que la valence et l'activation émotionnelles vécues en début et fin d'interaction correspondent à la valence et à l'activation rapportées par les utilisateurs (i.e., première et dernière impression) à l'aide d'outils de report continu ?

Question 2 : Est-ce que la valence et l'activation émotionnelles vécues lors des pics émotionnels correspondent à la valence et à l'activation rapportées par les utilisateurs (i.e., pics émotionnels) à l'aide d'outils de report continu ?

Pour ce faire, nous avons évalué l'expérience utilisateur grâce à deux construits : la valence et l'activation émotionnelle. Nos hypothèses tournent autour de la validité concernant les outils de report continu des émotions et de la capacité des utilisateurs à reporter leurs émotions dans des moments sujets aux biais. Ainsi, à travers les deux articles nous avions 5 hypothèses résumées ci-dessous :

H1 : La première et dernière impression ont la même influence sur la capacité (précision) de report émotionnel.

H2: L'activation émotionnelle a plus d'influence que la valence lors de la dernière impression.

H3: Les utilisateurs reportent avec plus de précision leurs pics d'activation que leurs vallées lorsqu'ils utilisent un outil de report continu des émotions.

H4 : Les utilisateurs reportent avec plus de précision leurs vallées de valence que leurs pics lorsqu'ils utilisent un outil de report continu des émotions.

H5 : Il y a une relation positive entre les pics et vallées émotionnels (activation et valence) et le report émotionnel continu des pics et vallées (activation et valence).

D'un point de vue méthodologique, les deux expériences duraient 90 minutes chacune. L'objectif était d'utiliser deux mesures d'expérience utilisateur (pendant et après la tâche) pour tester la validité convergente de deux construits : la valence et l'activation. Aussi, nous nous intéressions à des moments précis comme la première/dernière impression ainsi que les pics émotionnels. Les données ont été codifiées et analysées grâce à des logiciels statistiques comme Stata et SPSS.

Principaux résultats

Les résultats de nos deux expériences ont permis de répondre à nos questions de recherche. Étant donné que les hypothèses sont une suite logique à nos questions de recherche, nous avons présenté les résultats ci-dessous par hypothèse.

H1 : La première et dernière impression ont la même influence sur la capacité (précision) de report émotionnel. (Supportée).

La capacité de mémoire des êtres humains a été décrite comme formant un U. Ainsi, les première et dernière impression ont biaisé les utilisateurs. En effet, seules les première et dernière impression ont été rapporté de manière presque précise.

H2: L'activation émotionnelle a plus d'influence que la valence lors de la dernière impression. (Non supportée).

L'activation émotionnelle a été reportée de manière plus précise lors de la première impression. En effet, la distance entre l'émotion vécue et rapportée est plus petite lors de la première impression. C'est la valence qui fut le mieux rapportée lors de la dernière impression.

H3 : Les utilisateurs reportent avec plus de précision leurs pics d'activation que leurs vallées lorsqu'ils utilisent un outil de report continu des émotions. (Supporté).

Les utilisateurs ont majoritairement mieux performé pour reporter leurs pics d'activation que leurs vallées. En effet, la littérature démontre une meilleure capacité des humains à reporter leurs émotions les plus intenses (en termes d'activation). En utilisant l'outil de report continu des émotions, les résultats sont confirmés.

H4 : Les utilisateurs reportent avec plus de précision leurs vallées de valence que leurs pics lorsqu'ils utilisent un outil de report continu des émotions. (Supportée).

Les utilisateurs ont reporté avec plus de précision les vallées de valence c'est-à-dire les émotions négatives lors de leurs expériences. La littérature étaye cette conclusion car le négatif a toujours été beaucoup plus marquant dans les études scientifiques qui ont été conduites au cours des deux dernières décennies.

H5 : Il y a une relation positive entre les pics et vallées émotionnels (activation et valence) et le report émotionnel continu des pics et vallée (activation et valence). (Non supportée).

Les utilisateurs n'ont pas réussi à reporter correctement leurs émotions (moins de 30% de précision lors des cas les plus performants et généralement autour de 10% ou moins). En effet, en comparant les données psychophysiologues avec les données de report continu des émotions, il n'existe pas de corrélation positive entre les deux instruments, que ce soit pour l'activation ou la valence.

Contributions du mémoire

Nos deux articles contribuent à des littératures différentes. Premièrement, ce mémoire donne un aperçu du potentiel des outils de report continu des émotions ainsi que leurs limites. En effet, ces outils donnent des possibilités en termes de rétrospections auprès d'utilisateurs, mais présentent aussi des limites, notamment lorsqu'on les compare avec des outils psychophysiologiques. Nous invitons ainsi les chercheurs à continuer le travail d'investigation sur ces outils et mettons en avant l'avantage d'utiliser des multi méthodes lors d'analyses de l'expérience utilisateur (Ortiz de Guinea et al, 2013). En effet, en n'utilisant que les outils de report continu des émotions, nos conclusions sur l'expérience utilisateur de l'interface auraient été faussées, menant à de mauvaises décisions UX dans un contexte utilitaire.

Deuxièmement, nous contribuons à la littérature sur les biais systématiques apparaissant lors de l'utilisation de ces outils (Kahneman et al, 1993). En effet, dans nos résultats nous trouvions que la valence négative était mieux rapportée que la valence positive, contribuant aux recherches de Baumeister et al. (2001). De plus, l'activation émotionnelle était mieux rapportée lorsqu'elle était intense plutôt que calme contribuant aux recherches de Fredrickson & Kahneman, (1993) ainsi que celles de Kahneman et al, (1993). Aussi, l'activation émotionnelle est mieux rapportée en début de tâche qu'en fin, et semble contredire cette fois-ci les travaux de Kahneman et al, (1993).

Implications managériales

Pour l'industrie, cette recherche est enrichissante sous plusieurs aspects. Premièrement, elle permet aux praticiens de découvrir ou d'approfondir leurs connaissances sur les outils encore nouveaux de report continu des émotions. Ainsi, dans un contexte utilitaire, les praticiens seront désormais en mesure de connaître les différences (écarts) entre des résultats obtenus lors de l'interaction avec l'interface et lors d'une rétrospection grâce à des outils de report. Cette distinction est primordiale pour les praticiens souhaitant s'équiper d'outils au sein de leur département de R&D en expérience utilisateurs. De plus, cette recherche a permis d'approfondir des biais qui impactaient l'évaluation lors de la rétrospection. En effet, cette recherche a su mettre en lumière l'importance de diminuer la valence négative en fin de tâche/expérience, car les utilisateurs s'en souvenaient particulièrement bien. Nous avons aussi mis en lumière l'importance de minimiser ou maximiser l'activation en début de tâche/expérience, car elle est très bien reportée par les utilisateurs. Ces résultats permettent à des concepteurs d'interface de prévoir en amont des expériences optimisées. Enfin, cette recherche a aussi mis en perspective l'impact des pics émotionnels et la capacité des utilisateurs à les rapporter. Les concepteurs d'interfaces pourront encore une fois anticiper, grâce à nos résultats, la manière dont ils construisent leurs interactions. En effet, les émotions négatives sont mieux reportées que les émotions positives. Enfin, les pics d'activations émotionnelles sont beaucoup mieux rapportés que les vallées.

L'ensemble de ces résultats permet de concevoir des interactions beaucoup plus optimisées et de réduire les nombres de tests utilisateurs.

Limites du mémoire et pistes de recherches futures

Notre recherche fait face à plusieurs limites. Premièrement, nous avions un panel d'utilisateurs restreint (n=24). Deuxièmement, les deux expériences de ce mémoire se sont déroulées dans un contexte utilitaire, nous ne pouvons donc pas généraliser nos propos à l'ensemble des industries. Enfin, nous avions deux interfaces similaires, nous pourrions aussi changer les interfaces pour plus de robustesse dans nos résultats.

Concernant les avenues de recherche, il serait pertinent de s'intéresser à d'autres biais pouvant intervenir lors d'un processus de rétrospection. Par exemple, il pourrait être intéressant d'étudier le biais de désirabilité social et d'observer son impact sur les construits de valence et d'activation. En effet, notre étude se limite aux biais de première/dernière impression, mais il existe de nombreux biais intervenant lors d'un processus de rétrospection. Aussi, il serait utile dans une perspective plus longitudinale d'étudier l'effet du temps sur les biais systématiques. Ainsi, les émotions s'estompent peu à peu ainsi que la capacité à s'en rappeler (Walker et al, 1997). Nous pourrions donc observer sur plusieurs semaines ou mois, l'effet de ce biais avec les outils de report continu des émotions dans un contexte de rappel sans indice. Aussi, nous avons investigué l'impact de ces biais à travers les outils de report continu des émotions (non verbal), mais probablement que ce dernier constitue lui aussi un biais dans la manière de reporter ses émotions. Nous pourrions étudier la corrélation entre ce qui est rapporté avec l'outil et ce qui est rapporté de façon verbale.

En effet, la capacité d'un utilisateur à reporter de manière précise ses émotions est une chose, savoir si ce même utilisateur a la volonté de retourner sur la plateforme en est une autre. Ainsi, dans cette expérience nous ne nous intéressons pas à la volonté perçue de ré-achat par l'utilisateur. Il serait intéressant d'étudier la corrélation entre la précision de report des émotions et la volonté ou non de refaire l'expérience/de ré-achat.

Enfin, la validité des CSPMS doit être sollicité au travers d'un test de multi traits multi méthodes (MTMM) pour confirmer les résultats (Ortiz de Guinea et al, 2013). En effet, en intégrant une échelle psychométrique comme le SAM Scale (Bradley & Lang 1994) nous pourrions aussi comparer les outils de report continu des émotions avec les outils de report des émotions traditionnels ainsi que les outils psychophysiologiques.

Bibliographie

Betella, A. & Verschure, P.F.M.J. (2016) “The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions.” PLoS ONE, 11(2): e0148037. doi:10.1371/journal.pone.0148037

Bradley, M.M. & Lang, P.J. (1994) “Measuring emotion: The Self-Assessment Manikin and the Semantic Differential.” Journal of Behavioral Therapy and Experimental Psychiatry, 25(1), 49-59.

Broekens, J. & Brinkman, W. (2013) “Affect button: reliable, valid and usable affect self-report” International Journal of Human-Computer Studies, 71(6), 641-667.

CEFRIO. (2015). ICEQ - Portrait de la situation dans les entreprises et pistes pour réussir son passage au commerce électronique. Montréal, Québec.

Christianson, S.A. & Safer, M.A. (1996) “Emotional events and emotions in autobiographical memories” In: D.C., Rubin (ed). Remembering our past. New York: Cambridge University Press, 218–243.

Cowie, R., Douglas-Cowie, E., Savvidou, S., Mcmahon, E., Sawey, M., Schröder, M. (2000) “Feeltrace”: An instrument for recording perceived emotion in real time”ISCA ITRW on Speech and Emotion, 19–24.

Dimoka, A., Pavlou, P. A., & Davis, F. D. (2011). Neurois: The potential of cognitive neuroscience for information systems research. Information Systems Research, 22(4), 687–702.

Fredrickson, B. L. & Kahneman, D. (1993) “Duration neglect in retrospective evaluation of affective episodes” Journal of Personality and Social Psychology, 65, 45–55.

Girard J.M. & Wright A.G. (2017) “DARMA: Software for dual axis rating and media annotation.” Behavior research methods, p 1–8.

Hartson, R. & Pyla, P. (2012) “The UX Book: Process and guidelines for ensuring a quality of user experience” Elsevier.

Isbister, K., Höök, K., Laaksolahti, J., & Sharp, D. (2007) “The Sensual Evaluation Instrument: Developing a Trans-cultural Self-Report Measure of Affect.” International Journal of Human-Computer Studies, 65(4), 315-328.

ISO DIS 9241-210:2008. Ergonomics of human system interaction - Part 210: Human-centred design for interactive systems (formerly known as 13407). International Organization for Standardization (ISO). Switzerland.

Kahneman, D., Fredrickson, D. L., Schreiber, C. A., & Redelmeier, D. A. (1993) “When more pain is preferred to less: Adding a better end” Psychological Science, 4, 401– 405. doi:10.1111/j.1467-9280.1993.tb00589.x

Kahneman, D. (2000) “Evaluation by moments: Past and future” In D. Kahneman & A. Tversky (Eds.), Choices, values and frames. New York, NY: Cambridge University Press and the Russell Sage Foundation, 693– 708

Léger P.M., Davis F.D., Cronan T.P., Perret J. (2012) “Neurophysiological correlates of cognitive absorption in an enactive training context” Computers in Human Behavior, 34, 273–283.

Levine, L.J. (1997) “Reconstructing memory for emotions” Journal of Experimental Psychology: General, 126, 165–177.

Lourties S., Léger PM., Sénécal S., Fredette M., Chen S.L. (2018) “Testing the Convergent Validity of Continuous Self-Perceived Measurement Systems: An Exploratory Study.” In:

Nah FH., Xiao B. (eds) HCI in Business, Government, and Organizations. HCIBGO 2018. Lecture Notes in Computer Science, vol 10923. Springer, Cham

McCarthy, J. & Wright, P. (2004), "Technology as Experience" The MIT Press, Cambridge, Massachusetts London, England

Mehrabian, A. (1998) "Manual for a comprehensive system of measures of emotional states: The PAD Model."
Menona, S. & Kahn, B. (2002) "Cross-category effects of induced arousal and pleasure on the internet shopping experience" Journal of Retailing, 78(1), 31-40

Nagel, F., Kopiez, R., Grewe, O., Altenmuller, E. (2007) "EMuJoy : Software for continuous measurement." Behavior Research Methods 39(2), 283-290.

Ortiz de Guinea, A. & Markus, M. L. (2009) "Why break the habit of a lifetime? Rethinking the roles of intention, habit, and emotion in continuing information technology use" MIS Quarterly, 33(3), 433–444.

Ortiz de Guinea, A., Titah, R., Leger, P.M. (2014) "Explicit and Implicit Antecedents of Users' Information Systems Behavioral Beliefs: A Neuropsychological Investigation" Journal of Management Information Systems 30(4), 179–210.

Pollak, J.P., Adams, P., Gay, G. (2011) "PAM: A photographic affect meter for frequent, in situ measurement of affect." Proc. of CHI 2011. New York, USA: ACM

Riedl, R., Banker, R. D., Benbasat, I., Davis, F. D., Dennis, A. R., Dimoka, A., et al. (2009) "On the foundations of neurois: Reflections on the gmundem retreat 2009" Communication of the Association for Information Systems, 27(15), 243–264.

Scherer, K.R. (2005) "What are emotions? And how can they be measured." Soc. Sci. Inf., 44(4), 695–729.

Scherer, K.R., Shuman, V., Fontaine, J.R.J., Soriano, C. (2013) “The GRID meets the Wheel: Assessing emotional feeling via self-report.” In J.R.J. Fontaine, K.R. Scherer & C. Soriano (Eds.). Components of emotional meaning: A sourcebook, 281-298, Oxford: Oxford University Press.

Walker, W.R., Vogl, R.G., Thompson, C.P. (1997) “Autobiographical memory: Unpleasantness fades faster than pleasantness over time” Applied Cognitive Psychology, 11, 399–413