HEC MONTRÉAL

Equipping UX practitioners with psychophysiological trends during moderated user testing: The impact of visualized emotions on inference and empathy

by

Pascal Snow

Pierre-Majorique Léger

Sylvain Sénécal

HEC Montréal

Research Directors

Master of Science (M.Sc.) - User Experience

Thesis presented to obtain the Master of Science – User Experience (M. Sc.)

August 2023

Résumé

L'intégration de mesure psychophysiologique des tests utilisateurs fournissent une représentation inégalée de l'expérience émotionnelle tout au long d'une interaction. Cependant, en raison des défis techniques liés au traitement des données implicites brutes et des contraintes de temps inhérentes à une session de test utilisateur typique, les praticiens de l'expérience utilisateur (UXP) ont principalement recours à l'analyse de ces informations seulement après la conclusion du test et de l'entretien. Étant donné qu'ils sont contraints de dériver des informations UX rétrospectivement, ils perdent l'opportunité d'exploiter ces données pour faire des déductions et améliorer leur compréhension de la session de test utilisateur au fur et à mesure qu'elle se déroule. Compte tenu de ces limites, nous proposons que le cadre prédominant qui sous-tend l'utilisation et l'analyse des mesures implicites soit amélioré afin de fournir un soutien plus immédiat au test utilisateur à partir duquel elles sont collectées. Comme solution, nous proposons que l'agrégation des données implicites en tendances émotionnelles représentées visuellement soutienne la performance immédiate tout au long d'un test d'utilisateur, et conduise ainsi à une compréhension plus factuelle du test d'utilisateur au moment où il se déroule.

Cette recherche étudie l'impact de l'utilisation des données psychophysiologiques à une approche de triangulation simultanée des tests utilisateurs, parallèlement aux mesures traditionnelles, en particulier les mesures auto-déclarées et l'observation comportementale. Nous évaluons l'impact de ces données sur les performances inférentielles du modérateur du test et sur l'empathie perçue à l'égard de l'utilisateur. Pour mesurer cet impact, nous avons réalisé un plan expérimental à un facteur inter-sujets impliquant 22 professionnels ayant une formation UX. Les 22 participants ont été invités à naviguer dans une session de test utilisateur simulée numériquement, au cours de laquelle ils ont été chargés des responsabilités décisionnelles impliquées dans une procédure de test typique.

Les résultats de l'expérience suggèrent que le fait de fournir aux UXP les tendances émotionnelles psychophysiologiques d'un utilisateur ainsi que des résultats psychométriques autodéclarés améliore la précision déductive en termes d'identification des problèmes d'utilisabilité et de priorisation des aspects du test utilisateur qui contiennent des incohérences entre l'interaction comportementale de l'utilisateur et la réponse autodéclarés. Enfin, les participants qui ont reçu des tendances émotionnelles psychophysiologiques ont signalé des niveaux plus élevés d'empathie cognitive et émotionnelle à l'égard de cet utilisateur.

Mots clés: triangulation simultanée, méthodes de test utilisateur, entretien avec l'utilisateur, empathie, mesures implicites, données psychophysiologiques agrégées, tendance émotionnelle, hiérarchisation, précision inférentielle.

Abstract

Incorporating psychophysiological measurements into user testing provides an unparalleled representation of emotional experience throughout an interaction. However, because of the technical challenges involved in processing raw implicit data combined with the inherent time constraints of a typical user testing session, user experience practitioners (UXPs) resort to analyzing this information only after concluding the test and interview. Since they are forced to derive UX insights retrospectively, they lose the opportunity to leverage this data to make inferences and enhance their understanding of the user testing session as it takes place. Considering these limitations, we propose that the predominant framework underlying the usage and analysis of implicit measures could be improved to provide more immediate support to the user test from which it is collected. As a solution, we propose that aggregating implicit data into visually represented emotional trends would support immediate performance throughout a user test, and thus lead to more evidence-based understanding of the user test as it takes place.

This research investigates the impact of adding psychophysiological data to a concurrent triangulation approach to user testing alongside traditional measures; specifically, self-reported measures and behavioural observation. We assess how this impacts the test moderator's inferential performance and perceived empathy towards the user. To measure this impact, we performed a one-factor between-subject experimental design involving 22 professionals with UX backgrounds. The 22 participants were asked to navigate through a digitally simulated user testing session in which they were tasked with the decision-making responsibilities involved in a typical test procedure.

Results from the experiment suggest that providing UXPs with a user's psychophysiological emotional trends alongside self-perceived psychometric scales enhances inferential accuracy in terms of identifying useability issues and prioritizing aspects of the user test that contained inconsistencies between the user's behavioural interaction and self-perceived response. Finally, participants who received psychophysiological emotional trends reported higher levels of cognitive and emotional empathy towards that user.

Keywords: concurrent triangulation, user testing methods, user interview, empathy, implicit measures, aggregated psychophysiological data, emotional trend, prioritization, inferential accuracy

Table of contents

List of tables and figures

List of figures

List of tables

List of Abbreviations and Acronyms

HCI: Human-computer interaction

UX: User experience

UXP: User experience practitioner

UXR: User experience research

PMT: Physiological measurement tools

LMUT: Live moderated user testing

SaaS: Software as a service

BRAT: Bias reducing analytic technique

ECG: Electrocardiography

EDA: Electrodermal activity

EEG: Electroencephalogram

EI: Emotional interest

UCDP: User-centered design process

UCD: User-centered design

Preface

In December 2022, the HEC Montréal Research Ethics Board (CER) approved this project (Certificate # 2023-4971) and granted authorization to conduct the experiment outlined within this thesis.

The Administrative Director of the Master of Science in User Experience authorized the following dissertation to be written in article form. The article explores the impact of providing psychophysiological emotional trends to UX practitioners while moderating a user testing.

Acknowledgements

My time at Tech3lab has been such a stark contrast to my former education. It completely shattered all preconceived notions that I had regarding academia while expanding my horizons on the interplay between research and industry. Thank you to everyone who maintains the foundation of this special institution and those who provided technical support throughout the execution of this project. Additionally, I am grateful to the Tech3lab alumni at Deloitte Digital for the guidance and friendly encouragement throughout the early stages of this research.

To my codirectors, Pierre-Majorique and Sylvain, I want to convey my admiration for the way you handle the job. Not only do you manage this environment with exemplary leadership, but you also keep it fresh; cracking jokes with stoic optimism no matter how hectic the day turns out to be. The positivity you bring to the lab is infectious and really helped with my own tenacity in the face of setbacks. Besides all the mental notes that I have taken on business and research acumen, there is one lesson you taught me that stands out above the rest: trust the process. I will carry this with me long into the future – merci beaucoup.

Next, I would like to recognize my classmates for the solidarity. It was a pleasure to embark on this journey alongside you. The work ethic and passion displayed throughout these past couple of years nourished my own perseverance and I am thankful to have had such inspiring colleagues to look up to along the way. Special thanks to Barbara for playing an instrumental role in this project and being such a compassionate, patient, and overall stellar human being. I nominate you for Best Supporting Actor in this psychological thriller.

To my friends: Thank you for listening throughout the peaks and valleys and always coming across as convincingly interested in the progression. I do not know where I would be without your corny motivational speeches. To my family – team – we have something special. Thank you for your unconditional support throughout this stage of my adult life. I will soon have this degree, and that is nice, but it pales in comparison to knowing you would support me no matter what path I take.

Finally, I would like to express my appreciation to the Natural Sciences and Engineering Research Council of Canada (NSERC) – Prompt Industrial Research Chair in User Experience and Deloitte Digital for their financial contribution and scholarship.

Chapter 1: Introduction

Don Norman first coined the term 'user experience' (UX) in 1993 upon joining the product development team at Apple (Nielson, 2017). Exactly 30 years later, this same company that first introduced and subsequently championed the importance of UX in digital product design now represents the largest market capitalization in the world (Statista, 2023). Numerous studies have demonstrated that design-driven companies outperform the market by approximately 2:1 (Meyer & Norman, 2020). In fact, companies that score highly on the Design Value Index, an indication of their overall commitment to design principles, have outperformed the S&P 500 by 228% over a ten-year period (Westcott, 2014). Companies focused on design capabilities, including their design methodology, achieve 32% more revenue growth and 56% higher shareholder returns over a 5-year period (Sheppard et al., 2018). Throughout the three decades following Apple's unprecedented decision to establish a designated UX team to lead the development of the Apple Computer, the concept of UX has continued to play a leading role in today's digital economy.

User experience has a wide range of connotations, ranging from traditional product useability to more ephemeral interpretations of experiential and emotional responses to digital interactions (Forlizzi & Battarbee, 2004). While product useability focuses primarily on quantitative metrics, individual responses to UX are inherently subjective, temporally situated, and contextually dependent (Scapin et al., 2012). Since the user's emotional state and real-world circumstances are in a state of constant flux, UX is seen as being inherently dynamic (Hassenzahl, 2008; Law et al., 2009). Hence, knowing how and why an interaction evolves over time can be as important as the outcome itself (Karapanos, 2009). As Hassenzahl and Tractinsky (2006) state in a widely cited publication, "UX takes a 'human' perspective. It is interested in understanding the role of affect as an antecedent, a consequence, and a mediator of technology use" (p. 93). For instance, useability might measure the extent to which a product supports goal-directed behaviour, while UX might assess whether a product's ease of use sparks joy or whether a convoluted onboarded process causes impatience. Therefore, UX encompasses useability, but it is also concerned with the underlying humanity of an interaction. Beyond addressing the technological needs of consumers – otherwise affectionately referred to as 'users' – they must also provide solutions that offer innovative ways to empower and improve their quality of life (Jain et al., 2019). Having a useable product is vital for the success of a product, but it is insufficient in terms of what is needed to drive positive human experiences that lead to long-term satisfaction (Forlizzi & Battarbee, 2004), and thus confidence-building experiences and brand loyalty (Hassenzahl et al., 2020; Hassenzahl, 2014; Jordan, 2000). To produce the insights needed to identify what these experiences might look like, UXPs need to go beyond useability and make a deliberate effort to empathize with users and understand their needs. Considering this, it is important to ask: Who manages the process of exploring and developing insights on end-users?

Those managing this process can be referred to as UX practitioners (UXPs). When UXPs are carrying out a test or evaluation with a user, they are *moderating* the session. Thus, the term UXP and 'moderator' will be used interchangeably depending on the discussion context. This research focuses on the role of UXP in the context of their role as moderator of a user testing session, which represents an essential component of the user-centered design process. Another key responsibility of UXPs alongside user testing is to cultivate empathy towards users, since empathizing with end users plays an instrumental role in modern business (Weichert, 2018; Temkin, 2010). It leads to the contextual understanding that is a necessary part of designing products and represents a foundational component of the user-centered Design Thinking Process (Hasso Plattner Institute of Design at Stanford University, 2010). When UX designers cultivate empathy throughout a product [re]design, it enables them to uncover hidden problems and latent concerns that neither users nor development teams are consciously aware of (Makki, 2020), insights that are not discovered through quantitative data alone (Suri, 2003). The underlying useability of a product is rudimentary, but truly empathizing with users is said to be the core propulsion mechanism for product innovation (Leonard & Rayport, 1997). This notion of going beyond the functional aspects of product development to understand the user's core values and experience with products is referred to as empathic design (Mattelmäki & Battarbee, 2002; Kouprie & Visser, 2009). Empathy has come to define the early generative ideation phase of the User-Centered Design Thinking Process (UCDP). Despite the framework's iterative nature (Dwivedi et al., 2012), evaluation and testing guidelines frequently omit referencing the importance of an empathic focus. Rather than emphasizing the importance of cultivating empathy throughout all design activities, this arbitrary rule sequesters empathic processes as existing separate from UX testing and evaluation. This undermines its importance throughout the entire process, especially user testing and interviews when UX designers are working in proximity with users.

As UXPs work towards understanding and empathizing with users across diverse contexts, it has become of strategic importance to innovate and develop the underlying methods that support designers with their work (Pine & Gilmore, 2013; Brown, 2009; Martin, 2009). Indeed, the reciprocal relationship between innovations in technology and novel user expectations perpetuates a continuous demand for technological change, which is further exasperated by the rapid pace of the digital economy (Djamasbi & Strong, 2019). To capture the dynamic interactions that users have with complex products, it becomes increasingly important that UX evaluations employ a mixed-method approach to understanding users. Questionnaires are convenient and easily administered, however they are ineffective at finding complex patterns and are found to poorly correspond with the user's actual experience (Marshall & Rossman, 1999). Other subjective reporting such as interviews provide rich qualitative insights, but are similarly cognitively mediated, and thus oftentimes fail to accurately depict the full experience (Wilson & Sasse, 2000). These are both examples of explicit measures, which can be understood as self-disclosed assessments reported by the users. These measurements dominate both in terms of UX academic research (Riedl & Léger 2016; de Guinea et al., 2014) and within corporate UX practice (Bargas-Avila & Hornbæk 2011). While this data is informative and easily assessable, when used alone, it fails to illustrate the full depth of a user's experience throughout an interaction.

In fact, when used in isolation, explicit measures are subject to various limitations that arise from their intrinsic quality: They come from the user's perception. Explicit measures are collected after the experience or task. Therefore, they have the tendency to overlook the mental thought processes that took place throughout the actual usage of the product (Ortiz de Guinea & Webster, 2013). As a result, they fail to capture the automatic mental and emotional responses that run parallel to the user's conscious awareness throughout an interaction (Guinea & Markus, 2009), which play an influential role in their final outlook and appraisal of a product (Dimoka et al., 2011). Naturally, if users were preoccupied with noting the unconscious responses taking place throughout the interaction, they would be distracted from fully engaging in the interaction itself (Ortiz de Guinea et al., 2013). To make matters worse, explicit measures are acutely susceptible to flawed judgement, erroneous memory, and various other cognitive biases that unconsciously shape their perception of a digital product (Purdy, 2021). Human memory is constantly mediated by cognitive biases, where memories are encoded by a dynamic and subjective assessment of a moment's relative importance (Kahneman & Knetsch, 1993; Kahneman & Tversky, 1997; Redelmeier &

Kahneman, 1996). These shortcomings manifest frequently in the context of user testing. For instance, the peak effect rule results in users taking selective snapshots of moments that produce a high positive or negative emotion, whereas the peak-end rule is a phenomenon where users project the end of an impression onto the entire experience (Cockburn et al., 2015). In addition, moments of the interaction that are characterized by negative emotions tend to be recalled at a higher rate than positive ones (Ariely, 1998; Baumeister et al., 2001), while the length of time post-experience is also negatively related to the likelihood to recall (Schooler & Eich, 2000). Furthermore, the success or failure of a goal-directed task will influence their overall judgement of the entire task, even if they were isolated moments that differed from the culminating sentiment (Zaman et al., 2006), making it challenging to rely on users to accurately depict how they experienced useability problems and pain points. These useability problems range from obvious product failures impeding intended action (i.e., explicit pain point), all the way to subtle discomfort that falls beneath conscious perception (i.e., implicit pain point), and thus more challenging to identify accurately and conscientiously through explicit measures (Platzer, 2018). Therefore, it is essential that UXPs factor in the possible influence of bias when interpreting the information they collect through explicit measures, since the user's interpretation of the product is mediated by countless unconscious forces.

To mitigate the limitations of explicit measures borne from cognitive failures and bias, UX researchers found that combining them with implicit measures produces a synergistic effect (Tams et al., 2014). Psychophysiological measurement tools capture physiological metrics (i.e., heart rate, perspiration, pupil dilation, brain waves, facial expressions, etc.) that indicate emotional responses throughout an interaction (Charles & Nixon, 2019) and provide temporally contextualized cues that indicate their relative pertinence and how they relate to technological features. This promotes ecological validity of the evaluation by capturing the experience as it unfolds (Bruun, 2018), while avoiding the negative emotional responses proven to occur when interrupting users throughout an interaction (Bailey et al., 2006). Existing literature has distilled emotional responses to the environment into self-reported scales across 6 dimensions [i.e., surprised-indifferent; nervous-relaxed; cheerful-depressed; quiet-anxious; enthusiastic-calm; active-passive; quiet-anxious], but this has been proven to have low reliability in terms of accurately representing the intended emotion (Bigné et al., 2005; Kumar & Oliver, 1997). Psychophysiological measures of emotion have significantly higher predictive power than self-reported measures, arising from the fact that they capture objective automatic emotional processes in real time (Lewinski et al., 2014; Poels & Dewitte, 2006). In other words, UXPs can more accurately infer aspects of the user test when incorporating implicit measures to inform their understanding. Data triangulation refers to the mix of data sources in a study, while methodological triangulation describes the use of more than one method to study a particular phenomenon (Pettersson et al., 2018). When various methods are used concurrently, it allows UXPs to cross-analyze data sets to improve data validity (Mandryk et al., 2006) and generate more quantitatively robust findings that complement the qualitative nature of explicit UX measures (de Guinea et al., 2009). This contrasts with subsequent triangulation which analyzes distinct sources of data at different points in time, rather than employing a more simultaneous approach. Prior work has shown the efficacy of using an implicit data-driven approach to identifying pain point occurrences during user testing (Giroux-Huppé et al., 2019; Mirhoseini et al., 2017), while others have conversely demonstrated the ineffectiveness of relying on explicit measures such as interviews or surveys in isolation (Fang et al., 2014). Combining methods and thus employing a triangulation approach contributes to the overall quality and reliability of UX insights.

While physiological measurement tools (PMTs) offer many solutions to the shortcomings of explicit measurements, they come with their own dilemmas when integrating them into a typical UX toolkit or workflow. In addition to the expensive technology and specialized knowledge needed to work with them, the actual raw data generated from a user test is cumbersome to interpret (Semmer et al., 2003). These tools generate massive amounts of raw data, necessitating advanced analysis tools that support their overall ease of use (Georges et al., 2017). Consequently, UX and human computer interaction (HCI) research efforts typically process, analyze, and draw conclusions from this data only after the user has left the vicinity of testing space, therefore restricting the extent to which they can leverage this data to support their understanding of the user testing session as it unfolds. These challenges impede the integration of PMTs into user testing and explain why traditional methods remain prevalent across industrial UX workflows. However, as these tools become increasingly democratized, there is an enormous opportunity to redefine how they are used to support user testing (Léger et al., 2018). There have been attempts to expedite the delivery of implicit data through live representation of raw data. For example, iMotions (Copenhagen, Denmark) and NoldusHub (Wageningen, Netherlands) both offer SaaS products aimed at providing moderators with immediate feedback from a user test. While informative, this representation framework is subject to the inherently dynamic and unsteady fluctuations of biosensors, and thus represents a sub-optimal representation format that is not conducive to effective concurrent triangulation approaches to making inferences with this genre of data.

A substantial body of research has proven the reliability of psychophysiological measurements to support the retrospective analysis of a user's emotional response throughout digital interactions (Léger et al., 2019; Giroux-Hubbé et al., 2019). Furthermore, others have investigated subsequent triangulation approaches that incorporate physiological measures into an overall evaluation approach to user testing (Mandolfo et al., 2020). However, to our knowledge, there has yet to be research that explores how providing moderators with the user's psychophysiological emotional trends impacts their immediate performance throughout a user test when employing a concurrent triangulation approach. More specifically, whether equipping moderators with a user's psychophysiological emotional trends enhances performance outcomes in terms of (1) their ability to discern useability issues as they emerge, and (2) more effectively prioritize aspects of the user test that were misreported or inconsistent. To determine whether engaging in concurrent triangulation of implicit measures, explicit measures, and behavioural observation leads to a more robust ongoing understanding of the user test as it takes place, we have developed the research question below:

RQ1: *To what extent does providing UX practitioners with a visual representation of the user's psychophysiological trends impact the practitioner's performance outcomes while moderating a user test?*

In addition to understanding performance outcomes, this research aims to explore another peripheral effect of providing moderators with implicit measures during a user test. More specifically, it aims to explore whether it plays a role in the formation of empathy. Research has demonstrated that sharing physiological biosignals can promote various prosocial behaviours between individuals ranging from empathy to trust (Winters et al., 2021; Liu et al., 2019; Curran et al, 2019). However, given that building and maintaining empathy is particularly challenging when the user is absent (Morrow, 2000), it would be important to investigate whether displaying psychophysiological data as emotional trends is enough to instigate empathic responses, especially in a digitally mediated user test and interview. Exploring emotional trends while the user is present may offer the opportunity to enhance empathy and substantiate overall understanding of the implicit data-driven findings in the context of the post-test interview. If this is the case, engaging with this data while the user is still present would represent a more valuable approach than analyzing it in isolation after the user has left the space. Stemming from this unexplored relationship, the following secondary research question was developed:

RQ2: *To what extent does providing UX practitioners with a visual representation of the user's psychophysiological trends impact the practitioner's perceived empathy towards that user while moderating a user test?*

By conducting a between-subject experiment consisting of a simulation that is meant to emulate many of the decision-making responsibilities that would be necessary during a typical user test, this thesis aims to answer the preceding research questions. Through a comparison of performance outcomes of UX practitioners who received either exclusively self-perceived scales [explicit measure] or self-perceived scales [explicit measure] in addition to psychophysiological trends [implicit measure] during a user test, we may be able to highlight the benefit of integrating physiological measures earlier in the UX testing process. Leveraging this data to immediately support an ongoing user test, in contrast to analyzing it retrospectively after the test has concluded, contributes a novel approach to the existing frameworks for implementing physiological measurement tools in the context of user testing and concurrent triangulation approaches more broadly.

Contributions

The following table summarizes the contributions made by various team members throughout this project's execution. The sections are itemized and represented as a percentage of the work done by the primary author.

Table A: Personnel contribution as a percentage executed by primary student researcher

Component	Contribution
Research questions	▫ Identifying research questions according to project objectives and existing literature – 60%
Literature review	▫ Synthesizing relevant research to substantiate the constructs and themes in research – 100%
Conception and experimental design	▫ Qualtrics simulation revision and iteration; received invaluable feedback from classmates, Tech3lab staff, and industrial research partner team – 60% ▫ Creating the Qualtrics simulation – 100% ▫ Video direction and conceptualizing user flows – 100% ▫ Video recordings and editing – HEC Montréal video production team ▫ Acting the role of the user in the experiential behavioural stimuli videos – Barbara Scheed
Pre-tests	▫ Conducted pre-tests to ensure that instructions contained within the experiment were clear; received invaluable feedback from classmates and Tech3lab staff – 80%
Participant Recruitment	▫ Recruiting participants for the study – 80% ▫ Participant screening, scheduling, and compensation management – 100%
Data collection	▫ Met all participants at Tech3lab where the experiment was conducted in a controlled lab environment – 100%
Data Analysis	

	▫ Exporting and formatting data from Qualtrics into interpretable Excel sheets – 100% ▫ Establishing test parameters – 100% ▫ Collaborative data analysis with Tech3lab statistician, Dr. Shang Lin Chen – 50%
Drafting the thesis	▫ Writing introduction, literature review, scientific article, managerial article, and conclusion – 100%

This thesis began with an introduction to the research, presenting a high-level overview of the main ideas. Following this initial introductory section, Chapter 2 will provide a literature review of the main subjects related to the research with an emphasis on justifying the relevance of this study as it pertains to UX design methodology enhancements. Chapter 3 will outline the procedural aspects of carrying out the research, as well as the results of the experiment itself. Chapter 4 will consist of a brief managerial article that is meant to serve as a framework for implementing this tool-based approach into a typical UX workflow and highlight its effect on performance outcomes and empathy. Finally, the 5th and last chapter will summarize the findings from this research and propose concluding remarks including limitations, future directions, and most importantly, its contributions to theory and practice.

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.

Bailey, B. P., Adamczyk, P. D., Chang, T. Y., Chilson, N. A.: A framework for specifying and monitoring user tasks. Computers in Human Behavior, 22(4), 709-732 (2006)

Bargas-Avila, J. A., & Hornbæk, K. (2011, May). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2689-2698)

Baumeister, R., Bratslavsky, E., Finkenauer, C., Vohs, K.: Bad is stronger than good. Review of General Psychology. 5, 323-370 (2001)

Bigné, J. E., Andreu, L., & Gnoth, J. (2005). The theme park experience: An analysis of pleasure, arousal, and satisfaction. *Tourism management*, *26*(6), 833-844.

Brown. T. 2009. Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation. New York: Harper Collins.

Bruun, A. (2018, September). It's not complicated: A study of non-specialists analyzing GSR sensor data to detect UX related events. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (pp. 170-183).

Charles, R., Nixon, J.: Measuring mental workload using physiological measures: A systematic review. Applied Ergonomics. 74, 221-232 (2019)

Cockburn A, Quinn P, Gutwin C.: Examining the Peak-End Effects of Subjective Experience. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp. 357–366 (2015)

Curran, M. T., Gordon, J. R., Lin, L., Sridhar, P. K., & Chuang, J. (2019, May). Understanding digitally-mediated empathy: An exploration of visual, narrative, and biosensory informational cues. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13)

d.school, *An Introduction to Design Thinking PROCESS GUIDE*, 2010.

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

Djamasbi, S., & Strong, D. (2019). User experience-driven innovation in smart and connected worlds. *AIS Transactions on Human-Computer Interaction*, *11*(4), 215-231.

Dwivedi, M. S. K. D., Upadhyay, M. S., & Tripathi, A. (2012). A working framework for the user-centered design approach and a survey of the available methods. *International Journal of Scientific and Research Publications*, *2*(4), 12-19.

Fang, Yulin, Israr Qureshi, Heshan Sun, Patrick McCole, Elaine Ramsey et Kai H Lim (2014). « Trust, satisfaction, and online repurchase intention: The moderating role of perceived effectiveness of e-commerce institutional mechanisms », Mis Quarterly, vol. 38, no 2.

Forlizzi, J. & Battarbee, K., 2004, Understanding experience in interactive systems. In Proceedings of the 2004 conference on Designing Interactive Systems (DIS 04): processes, practices, methods, and techniques (New York: ACM), p. 261.

Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., & Léger, P. M. (2019). Identifying psychophysiological pain points in the online user journey: the case of online grocery. In *Design, User Experience, and Useability. Practice and Case Studies: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part IV 21* (pp. 459-473). Springer International Publishing.

Hassenzahl, M.; Tractinsky, N. User experience—A research agenda. Behav. Inf. Technol. 2006, 25, 91– 97.

Hassenzahl, M. 2008. User experience (UX): towards an experiential perspective on product quality. In Proc. of the 20th international Conference of the Association Francophone D'interaction Homme-Machine. IHM '08, vol. 339. (2008) ACM, New York, NY, 11-15.

Hassenzahl, M. The Thing and I: Understanding the Relationship between User and Product. In Funology: From Useability to Enjoyment; Kluwer Academic Publishers:

Al-Azzawi, A. Experience with Technology; Springer: London, UK, 2014. Dordrecht, A2003; Volume 3, pp. 31–42. 5.

Hassenzahl, M.; Burmester, M.; Koller, F. Der User Experience (UX) auf der Spur: Zum Einsatz von. Available online: www.attrakdiff (accessed on 12 February 2020).

Jain, P., Djamasbi, S., & Wyatt, J. (2019). Creating value with proto-research persona development. In Proceedings of HCI in Business, Government and Organizations.

Jordan, P. W. (2000). Contemporary Trends and Product Design. *Contemporary Ergonomics*.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions.

Kahneman, D., & Knetsch, J. (1993). Strong influences and shallow inferences: An analysis of some anchoring effects. *Unpublished manuscript, University of California, Berkeley*.

Kouprie, M., & Visser, F. S. (2009). A framework for empathy in design: stepping into and out of the user's life. Journal of Engineering Design, 20(5), 437-448. DOI: 10.1080/09544820902875033. https://doi.org/10.1080/09544820902875033

Kumar, A., & Oliver, R. L. (1997). Special session summary cognitive appraisals, consumer emotions, and consumer response. *ACR North American Advances*.

Law, E., Roto, V., Hassenzahl, M., Vermeeren, A., and Kort, J. (2009). Understanding, Scoping and Defining User eXperience: A Survey Approach. Proc. CHI'09, ACM SIGCHI conference on Human Factors in Computing Systems.

Léger, P. M., Courtemanche, F., Fredette, M., & Sénécal, S. (2019). A cloud-based lab management and analytics software for triangulated human-centered research. In *Information Systems and Neuroscience: NeuroIS Retreat 2018* (pp. 93-99). Springer International Publishing.

Leonard, D., & Rayport, J. F. (1997). Spark innovation through empathic design. Harvard Business Review, 75(6), 102–113.

Lewinski, P., Fransen, M. L., & Tan, E. S. (2014). Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli. *Journal of Neuroscience, Psychology, and Economics*, *7*(1), 1.

Liu, F., Kaufman, G., & Dabbish, L. (2019). The effect of expressive biosignals on empathy and closeness for a stigmatized group member. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-17.

Makki, A. H. (2020). *Design Method to Enhance Empathy for User-Centered Design: Improving the Imagination of the User Experience* (Doctoral dissertation, Carleton University).

Mandryk, R. L., Inkpen, K. M., & Calvert, T. W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & information technology*, *25*(2), 141-158.

Marshall, C. & Rossman, G.B., 1999, Designing Qualitative Research, (Thousand Oaks, CA: Sage

Martin, R. 2009. The Design of Business: Why Design Thinking Is the Next Competitive Advantage. Cambridge MA: Harvard Business Press.

Mattelmäki, T., & Battarbee, K. (2002). "Empathy Probes." In PDC 02 Proceedings of the Participatory Design Conference, edited by T. Binder, J. Gregory, and Wagner, 266–271. Retrieved from http://ojs.ruc.dk/index.php/pdc/article/viewFile/265/257

Meyer, M. W., & Norman, D. (2020). Changing design education for the 21st century. *She Ji: The Journal of Design, Economics, and Innovation*, *6*(1), 13-49

Mirhoseini, S. M. M., Léger, P. M., & Sénécal, S. (2017). The influence of task characteristics on multiple objective and subjective cognitive load measures. In *Information Systems and Neuroscience: Gmunden Retreat on NeuroIS 2016* (pp. 149-156). Springer International Publishing.

Morrow, S. L., & Smith, M. L. (2000). Qualitative research for counseling psychology. *Handbook of counseling psychology*, *3*, 199-230.

Nielson, J. (2017). *A 100-year view of user experience.* Nielsen Norman Group. https://www.nngroup.com/articles/100-years-ux/

NoldusHub. (n.d.). *Noldus | Advance your behavioral research*. Retrieved from https://www.noldus.com/noldushub.

Ortiz de Guinea, A. O., Titah, R., & Léger, P. M. (2014). Explicit and implicit antecedents of users' behavioral beliefs in information systems: A neuropsychological investigation. *Journal of Management Information Systems*, *30*(4), 179-210.

Ortiz de Guinea, A. O, and Webster, j. An investigation of information systems use patterns: technological events as triggers, the effects of time, and consequences for performance. MIS Quarterly, 37, 4 (2013), 1165–1188

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

Pettersson, I., Lachner, F., Frison, A. K., Riener, A., & Butz, A. (2018, April). A Bermuda triangle? A Review of method application and triangulation in user experience evaluation. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-16).

Pine, B. J., & Gilmore, J. H. (2013). The experience economy: Past, present, and future. *Handbook on the Experience Economy*. https://doi.org/10.4337/9781781004227.00007

Platzer, D. (2018, October). Regarding the pain of users: towards a genealogy of the "pain point." In *Ethnographic Praxis in Industry Conference Proceedings* (Vol. 2018, No. 1, pp. 301-315).

Poels, K., & Dewitte, S. (2006). How to capture the heart? Reviewing 20 years of emotion measurement in advertising. *Journal of Advertising Research*, *46*(1), 18-37.

Purdy, C. (2021). *Bias in research for design: Considerations for designers when conducting user experience research* (Master's thesis, NTNU).

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.: Fundamentals of NeuroIS Studies in Neuroscience, Psychology and Behavioral Economics. Springer, Berlin, Heidelberg (2016)

Scapin, D. L., Senach, B., Trousse, B., & Pallot, M. (2012). User experience: Buzzword or new paradigm? In Proceedings of the ACHI Fifth International Conference on Advances in Computer-Human Interactions.

Schooler, J. W., & Eich, E. (2000). Memory for emotional events.

Semmer, N. K., Grebner, S., & Elfering, A. (2003). Beyond self-report: Using observational, physiological, and situation-based measures in research on occupational stress. In *Emotional and physiological processes and positive intervention strategies* (pp. 205-263). Emerald Group Publishing Limited.

Sheppard, B., Sarrazin, H., Kouyoumjian, G., & Dore, F. (2018, October 25). *The Business Value of Design*. McKinsey & Company.

https://www.mckinsey.com/capabilities/mckinsey-design/our-insights/the-business-value-of-design

Statista Research Department, & 8, A. (2023, August 8). *Biggest companies in the world by market Cap 2023*. Statista. https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-capitalization/

Suri, F. (2003). The Experience of Evolution: Developments in Design Practice, The Design Journal, 6(2), 39-48. DOI: 10.2752/146069203789355471.Retrieved from: https://doi.org/10.2752/146069203789355471

Temkin, B. D. (2010). Mapping the customer journey. *Forrester Research*, *3*, 20.

Westcott, M. (2014, March 10). *Design-driven companies outperform S&P by 228% over ten years*. https://www.dmi.org/blogpost/1093220/182956/Design-Driven-Companies-Outperform-S-P-by-228-Over-Ten-Years--The-DMI-Design-Value-Index.

Weichert, S.; Quint, G.; Bartel, T. Quick guide UX Management: So Verankern Sie Useability und User Experience im Unternehmen; Springer: Wiesbaden, Germany, 2018.

Wilson, G.M. and Sasse, M.A., 2000a, Do Users Always Know What's Good For Them? Utilizing Physiological Responses to Assess Media Quality. InHCI 2000: People and Computers XIV – Useability or Else,pp. 327 – 339 (Sunderland, UK: Springer).

Winters, R. M., Walker, B. N., & Leslie, G. (2021, May). Can you hear my heartbeat? hearing an expressive biosignal elicits empathy. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-11).

Zaman, B., & Shrimpton-Smith, T. (2006, October). The FaceReader: Measuring instant fun of use. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles* (pp. 457-460).

Chapter 2: Literature Review

Context

This chapter provides a high-level overview of UX design methods while emphasizing the relevance of this research by drawing attention to the gaps in existing literature. More specifically, it will outline core features of user experience, juxtapose methodological differences between implicit and explicit measures, and subsequently lay out a theoretical overview of empathic design as it relates to UX evaluation.

2.1 Understanding user experience

The domain of UX research continues to gain momentum, but it remains challenged by defining its overall scope (Law et al., 2014). HCI research initially focused on useability, where it was primarily concerned with behavioural goals in the context of typical product use but has since expanded into a more holistic study of UX more broadly (Hassenzahl, 2006). Over time, the narrow focus on the instrumental aspect of UX was challenged, as more research continues to pile up proving the importance of other more experiential aspects of digital interactions. For instance, the beauty of a product transcends instrumentality and is laden with intrinsic value (Postrel 2002) in terms of its contribution to our basic human needs for objects that are aesthetically pleasing (Maslow, 1954). Following this shift, scholars in the field of UX/HCI argued that future UX research must be expanded beyond the pragmatic (i.e., behavioural goals) outlook on UX to also encompass hedonic aspects such as stimulation (i.e., increasing skills), identification (i.e., self-expression), and other self-actualizing traits (Hassenzahl, 2003), which built on Logan's (1994) concept of emotional useability. This marks the beginning of more multidimensional models of UX that explicitly incorporate product attributes alongside the user's emotions and unique set of needs and values. This led to the integration of non-instrumental aspects of UX into the overall judgement of its quality (Hassenzahl, 2006).

ISO's definition (9241-11:2018) of useability explains that it is the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, learnability, ease of use and satisfaction in a specified context of use." Useability is tuned into optimizing human performance and pragmatic goals during an interaction (i.e., task success) and tends to be focused on behavioural aspects. On the other hand, ISO's definition for

UX specifies that it “includes all the users’ emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use of a product, system, or service” (9241—210:2010). Thus, UX analyzes the full range of human experience before, during, and after interaction with a digital product (Weichert, 2018; ISO, 2018; ISO, 2010; Kaye, 2007), as well as their perceptions, values, expectations, and motivations related to their product usage (Mäkelä & Suri, 2001). While ISO’s (2018) definition mentions that UX includes behaviour, implying that useability is a fundamental component of it, it is primarily concerned with optimizing overall human satisfaction according to a subjective set of personal values (Law et al., 2009). Scholars and industry practitioners alike tend to differentiate between the two, but these terms are inherently intertwined, with useability falling under the UX umbrella (Vermeeren et al., 2010; Weichert et al., 2018; Ketola & Roto, 2008; Roto, 2009; Robinson et al., 2018). Thus, for the general UX practitioner trying to understand UX, they are capturing behavioural metrics alongside more subjective and emotional dimensions, adjusting measurement approaches depending on area of inquiry (Robinson et al., 2018).

The experiential perspective of UX focuses on two fundamental aspects of the interaction: Situatedness and temporality (Hassenzahl et al., 2006). This stance defines experience as being influenced by the user’s internal state (I.e., mood, expectations, active goals) and product elements as they interact over time with a clear beginning and end (Hassenzahl et al., 2006). According to this framework, these dynamic variables interact and modify one another, resulting in the user’s actual experience. Similar research has expanded upon these dimensions to add an additional dimension: context (Djamasbi & Strong, 2019). This refers to the circumstantial conditions underlying the interaction, such as the task and physical environment. This illustrates a general trend for more modern theoretical frameworks of UX: An approach to understanding that is more holistic and involves analyzing the interplay between the user, product, and broader socio-technological context. This notion of experiential UX views the user’s experience throughout the interaction as a constant stream of inner dialogue where the assessment is dynamic and contextually dependent, and overall judgment depends on when issues are encountered throughout the interaction (Forlizzi et al., 2004). Thus, when assessing UX according to this framework, UXPs are faced with a predicament (Ariely & Carmon, 2003): How accurate is this stream of consciousness when translated into a single retrospective assessment of their experience?

Although UX researchers might expect that the user's summary would accurately depict their experience from beginning to end, countless studies have demonstrated that this is not the case (Purdy, 2021; Giroux-Hubbé et al., 2019; Cockburn et al., 2015; Ortiz de Guinea & Webster, 2013; Zaman et al., 2006; Baumeister et al., 2001; Ariely, 1998; Kahneman & Tversky, 1997). For instance, summarized assessments of a website's useability disproportionately reflect problems encountered at the end of a usage episode (Hassenzahl et al., 2004). Therefore, it is important to interpret UX as an ongoing experience rather than capturing isolated moments, or simply the culminating sentiment. Finding solutions to this challenge will lead to improved UX processes, methods, and tools used in the context of user testing.

2.2 User Experience design methods and triangulation

Design methods are any practical tool, approach, or tactic used to support the design process, ranging from creative ideation activities to measurement instruments. UX researchers continuously develop new methods, and thus keep pace with increasingly complicated user-technology interaction contexts (Robinson, 2017). However, when it comes to actual implementation, designers seem more reluctant to change established habits and have been found to arbitrarily restrict themselves to a few methods that the design team is familiar with (Rohrer, 2014). Time and resource represent additional constraints that limit their ability to iterate or introduce new methods in existing workflows (Gray, 2016). Furthermore, the majority of research on the topic is incompatible with real-world implementation contexts, resulting in scholars and professionals alike advocating for a focus on integrating academic contributions with authentic practice contexts (Gray, 2016).

Fundamentally, methods differ in their ease of use and the capital that is necessary to deploy them. Many of the tools developed by the research community, versus commercially available tools, tend to be overly complex and challenging to grasp (Følstad et al., 2012). In fact, the useability of IT tools has a drastic effect on which software services are used across industry (Cajander, 2022). Practicing designers find that the in-situ use of methods is often misaligned with the original intent of the method, implying a common disconnect between the reported intended use of methods compared to the actual design activity (Lallemand et al., 2015; Goodman, 2013; Chang, 2008; Rogers, 2004). Oftentimes, the complexity of academically sourced methods is not compatible with the temporal constraints of a practical evaluation context (Følstad et al., 2012). They are challenging to apply to practice because of their disregard for real-world conditions, which obstructs them from being efficiently integrated into bona fide business contexts (Roedl & Stolterman, 2013). Hence, the lack of established knowledge on analysis practices suggests that it will be challenging for academics to generate tools and methods that support “real-world” needs (Følstad et al., 2012).

The majority of UX research discusses methods through codified descriptions of their implementation, applying them to practice in a generic and context-free manner (Hanington et al., 2012), oftentimes through extensive lists (Gerea et al., 2015; Alves et al., 2014; Hussein et al, 2014; Vermeeren, 2010; Bevan, 2009; Ketola et al, 2008; Vredenburg, et al., 2002). In fact, studies have shown that there is stark divide in how designers use qualitative and quantitative data to inform their understanding, leveraging them in isolation instead of developing holistic and multidimensional approaches to understanding UX (Robinson, 2017). Gray (2016) points out that method descriptions are less important than understanding how they can be adapted and combined according to emergent questions that come up throughout a user test. Others have echoed this sentiment, proposing that methods should be thought of as components of a whole rather than as isolated and codified units (Rosenbaum et al., 2008; Woolrych et al., 2011). Thus, UX designers should be more intentional in how they leverage qualitative and quantitative methods alongside one another, especially given the evidence that employing both quantitative and qualitative methods are necessary to adequately capture UX (Sauro, 2016). Consequently, the absence of research that contextualizes methods in terms of alternative use cases and synergistic potential impedes the transferability of tools and methods from research to professional practice (Furniss, 2008; Følstad et al., 2012) not only creating a disconnect between the domain of research and industry practice (Buie et al., 2010), but also between qualitative and quantitative data (Sauro, 2016).

The analysis involved in interpreting data across methods is demanding (Følstad et al., 2012). Described by one researcher as the “ultimate detective work” (Rubin & Chisnell, 2008), analysis in user testing involves making observations that are subsequently converted into organized explanations of useability issues or more general UX insights. Triangularization is a valuable analysis approach that can be implemented by UXPs to help them develop confidence in the validity of their data across methods and categories of data (Pettersson, 2018). Triangulation typically occurs sequentially; where initial findings are explained or validated by future data. This is the predominant form of triangulation used in UX practice, with numerous studies emphasizing its usefulness at validating data points across sources (Leong et al., 2012; Hayashi & Hong 2015; Lederman et al., 2014). However, triangulation can also occur concurrently when quantitative and qualitative data is closely matched, leading to a truly joint analysis (Pettersson, 2018). This is especially important given that user experience is accessible in layers, where overall richness of understanding can be improved by encompassing explicit knowledge that is expressed during interviews alongside more tacit knowledge attained by observing behaviour (Visser et al., 2005; Sanders, 2002; Polanyi, 1966). Concurrent triangulation allows data to be strengthened based on correlations or contradictions, facilitating a more critical approach to corroborating, or investigating data from a user test. Woo et al., (2015) finds that making connections across data points from various sources of data has been found to strengthen research outcomes, and Pettersson (2018) corroborates that it leads to more richly defined and better validated knowledge. However, they note that concurrent triangulation is rare across both research and practice contexts, and there have been few practical contributions to understanding this cross-analysis approach to managing multiple parallel datasets (Pettersson, 2018). For concurrent triangulation to be feasible, analysis tools need to fit the quick pace of user testing (Følstad et al., 2012), while also reducing post-processing and time needed to interpret results (Georges et al., 2017).

Further research on UX design methods is important for several reasons. First, understanding the appropriate use and combination of methods is essential to adequately identify user needs (Sauro, 2016). Second, the misalignment between academically sourced methods and real-world conditions hinders their efficient integration into practical evaluation contexts, highlighting the need for research that generates tools and methods that support "real-world" needs (Følstad et al., 2012). This disconnect between academic research and industry practice in terms of contextualizing methods in the UX design process impedes the transferability of tools and methods from research to professional practice (Buie et al., 2010). Existing studies (ex. Giroux-Hubbé et al., 2019) measure how physiological measurements can be used to support retrospective understanding following a completed user test. Similarly, Mandolfo et al., (2019) examine subsequent quadrangulation that incorporates physiological measurements to contribute to understanding of a past user test. However, to our knowledge, there has yet to be research that explores the interpretation of physiological measurements in the context of concurrent triangulation in a way that mimics real-world user testing. Furthermore, critically revaluating the combinations and adaptations of methods has been found to enhance the effectiveness of UX methods in the context of user testing, allowing for a more holistic and integrated approach to evaluation (Rosenbaum et al., 2008; Woolrych et al., 2011). However, there has been little to no research exploring how physiological measurement tools can work synergistically alongside traditional methods throughout the user test, and how this may impact the UX designer's methodological approach to further inquiry. While it is true that business constraints make it unfeasible to implement the full gamut of methods for every single development cycle, further development on concurrent methodological approaches would offer valuable and timely contributions to UX evaluation practices.

2.2.1 Explicit Measures

Explicit measures are the most common category of data collected to support UX inquiries (Perrig et al., 2022; Bargas-Avila & Hornbæk, 2011; Pettersson et al., 2018). The term 'measure' can be used to describe the outcome of a given construct (i.e., score), but it can also describe the measurement technique itself (i.e., survey) (Houwer, 2006). Survey scales are the most popular form of explicit measure in UX and continue to be the most frequently used method for understanding and predicting human behaviour (Perrig, 2022; Houwer, 2006; Bruun et al., 2015; Alves et al., 2014). Surveys are a valuable tool in UX testing because they are low-cost, easily deployed, and there are many existing platforms that can support the diffusion of surveys to respondents as well as the subsequent analysis of results. They are most helpful in understanding the user's self-reported outlook on an interaction which can support the identification of problematic or successful aspects of the product. Surveys are presented to the user to elicit their perception on any established dimension of UX such as useability (i.e., SUS) (Brooke, 1996; Loiacono et al., 2002), attractiveness (Hassenzahl et al., 2003), or any other system attribute, limited only by the creativity of the UX researchers developing them. Indeed, UX researchers are continuously developing and validating new measurement scales (ex. Phan et al., 2016; Stoyanov et al., 2015; Bernhaupt et al., 2013, etc.).

Calling upon a user's retrospection is the most straight-forward method to evaluate a user's subjective experience. However, being easily gathered does not guarantee reliability or validity.

In fact, industry practices often lead to surveys that are expedited according to time constraints, and oftentimes fail to abide by best practices. For example, failure to survey existing measurement instruments, choose appropriate items, create and modifying items as necessary, and finally undergo the extensive scale development process results in scales that lack a high degree of confidence in their content, construct validity, and overall reliability (Moore & Benbasat, 1991). In fact, relying on this method alone can also be inherently problematic. First, because of the way they are collected, which happens after an experience or task, it has the tendency to overlook the thought process that occurred during the actual product use, which is what leads users to their perceptions and culminating beliefs (Guinea et al., 2013). Self-reported measures fail to capture the internalized mental processes that run in parallel to the user's individual awareness of the experience (Guinea et al., 2009; Guinea et al., 2013) which play a leading role in their appraisal of the product (Dimoka et al., 2011). Furthermore, most scales emphasize goal-oriented behaviour, thus omitting and failing to account for the hedonic qualities of user experience (Bauer et al., 2006). Thus, UX research that continues to prioritize self-reported measures, including interviews, will perpetually be afflicted by validity issues stemming from common method biases (Sharma et al., 2009; Straub et al., 2007).

Self-reported measurement scales are limited by the user's conscious awareness and perception (Riedl & Léger, 2016). When using a digital product, users are constantly evaluating and making decisions about their interaction. These decisions, perceptions, and responses are grounded in logical reasoning; they use their stored knowledge and experience to make sense of their interaction. However, one's self-perceived evaluation of personal experience is mediated through innumerable cognitive, social, and methodologically related biases, whereby the recalled experience differs from the physiological account of it (Cockburn et al., 2017; Kahneman et al., 1993; Schooler & Eich, 2000). These various cognitive and emotional constructs are not always actively perceived by the user, and therefore cannot be factored into their own self-assessment of the interaction (Barki et al., 2008; Saadé et al., 2005; Venkatesh, 2000; Agarwal, 2000 Karahanna et al., 1999). Bias affects any instance related to perceiving and recounting information, but it also relates to how users recall emotion. Research suggests that it is challenging to discern your own emotions because humans view them as overlapping and fluid as opposed to isolated, discrete states (Posner et al., 2005). This results in consumers' retrospective evaluations of their experience being inconsistent with how they experienced it (Lourties et al., 2018). These subtle emotions are crucial to explore and account for since they represent aspects of user experience that cannot be accurately reported on by users (Ouellette, 1998). Since emotions are inherently fleeting, they are difficult to accurately represent at a later point in time (Scherer, 2005), which has been corroborated by many studies (Miron-Shatz et al.; Redelmeier & Kahneman). These studies have demonstrated the peak-end rule, a psychological heuristic that leads to users disproportionately representing emotion based on a moment of peak emotion or the ending sentiment, instead of accurately representing segments of the interaction or the average (Bruun & Ahm, 2015).

To summarize, self-reported questionnaires are widely accepted as the most popular method used to measure UX phenomena because of their accessibility and cost-effectiveness. However, they have limitations related to their inability to capture the user's internal thoughts and emotions throughout an interaction. This leads to biased reporting and problems with the validity of the information provided. To combat these limitations, UX researchers have increasingly been employing mixed-method approaches, combining self-reported data with other methods, and thus leveraging the advantage of each approach.

2.2.2 Implicit Measures

Implicit measures differ from self-reported [explicit] measures in that the subject has no conscious access to the data being collected (Asendorpf et al., 2002) and has no control over the measurement outcome (Fazio & Olson, 2003). They can be thought of as the anthesis of explicit measures in the sense that they occur through automatic and unconscious mental activities 'below' the user's perceptual awareness (Greenwald & Banaji, 1995). Implicit responses are more spontaneous and less explainable by behaviour (Wilson et al., 2000), and are strongly connected to emotions (Bartoszek & Cervone, 2017). They play a critical role in influencing a user's evaluative judgments and subsequent behaviours even if they are not directly perceived (Ortiz de Guinea, 2014). Therefore, in a digital economy increasingly characterized by more experiential interactions, implicit measures are increasingly important to promote understanding of the user's visceral experience.

One effective approach for measuring emotion is through the use of psychophysiological measures (Dirican et al., 2011). Psychophysiological instruments collect various biological metrics such heart rate (i.e., electrocardiography, ECG), perspiration (i.e., galvanic skin response, electrodermal activity, EDA), pupil dilation (i.e., oculometry), brain waves (i.e., EEG), and facial expressions, which in isolation, but especially collectively, can capture the user's implicit emotional response throughout an interaction (Charles et al., 2019). While emotion itself may not lead to valuable UX insights, when contextualized and interpreted, they act as a proxy for unconscious attitudes, responses, or cognitions that occur throughout the interaction (Houwer, 2006; Brunel et al., 2002), insights that would otherwise be inaccessible. Recent studies have demonstrated the efficacy of triangulating these measures to identify useability issues, where they found that less than 25% of pain points were perceived and subsequently self-disclosed by participants during an interview following a user test because they inadvertently forgot, marginalized, or failed to notice their transient discomfort (Giroux-Hubbé et al., 2019). By incorporating psychophysiological data, we can bridge the gap between subjective perceptions and objective analysis, enabling a more robust understanding of UX that is fortified with insights into the user's emotional experience.

Integrating implicit measures into user testing has numerous benefits. First, psychophysiological measurement instruments are a more effective approach to empirically measuring peak moments of emotion (Courtemanche et al., 2017; Giroux-Hubbé et al., 2019; Swoboda et al., 2022). These measures capture the precise emotional peaks that occurred throughout the interaction, thereby providing contextual cues that indicate how pertinent emotional responses relate to technological features of the interface, while also providing quantitative data that serve as benchmarks across participants and prototypes. Because these cues are temporally situated to the interaction, they provide a narrative structure that can support UXPs with their interpretation and analysis of the user journey by connecting pain points to the technological features that instigated them. This promotes overall ecological validity of the evaluation by capturing the experience as it unfolds without having to stop the interaction to ask the user to self-report on their experience. This avoids negative emotional responses proven to occur when interrupting a user journey (Bailey et al., 2006). Furthermore, it provides a more accurate account of the emotional journey, since relying on users to delineate their own emotional response throughout the task through self-perceived measures has been proven to be extremely ineffective (Giroux-Hubbé et al. et al., 2019; Agourram et al., 2019; Xiao & Nah, 2018). Therefore, physiological measures in UX testing generate additional data from which UXPs can derive UX insights, supporting designers' ability to identify useability issues, and thus serving as an extremely valuable tool in UX evaluation.

Despite their potential to enhance UX methods, it remains important to discuss their shortcomings. Firstly, psychophysiological measures generate copious amounts of raw quantitative data, meaning that as an output, it requires time and expertise to be converted into useful and actionable insights (Georges et al., 2017). In addition, the measurement of physiological constructs (i.e., electrocardiogram, electrodermal activity, electroencephalogram, etc.) varies across implementation contexts and studies. This lack of standardization makes it challenging to replicate and validate findings across research contexts (Charles et al., 2019). Furthermore, these measures are extremely reactive to environmental factors. For example, temperature, level of humidity, and individual biological tendencies all influence the accuracy of EDA measures (Kramer, 1990). While physiological measures permit data to be collected without interruption, it can occasionally cause concerns over ecological validity, since this type of data is often collected in laboratory settings and does not always replicate real-world use (Wilson et al., 1993; Johnston et al., 1990). Finally, psychophysiological measures are enhanced by triangulating various methods to produce an emotional proxy, therefore necessitating multiple simultaneous physiological collection instruments (Charles et al., 2019). These factors pose challenges to data collection, post-processing, and subsequent analysis of implicit measures, and can explain why psychophysiological measurement tools are not more widely used across user testing (Léger et al., 2018).

Considering these high capital prerequisites alongside obstacles that complicate the implementation of implicit measures, researchers have proposed frameworks for improving the application of physiological measurements in the context of user testing. For instance, existing research outlines key considerations underlying the development of new tools including the effectiveness of identifying pain points, the ease of use of leveraging such tools, and the reduction of time needed to post-process and analyze the results (George et al., 2017). While this framework is helpful, it remains focused on the retrospective analysis of data from past user tests. To our knowledge, no research has examined these dimensions in the context of supporting the extent to which these tools can support live inferences and ongoing understanding of a user test as it takes place. By enhancing these dimensions, specifically through automating data processes and finding visualization strategies to represent data, it would drastically support UXPs ability to interpret and analyze implicit measures during a user test. Since concurrent triangulation depends on rapid analysis and comparison of data, it would be helpful to develop streamlined approaches of representing implicit measures that support on the fly analysis throughout a user test.

2.2.2.1 Psychophysiological Inference

While psychophysiological inference in inherently tied to physiology, having knowledge of physiological systems (i.e., heart rate, perspiration, brain activity, etc.) is insufficient in terms of being able to deduct psychological meaning from biological responses (Cacioppo et al., 1990). Within psychophysiological inquiry and analysis, the focus is on integrating data from multiple sources not as isolated bodily reactions, but instead as interrelated reactionary psychological mechanisms rather than the physiological structures themselves. As stated by Cacioppo (1990, p.4), "psychophysiology is based on the assumptions that human perception, thought, emotion, and action are embodied phenomena; and that measures of physical processes can therefore shed light on the human mind."

Emotion can be quantified and represented through several implicit metrics, the most common being valence and arousal (Albert et al., 2013). Ekman (1978, 1984, 1997) initially introduced the notion of identifying emotions through the distinct patterns it elicits in physiological expression through facial expressions. Through this mechanism, as well as other bodily functions, one can objectively measure physiological responses and use them as a proxy for understanding its associated emotion (Dirican & Gökturk, 2011; Ortiz de Guinea & Webster, 2013; Ortiz de Guinea et al., 2014). Valence measures the degree to which something is pleasurable. For instance, a negative valence value would indicate something is very unenjoyable or disgusting, while a positive valence would indicate something very enjoyable and pleasurable, with a value of 0 representing a neutral emotion that is neither bad nor good (Posner et al., 2005). On the other hand, arousal is defined as a user's overall activation ranging from high to low, with high activation being characteristic of heightened excitement or frustration, and low arousal being characteristic of boredom or calmness (Wiem, 2017; De Guinea et al., 2013). Valence and arousal play a complimentary role since a high arousal can be characterized as a positive or negative thing depending on its associated valence throughout the interaction. Individual dimensions of physiological measures can be used in isolation; however, they are found to be most productive when used in combination with one another, becoming increasingly accurate as they are layered upon one another (Maia & Furtado, 2016). The circumplex model of affect demonstrates how emotions move around across dimensions, with valence and arousal measurements producing a pinpoint that acts as a proxy for the inferred emotion at a given time (Posner et al., 2005). The nature of user experience results in these metrics varying over time in direction and magnitude (Wimmer et al., 2010).

Traditionally, metrics such as the SAM scale provide proxies for valence and arousal that can be attained through self-reported psychometrics scales (Morris, 1995). Users are asked to reflect on their interaction and report how they recall feeling retrospectively. However, these self-reported scales have shown low reliability and suffer from methodological shortcomings due to recall failures (Bigné et al., 2005; Kumar & Oliver, 1997). Since a user's appraisal of a task is influenced by its overall success or failure, users have been found to ignore isolated moments of the interaction that differ from the concluding sentiment (Zaman et al., 2006). Automatic physiological measures of emotion have demonstrated higher predictive power compared to self-reported measures; they capture objective, real-time automatic emotional processes, and thus provide more accurate data from which UX professionals form their insights (Lewinski et al., 2014; Poels & Dewitte, 2006). Perrig (2022) found that few constructs are measured twice or more using two separate approaches. To our knowledge, little research has analyzed how moderators cope with receiving parallel self-reported and psychophysiological measures while engaging in concurrent triangulation. Therefore, by providing UX moderators with valence and arousal from psychometric scales and distilled from physiological measurement tools, we can further understand how moderators approach inconsistencies across data sets. Not only can we assess proclivities in data utilization, but we can also observe whether it provides benefits to the UXP's performance outcomes while moderating the user test. Theoretically, by leveraging implicit measures alongside traditional measures, UXPs can more effectively infer how they should allocate their time to better understand contradictions or nuances across measures, leading to more relevant prioritization and richer insights.

There have been many different approaches to expressing biosignals (Stepanova, 2023). Novel tools such as iMotions Lab (iMotions, Copenhagen, Denmark) and NoldusHub (Noldus, Wageningen, Netherlands) can precisely represent raw data in a live format. Both companies offer SaaS programs that offer live representations of emotion that are meant to support immediate feedback to moderators. These live representations, while highly specific and sensitive, can be challenging to interpret and derive insights from, limiting the possibility for effective psychophysiological inference in the context of a moderated user test employing the program (Stepanova 2023; Slovàk et al., 2012). In fact, the inherently sporadic and “noisy” nature of physiological data means that it can produce volatile expression patterns which are not conducive to effective inference and may not be the most useful representation format for physiological data. One approach to mitigating this might be to use affective trends, where markers are deliberately placed in intervals based on aggregated data. This may reduce overall specificity while enhancing the overall interpretive ease of use for moderators to employ in the context of user testing. This deliberate dilution of marker specificity and sensitivity may support the ease of use of inherently complicated psychophysiological data and enhance the UXPs ability to draw inferences.

To summarize, psychophysiological inquiry is based on the presumption that the user’s perception, emotion, actions, and thoughts are embodied phenomenon, meaning that bodily processes, when measured, can provide valuable insights into the user’s mind (Cacioppo, 1990). Valence and arousal are commonly used to represent a vast circumplex of emotions, and when matched across explicit and implicit measures, it can support UXPs with effectively understanding discrepancies across data and support their ability to make inferences. While existing tools have identified the merits of live representations of emotion during a user test, and have developed products to support this objective, they are so sensitive to bodily fluctuations that they are challenging to derive inferences from. Instead, aggregating emotion across intervals and subsequently displaying them as emotional trends may be a more effective way to represent emotion, striking a productive balance between live representation, ease of interpretation, and overall inferential potential.

2.3 Bias in user testing

In the context of user testing, participants must filter information and make quick judgements, leading to decisions that are unconsciously skewed and self-serving (Benson, 2016). While this may seem like a deliberate and concerted process, in practice, it often occurs subconsciously through heuristics, a mental process that makes decisions and evaluations based on limited available information (Bazerman et al., 2012). They are helpful since they help reduce information overload by selectively retaining information and support thought processes by filling in the blanks when information is not readily available. (Benson, 2016). Thus, heuristics help users make sense of their interaction while coping with constraints (Simon, 1957). Much of this information can be explained by Simon's concept of bounded rationality, which posits that humans, while intelligent, are inherently limited by their ability to process data and thus unable to consistently make truly rational decisions. These patterns of flawed judgement and decision making are referred to as cognitive biases (Ellis, 2018; Wilke et al., 2012). They are ubiquitous to cognition and perception and help users make sense of their technology interactions. (Benson, 2016).

Research has proven the discrepancy between a user's objective psychophysiological response to an experience compared to how they recount it retrospectively (Giroux-Hubbé et al., 2019; Cockburn et al., 2017; Eich et al., 2000). Although this approach to navigating the world is often sufficient in everyday life, it does pose certain challenges in the context of UX. One of the most pervasive and recognizable biases is recency bias, which is the term for disproportionately basing an overall interpretation on account of the most recent aspect of the experience. Another is the familiarity/availability bias, which involves estimating the likelihood of an event based on how easily one can recall a similar instance. Another common bias is confirmation bias, where individuals tend to seek information that confirms their existing beliefs rather than challenging them. Additionally, there is the overconfidence bias, which leads people to overestimate their own accuracy or performance compared to reality. Finally, hedonic, and utilitarian aspects of an interaction are remembered remarkably differently (Langer et al., 2005). The Cognitive Bias Codex (Manoogian & Benson, 2017) classifies 187 cognitive biases, and this number is trending upwards with each passing year. Academics have been identifying hundreds of cognitive biases that impact decision-making behaviour across various contexts, but there have been minimal advances towards diminishing their impact (Ellis et al., 2018). Bias remains an elusive construct in research given the challenges involved in detecting and operationalizing its presence, thereby limiting the progress on effective mitigation approaches (Ellis et al., 2018).

In fact, bias influences UXPs who moderate user tests as well. Even the most intellectually capable people are susceptible to cognitive bias stemming from their bounded rationality (Kretz, 2018), as seen by studies showing that even experts such as physicians were found to make decisions that were subject to bias (Meehl, 1954). Studies have highlighted the effect of cognitive biases as a major contribution to pathologies that impact analytical abilities (Cooper, 2005). One established approach to minimizing bias is by ensuring methodological triangulation when studying a subject, which has been found to provide increasingly reliable and holistic understandings of phenomena (Johnson & Onwuegbuzie, 2007) and counteract the biases that influence investigators and methods (Pettersson, 2018). Method bias relates to errors that are systematically introduced via the underlying method, design, or reporting of a phenomenon being studied that may promote one result over another. Other approaches investigate whether training and awareness could diminish the effect of cognitive biases but have yielded little success (Ellis, 2018). In fact, even people made aware of cognitive bias are reluctant to concede that they may be influenced by it, demonstrating blind spot bias itself (Pronin, 2002). Bias-reducing analytic techniques (BRATS) are a novel approach to this information dilemma. BRATS involve incorporating simple techniques to minimize the effects of cognitive biases (Kretz, 2018; Kretz, 2015) For example, using snapshot tools can be beneficial for recall and sensemaking by using data-driven evidence to support decision-making. Similarly, checklists ensure that individuals do not have to rely exclusively on memory and intuition to make decisions, while being subject to external review and anticipating having your decision making or work be externally evaluated has been found to make people perform better in the first place (Hackman, 2011). While results showing the effectiveness of BRATS are mixed (Kretz, 2018), they do offer a promising approach to tool-based de-biasing methods. These techniques could be applied to user testing to minimize bias impacting both the user and UXP's perception of the experience.

Rather than focusing on modifying cognitive behaviours, research on bias is evolving towards the modification of the decision-making environment through the implementation of tools or the modification of conditions. Ellis (2018) proposes visualization strategies to mitigate bias, since visual approaches are found to support cognitive evaluations, evidence-based reasoning, and support the recollection of experience (Khan et al., 2015; Micallef, 2012). Previous studies have demonstrated inconsistencies between actual and reported experience (Giroux-Hubbé et al., 2019; Cockburn et al., 2017; Eich et al., 2000). However, even with such tools that can quantitatively indicate problems, it is important to combine this with human judgement to determine the context of the insight and minimize bias (Giroux-Hubbé et al., 2019). Incorporating implicit measurements alongside traditional (i.e., explicit) methods could represent a novel approach to mitigating bias. Not only does it provide more quantitative robust findings (de Guinea et al., 2009), but it also provides alternative sources of information that could support UXPs with identifying discrepancies across explicit and implicit measurements that indicate biased reporting. To our knowledge, no studies have examined a moderator's ability to detect these inconsistencies in real-time during a user test. Incorporating implicit measurements alongside traditional (i.e., explicit) methods could represent a novel approach to mitigating bias. Visualizations of emotional trends during a user test may be an effective BRAT, not only helping the moderator identify inconsistencies in user reporting, but also to support overall sensemaking. Despite recognition of the pervasive effect of biases on data reliability, there has been limited progress on detection and mitigation specifically (Ellis, 2018). Therefore, investigating the effectiveness of BRATS, or any other tool-based support such as physiological measurement tools more broadly, may help foster a more 'supportive' decision-making environment and would be a helpful approach to indirectly minimizing bias through the improvement of decision-making processes and help bridge the gap between BRAT research and practically implementable approaches. With further research into the effects of visualizing implicit measures through psychophysiological trends and using it in the context of concurrent triangulation of a user test, we may be able to find ways to enhance the reliability and accuracy of data collection practices.

2.4 Identifying useability problems in user testing

A pain point occurs when the user experiences a negative reaction to the interaction artefact, typically arising due to some obstruction to goal directed behaviour (Platzer, 2018). Pain points can be explicit, which typically means that not only is the negative sentiment abundantly clear, but also the source of the discomfort. However, they can also be implicit, meaning that they occur below conscious perception, and are usually more subtle in their interference with goal-directed behaviour. Previous research has found that physiological measurement tools (PMT) can indicate useability issues by capturing implicit data and linking it to the feature that triggered the pain point (Ahlstrom et al., 2006; Giroux-Hubbé et al., 2019). This allows UXPs to have a more precise contextual understanding of the root cause of these pain points without disrupting the flow of a user testing session.

Giroux-Hubbé (2019) demonstrated the success of visually representing pain points in a temporal context, enabling precise identification of when they occurred within task-level user journeys. Most notably, participants self-disclosed less than 25% of the implicitly identified pain points during interviews after user testing (Giroux-Hubbé, 2019), in line with other research highlighting the ineffectiveness of relying on interviews or surveys to identify pain points (Fang et al., 2014). While Giroux-Hubbé's study focuses on pain points self-identified by the user against an objective count of implicit pain points detected, it would be insightful to explore how useability problems are detected by moderators depending on whether they have access to implicit data, essentially inverting the inquiry perspective. While this study focused on pain points, another approach would be to visualize data in a way that not only emphasizes pain points, but also highlights moments of peak positive emotion. As stated by prominent UX researchers, the field is moving beyond functionality; rather than merely minimizing pain points, it is also important to understand the nature of positive interactions that promote wellbeing and identify the strength of a product in terms of its capacity to spark joy (Hazzenzahl et al., 2006). Creating visualization tools that not only identify pain points, but also support the identification of peak positive moments would support this new wave perspective on UX.

Giroux-Hubbé's outgoing recommendation (2019, p.69) was to use the visualization of pain points to generate a deeper understanding and generalize the use of pain points. Because users forget, marginalize, or fail to notice pain points, equipping UXPs with visually represented psychophysiological emotional trends would be a productive enhancement to the user testing toolkit. Because of the transience and subtleties of pain points, implicit measures are helpful in the prediction and identification of useability problems that negatively affect UX. By visually representing emotional responses over time, whether high or low, UX practitioners can gain a better contextual understanding of user testing sessions without interrupting the user.

2.5 Empathy in design

Empathy is a socio-emotional phenomenon that occurs when one understands, reacts, and resonates with the thoughts and experiences of another, resulting in an improved understanding of their emotions and point of view (Yalçin, 2019). Empathy has been found to support professionals across a broad range of professional environments; patients of empathic doctors are more satisfied with their care (Kim et al., 2004) and employees of empathic managers feel less stress from their jobs (Scott et al., 2010). It plays an essential role in the formation and maintenance of social bonds because it promotes mutual understanding and prosocial behaviours (Omdahl, 1995). Cultivating an empathic approach to understanding the user's world facilitates the designer's ability to see the world through their eyes, and thus make better decisions throughout the design phase of product development (Kaasinen et al., 2015).

The theoretical framework for empathy research is divided across two categories: The emotional and cognitive dimensions of empathy (Yalçin, 2019). Affective [emotional] empathy is the automatic mimicry that occurs in response to another's emotional expression, which is speculated to be instigated by the perception-action mechanism (PAM) (Preston & De Waal, 2022; Batson, 2009). It is an instinctual and shared response where the empathizer mirrors the experience of the person experiencing the emotion. While emotional empathy involves *feeling* what the other person is experiencing, cognitive empathy is used to describe circumstances where one *understands* the feelings of another from their perspective and point of view (New et al., 2013). It plays a greater role in making sense of another person's firsthand experience (Gasparini, 2015). Thus, cognitive empathy bolsters understanding of the needs throughout the design process, particularly in terms of discovery and interpretation (Brown, 2009; Brown et al., 2010), to provide helpful context for design-related decision making. By cognitively empathizing with the perspective of users, it incorporates their firsthand experience into the Design Thinking process (Gasparini, 2015).

In the context of UX, empathic design (ED) refers to the designer's ability to understand the environment, wants, needs, and feelings of users such that the resulting product reflects the requirements and desires of current and future users (Drouet et al., 2022; Surma-Aho & Hölttа-Otto, 2022; Wright & McCarthy, 2008; Mattelmäki et al., 2014). ED has been found to promote creativity, foster innovation in product design (Dorst & Cross, 2001; Gasparini, 2015), and enhance designers' ability to generate effective design insights (Ho et al., 2011; Zingoni, 2019). Despite the widely regarded importance of empathy in the design thinking process (The Hasso The Interaction Design Foundation, 2023; Plattner Institute of Design at Stanford, 2010), empathy in design research remains operationally vague and challenging to document (Wright et al., 2008) and is thus broadly understudied (Chang-Arana et al., 2020; Heylighen & Dong, 2019; Kouprie & Visser, 2009). Despite research outlining how engaging users in meaningful ways contributes to the 'cross-fertilization of knowledge' (Bogers & Horst, 2013) and interweaves all parties' individual knowledge into the design process (Heylighen & Devlieger, 2007), design teams tend to disproportionately rely on their own values, prior knowledge, and expert understanding (Strickfaden et al., 2009). Rather than engaging in collaborative knowledge transfers between users and UXPs, design teams have been found to rely on distanced approaches, as seen by the fact that existing empathy tools tend to either refer to conceptual approaches (i.e., personas) or more literal tools that simulate disability or impairment. While these tools are useful, they do not directly implicate users as experts or foster meaningful connections (Strickfaden et al., 2009). As pointed out by Følstad et al., (2012, p.2133) in their study on UX evaluation methods, they identify how calling on the user's own experience is a method which they describe as the "pragmatic exploitation of the available resources." Therefore, to develop empathy with users, it is essential that designers engage, listen, and apply their understanding in a way that implicates users (Cipolla & Bartholo, 2014). Empathy is widely regarded as a key component of the user-centered design process and contributes to successful design outcomes (Gasparini, 2015). However, Design Thinking frameworks consistently constrain empathy to the first step of the design process, and existing literature rarely emphasizes the importance of empathy during user testing specifically.

The notion that empathy can be stimulated through shared emotional signals has been established (Singer et al., 2009; Goldman, 2011; de Waal et al., 2017; Preston et al., 2002). This innate biological and emotional response has an evolutionary basis, where to be empathic of one another had utility to human culture (Davis, 1994). In much the same way that it is unlikely that one can entirely regulate their implicit response (i.e., the emotions that surface in response to some stimuli), research has found that empathy may also have a similarly involuntary and affective feedback mechanism. Notably, observing the social biosignals of another person can influence one's own physiology (Feijt et al., 2023; Liu, 2019; Howell et al., 2016). Similarly, there are prominent claims that virtual reality can foster empathy by representing expressed emotion through digital simulations (Bollmer, 2017). However, to our knowledge, there has been no research that has explored this relationship in the context of the user-moderator relationship within a digitally mediated user-testing session or as a tool that can be implemented to stimulate empathy throughout design processes.

To summarize, empathy is an indispensable quality in design, and remains an important consideration that underlies the whole design process, supporting the formation of user insights from generative phases all the way to user testing (Brown, 2009). This sensitivity towards users can even be as important as other forms of design competence and knowledge, representing an especially important quality in UX designers (Kress et al., 2012). By developing an empathic understanding of the needs and values of users, among other previously discussed aspects of useability and UX, it supports the Design Thinking process by generating more relevant solutions (Gasparini, 2015) and fostering innovation and creativity in product development (Cross, 2011). Thus, empathy should be thought of as a helpful trait that can be leveraged like any skill to support the generation of relevant and innovative design solutions. Finally, while existing tools meant to stimulate empathy either relate to conceptual understanding (i.e., personas) or literal impediments (i.e., vision impairing glasses), purposefully using visual expressions of emotion via psychophysiological trends may represent a novel category of empathy tool to support UX design processes.

2.6 Summarizing the gap

As user experience has transitioned from useability-focused to more holistic interpretations of human experience, it necessitates more sophisticated approaches to measuring human experience, especially as interaction contexts become increasingly complex and multidimensional. As an industry, researchers are calling for a move away from inherently divided quantitative and qualitative UX research methods, instead opting for a more integrated approach to understanding (Robinson et al., 2018). Thus, it would be valuable to investigate whether combining implicit measures, explicit measures, and behavioural observation produces richer understandings of experience such that it contributes to a more effective posttest interview.

While psychophysiological measurement approaches to support UX are increasingly popular, they come with many drawbacks related to the complexity of integrating them into typical UX workflows. They generate copious amounts of raw quantitative data and require extensive time and expertise to be converted into useful and actionable insights (Georges et al., 2017). This is in line with Gray's (2016) observation that many contributions to methodological research fail to provide solutions that are compatible with authentic practice contexts. Indeed, many of the tools developed across research contexts tend to be overly complex and challenging to grasp compared to commercially available tools (Følstad et al., 2012), implying a common disconnect between the reported intended use of methods compared to the actual design activity (Lallemand, 2015; Goodman, 2013; Chang, 2008; Rogers, 2004). Even the most cutting edge commercially available tools represent live data in a way that is not necessarily optimal for the time constraints of a typical user test, where highly fluctuating data is challenging to make inferences from. Considering the critiques of existing methods, we aim to investigate tools that are specifically designed for in situ approaches. To our knowledge, there has yet to be research that explores the use of aggregated psychophysiological emotional trends in the context of a concurrent triangulation, and whether this represents an effective representation format for UXPs moderating a user test.

Arhippainen et al. (2013) have shown that applying several methods over the course of a user test can help researchers catch user experience “piece by piece” but insist that there is a lack of research on how multiple methods work together synergistically to produce knowledge that is bigger than the sum of the individual parts. While many studies have explored a subsequent triangulation approach to analyzing various methods (Leong et al., 2012; Hayashi & Hong 2015; Lederman et al. 2014), less attention has been paid to concurrent approaches that focus on truly matched joint analysis of qualitative and quantitative data (Pettersson et al., 2018). Considering that data triangulation provides so many positive effects to research outcomes (Woo et al., 2015), it is especially important to note that the research and implementation case studies on concurrent triangulation remain sparse (Pettersson et al., 2018). Thus, we aim to contribute to this gap in existing literature on concurrent triangulation; more specifically, incorporating the psychophysiological dimension alongside traditional methods.

Given that users demonstrably lack the capacity to adequately account for all useability issues (Giroux-Hubbé et al., 2019), it is neither fair nor accurate to rely on them to provide insights on everything. These cognitive failures are commonly identified, but surprisingly, there has been little research on mitigating their effect (Ellis, 2018). Ellis (2018) proposes visualization strategies to mitigate bias, since visual approaches are found to support sensemaking in cognitive evaluations (Khan, 2015; Micallef, 2012). It is suggested that they help externalize the recollection of an experience, and thus may be able to diminish shortcomings in cognition through evidence-based reasoning (Kretz, 2013). In this sense, data layering is not just combining various sources of data to derive distinct insights. Instead, it is the process of combining sources of information to enhance data validity. Investigating how moderators leverage matched sets of data (i.e., valence and arousal matched across implicit and explicit measures) to identify and subsequently respond to inconsistencies in user reporting may shed light on how this may be used as a bias-reducing analytic technique to support user testing.

Finally, UX designers tend to disproportionately leverage their own values, prior knowledge, and expert understanding in isolation (Strickfaden et al., 2009). Designers oftentimes research users and gather data without necessarily using it to support collaborative approaches (Strickfaden et al., 2009). Similarly, the predominant approach to leveraging psychophysiological measures involves deriving insights retrospectively on a past user test, and thus impedes the designer's ability to have meaningful dialogue with the user that is supported by the insights afforded by implicit data. Considering that building and maintaining empathy is particularly challenging when the user is absent (Morrow, 2000), it would be valuable to investigate whether having access to this information while the user is still present would support empathic tendencies. Finally, given that empathy can be stimulated by visual expressions of emotion (Singer et al., 2009; Goldman, 2011; de Waal et al., 2017; Preston et al., 2002) it would be valuable to determine whether emotions expressed as psychophysiological trends would be a sufficient form of emotional expression to trigger a similar empathic response.

References

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly*, 665-694.

Agourram, H., Alvarez, J., Sénécal, S., Lachize, S., Gagné, J., & Léger, P. M. (2019). The relationship between technology self-efficacy beliefs and user satisfaction–use experience perspective. In *Human-Computer Interaction. Design Practice in Contemporary Societies: Thematic Area, HCI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part III 21* (pp. 389-397). Springer International Publishing.

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International journal of industrial ergonomics, 36*(7), 623-636.

Albert, J., Diéguez-Risco, T., Aguado, L., & Hinojosa, J. A. (2013). Faces in context: M odulation of expression processing by situational information. *Social neuroscience, 8*(6), 601-620.

Arhippainen, L., Pakanen, M., & Hickey, S. (2013). Mixed UX methods can help to achieve triumphs. In *Proceedings of CHI 2013 Workshop "Made for Sharing: HCI Stories for Transfer, Triumph, and Tragedy* (pp. 83-88).

Ariely, D., & Carmon, Z. (2003). *Summary assessment of experiences: The whole is different from the sum of its parts*. Russell Sage Foundation.

Al-Azzawi, A. Experience with Technology; Springer: London,UK, 2014. Dordrecht, A2003; Volume 3, pp. 31–42. 5.

Alves, R., Valente, P., & Nunes, N. J. (2014, October). The state of user experience evaluation practice. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (pp. 93-102).

Bailey, B. P., & Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behavior*, *22*(4), 685-708.

Barki, h.; Paré, G.; and Sicotte, C. Linking It implementation and acceptance via the construct of psychological ownership of information technology. Journal of Information Technology, 23, 4 (2008), 269–280.

Bartoszek, G., & Cervone, D. (2017). Toward an implicit measure of emotions: Ratings of abstract images reveal distinct emotional states. *Cognition and Emotion*, *31*(7), 1377-1391.

Batson, D.; Fultz, J; & Schoenrade, P. "Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences", *J. Pers.*, vol. 55, no. 1, pp. 19-39, 1987.

Bauer, H. H., Falk, T., & Hammerschmidt, M. (2006). eTransQual: A transaction process-based approach for capturing service quality in online shopping. *Journal of business research*, *59*(7), 866-875.

Bazerman, M. H., & Moore, D. A. (2012). *Judgment in managerial decision making*. John Wiley & Sons.

Benson, B. (2016). Cognitive bias cheat sheet. better humans. https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18.

Bernhaupt, R., & Pirker, M. (2013). Evaluating user experience for interactive television: towards the development of a domain-specific user experience questionnaire. In *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part II 14* (pp. 642-659). Springer Berlin Heidelberg.

Bevan, N. (2009, August). What is the difference between the purpose of useability and user experience evaluation methods. In *Proceedings of the Workshop UXEM* (Vol. 9, No. 1, pp. 1-4).

Bigné, J. E., Andreu, L., & Gnoth, J. (2005). The theme park experience: An analysis of pleasure, arousal, and satisfaction. *Tourism management*, *26*(6), 833-844.

Bogers, M., & Horst, W. (2014). Collaborative prototyping: Cross-fertilization of knowledge in prototype-driven problem solving. *Journal of Product Innovation Management*, *31*(4), 744-764.

Brooke, J. (1996). Sus: a "quick and dirty'useability. *Useability evaluation in industry*, *189*(3), 189-194.

Brown, T. A., Sautter, J. A., Littvay, L., Sautter, A. C., & Bearnes, B. (2010). Ethics and personality: Empathy and narcissism as moderators of ethical decision making in business students. *Journal of Education for Business*, *85*(4), 203-208.

Brunel, F. F., Ruth, J. A., & Otnes, C. C. (2002). Linking thoughts to feelings: Investigating cognitive appraisals and consumption emotions in a mixed-emotions context. *Journal of the Academy of Marketing Science*, *30*, 44-58.

Bruun, A. (2018, September). It's not complicated: A study of non-specialists analyzing GSR sensor data to detect UX related events. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (pp. 170-183).

Buie, E., Dray, S., Instone, K., Jain, J., Lindgaard, G., & Lund, A. (2010). How to bring HCI research and practice closer together. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 3181-3184).

Johnson, R. & Onwuegbuzie, A., and Lisa A Turner. 2007. Toward a definition of mixed methods research. Journal of mixed methods research 1, 2 (2007), 112–133.

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). Psychophysiological science: Interdisciplinary approaches to classic questions about the mind. *Handbook of psychophysiology*, *3*, 1-16.

Cajander, Å., Larusdottir, M., & Geiser, J. L. (2022). UX professionals' learning and usage of UX methods in agile. *Information and Software Technology*, *151*, 107005

Chang, Y., Lim, K., and Stolterman, E. 2008. Personas: From theory to practices. In Proceedings of the 5th Nordic Conference on Humancomputer Interaction: Building Bridges, 439-442.

Cipolla, C., & Bartholo, R. (2014). Empathy or inclusion: A dialogical approach to socially responsible design. *International Journal of Design*, *8*(2).

Chang-Arana, Á. M., Piispanen, M., Himberg, T., Surma-aho, A., Alho, J., Sams, M., & Hölttä-Otto, K. (2020). Empathic accuracy in design: Exploring design outcomes through empathic performance and physiology. *Design Science*, *6*, e16.

Charles, R., Nixon, J.: Measuring mental workload using physiological measures: A systematic review. Applied Ergonomics. 74, 221-232 (2019)

Cockburn, A., Quinn, P., Gutwin, C.: The effects of interaction sequencing on user experience and preference. International Journal of Human-Computer Studies. 108, 89104 (2017)

Cooper, J. R. (2005). *Curing analytic pathologies: Pathways to improved intelligence analysis* (p. 6). Washington, DC: Center for the Study of Intelligence.

De Waal, F. B. (2007). The Russian doll model of empathy and imitation. *On Being Moved: From Mirror Neurons to Empathy.* Advances in consciousness research

De Waal, F. B., & Preston, S. D. (2017). Mammalian empathy: behavioural manifestations and neural basis. *Nature Reviews Neuroscience*, *18*(8), 498-509.

Decety, J., & Batson, C. D. (2009). Empathy and morality: Integrating social and neuroscience approaches. *The moral brain: Essays on the evolutionary and neuroscientific aspects of morality*, 109-127.

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

Dirican, A. C., & Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science*, *3*, 1361-1367.c

Djamasbi, S., & Strong, D. (2019). User experience-driven innovation in smart and connected worlds. *AIS Transactions on Human-Computer Interaction*, *11*(4), 215-231.

Dorst, K., & Cross, N. (2001). Creativity in the design process: co-evolution of problem–solution. *Design studies*, *22*(5), 425-437.

Drouet, L., Bongard-Blanchy, K., Koenig, V., & Lallemand, C. (2022, April). Empathy in design scale: development and initial insights. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1-7).

Dorst, K., & Cross, N. (2001). Creativity in the design process: co-evolution of problem–solution. *Design studies*, *22*(5), 425-437.

Ellis, G. (2018). So, what are cognitive biases?. *Cognitive biases in visualizations*, 1-10.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1978). Emotion in the human face: guide-lines for research and an integration of findings. *(No Title)*.

Ekman, P. (1984). Expression and the nature of emotion. *Approaches to emotion*, *3*(19), 344.

Ekman, P. (1997). Expression or communication about emotion.

Fang, Yulin, Israr Qureshi, Heshan Sun, Patrick McCole, Elaine Ramsey et Kai H Lim (2014). « Trust, satisfaction, and online repurchase intention: The moderating role of perceived effectiveness of e-commerce institutional mechanisms », Mis Quarterly, vol. 38, no 2.

Feijt, M. A., Westerink, J. H., De Kort, Y. A., & IJsselsteijn, W. A. (2023). Sharing biosignals: An analysis of the experiential and communication properties of interpersonal psychophysiology. *Human–Computer Interaction*, *38*(1), 49-78.

Følstad, A., Law, E., & Hornbæk, K. (2012, May). Analysis in practical useability evaluation: a survey study. In *proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2127-2136).

Forlizzi, J. and Battarbee, K., 2004, Understanding experience in interactive systems. In Proceedings of the 2004 conference on Designing Interactive Systems (DIS 04): processes, practices, methods, and techniques (New York: ACM), p. 261.

Fuentes, C., Gerea, C., Herskovic, V., Marques, M., Rodríguez, I., & Rossel, P. O. (2015). User interfaces for self-reporting emotions: a systematic literature review. In *Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information: 9th International Conference, UCAmI 2015, Puerto Varas, Chile, December 1-4, 2015, Proceedings 9*(pp. 321-333). Springer International Publishing.

Furniss, D. (2008). *Beyond Problem Identification: Valuing methods in a'system of useability practice'*. University of London, University College London (United Kingdom).

Gasparini, A. A. (2015). Perspective and Use of Empathy in Design Thinking. Advancements in Computer-Human Interaction, ACHI, Lisbon.

Georges, V., Courtemanche, F., Sénécal, S., Léger, P. M., Nacke, L., & Fredette, M. (2017). The Evaluation of a Physiological Data Visualization Toolkit for UX Practitioners: Challenges and Opportunities. In *Workshop on Strategies and Best Practices for Designing, Evaluating and Sharing Technical HCI Toolkits (HCI Tools)*.

Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., & Léger, P. M. (2019). Identifying psychophysiological pain points in the online user journey: the case of online grocery. In *Design, User Experience, and Useability. Practice and Case Studies: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part IV 21* (pp. 459-473). Springer International Publishing.

Goldman, A. (2011). Two routes to empathy. *Empathy: Philosophical and psychological perspectives*, 31-44.

Goodman, E. S. (2013). *Delivering design: Performance and materiality in professional interaction design*. University of California, Berkeley

Gray, C. M. (2016, May). " It's More of a Mindset Than a Method" UX Practitioners' Conception of Design Methods. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 4044-4055).

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4.

Hackman, J. R. (2011). Collaborative intelligence: Using teams to solve hard problems San Francisco, CA: Berrett-Koehler Publishers.

Houwer, J. (2006). What are implicit measures and why are we using them? *The handbook of implicit cognition and addiction*, 11-28.

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

Hassenzahl, M. The Thing and I: Understanding the Relationship between User and Product. In Funology: From Useability to Enjoyment; Kluwer Academic Publishers: Dordrecht, Germany, 2003; Volume 3, pp.

Heylighen, A., & Dong, A. (2019). To empathise or not to empathise? Empathy and its limits in design. *Design Studies*, *65*, 107-124.

Hayashi, E. & Hong, J. 2015. Knock x knock: the design and evaluation of a unified authentication management system. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 379–389.

Hoffman, M. L. (2000). Empathy, its arousal, and prosocial functioning. *Empathy and Moral Development*, *2*, 29-62.

Howell, N., Devendorf, L., Tian, R., Vega Galvez, T., Gong, N. W., Poupyrev, I., ... & Ryokai, K. (2016, June). Biosignals as social cues: Ambiguity and emotional interpretation in social displays of skin conductance. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (pp. 865-870).

Hussein, I., Mahmud, M., & Tap, A. O. M. (2014, September). A survey of user experience practice: a point of meet between academic and industry. In *2014 3rd International Conference on User Science and Engineering (i-USEr)* (pp. 62-67). IEEE.

ISO 9241-11:2018. ISO. (2023, April 15). https://www.iso.org/standard/63500.html

ISO DIS 9241-210. Ergonomics of human system interaction - part 210: Human-centred design for interactive systems. Tech. rep., International Organization for Standardization, Switzerland, 2010

Kaasinen, E., Roto, V., Hakulinen, J., Heimonen, T., Jokinen, J. P., Karvonen, H., ... & Turunen, M. (2015). Defining user experience goals to guide the design of industrial systems. *Behaviour & Information Technology*, *34*(10), 976-991.

Kahneman, D., & Knetsch, J. (1993). Strong influences and shallow inferences: An analysis of some anchoring effects. *Unpublished manuscript, University of California, Berkeley*.

Karahanna, E., and Straub, D.W. the psychological origins of perceived usefulness and ease-of-use. Information & Management, 35, 4 (1999), 237–250.

Kaye, J. Evaluating experience-focused HCI. In CHI '07 Extended Abstracts on Human Factors in Computing Systems CHI '07. (2007) ACM, New York, NY, 1661-1664. 19.

Ketola, P., Roto, V. (2008) Exploring User Experience Measurement Needs. 5th COST294-MAUSE Open Workshop on Valid Useful User Experience Measurement (VUUM). Reykjavik, Iceland.

Kim, S. S., Kaplowitz, S., & Johnston, M. V. (2004). The effects of physician empathy on patient satisfaction and compliance. *Evaluation & the health professions*, *27*(3), 237-251.

Khan A, Breslav S, Glueck M, Hornbæk K (2015) Benefits of visualization in the mammography problem. Int J Hum-Comput Stud 83:94–113

Kouprie, M., & Visser, F. S. (2009). A framework for empathy in design: stepping into and out of the user's life. *Journal of Engineering Design*, *20*(5), 437-448.

Kress, G. L. (2012). *The effects of team member intrinsic differences on emergent team dynamics and long-term innovative performance in engineering design teams*. Stanford University.

Kretz, D. R. (2018). Experimentally evaluating bias-reducing visual analytics techniques in intelligence analysis. *Cognitive Biases in Visualizations*, 111-135.

Kretz DR (2015) Strategies to reduce cognitive bias in intelligence analysis: can mild interventions improve analytic judgment? The University of Texas at Dallas

Kumar, A., & Oliver, R. L. (1997). Special session summary cognitive appraisals, consumer emotions, and consumer response. *ACR North American Advances*.

Ho, D. K. L., Ma, J., & Lee, Y. (2011). Empathy@ design research: a phenomenological study on young people experiencing participatory design for social inclusion. *CoDesign*, *7*(2), 95-106.

Lallemand, C., Gronier, G., & Koenig, V. (2015). User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in human behavior*, *43*, 35-48.

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.

Lawson, B. 2006. How designers think: the design process demystified. Architectural Press, Oxford, UK.

Law, E. L. C., Van Schaik, P., & Roto, V. (2014). Attitudes towards user experience (UX) measurement. *International Journal of Human-Computer Studies*, *72*(6), 526-541.

Lederman, R., Wadley, G., Gleeson, J., Bendall, S., and Álvarez-Jiménez, M., 2014. Moderated online social therapy: Designing and evaluating technology for mental health. ACM Transactions on Computer-Human Interaction (TOCHI) 21, 1 (2014), 5.

Léger, P. M., Courtemanche, F., Fredette, M., & Sénécal, S. (2019). A cloud-based lab management and analytics software for triangulated human-centered research. In Information Systems and Neuroscience: NeuroIS Retreat 2018 (pp. 93-99). Springer International Publishing.

Leong, T.W., Vetere, F., & Howard, S. 2012. Experiencing coincidence during digital music listening. ACM Transactions on Computer-Human Interaction (TOCHI) 19, 1 (2012), 6.

Lewinski, P., Fransen, M. L., & Tan, E. S. (2014). Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli. *Journal of Neuroscience, Psychology, and Economics*, *7*(1), 1.

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suarez-Orozco, C. (2018). Journal article reporting standards for qualitative research in psychology: The APA publications and communications board task force report. *American Psychologist*, *73*(1), 26-46.

Liu, F., Kaufman, G., & Dabbish, L. (2019). The effect of expressive biosignals on empathy and closeness for a stigmatized group member. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-17.

Logan, R. J., Augaitis, S., & Renk, T. (1994, October). Design of simplified television remote controls: a case for behavioral and emotional useability. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 38, No. 5, pp. 365-369). Sage CA: Los Angeles, CA: SAGE Publications.

Loiacono, E. T., Watson, R. T., & Goodhue, D. L. (2002). WebQual: A measure of website quality. *Marketing theory and applications*, *13*(3), 432-438.

Lourties, S., Léger, P. M., Sénécal, S., Fredette, M., & Chen, S. L. (2018). Testing the convergent validity of continuous self-perceived measurement systems: an exploratory study. In *HCI in Business, Government, and Organizations: 5th International Conference, HCIBGO 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15- 20, 2018, Proceedings 5* (pp. 132-144). Springer International Publishing.

Maia, C. L. B., & Furtado, E. S. (2016, October). A study about psychophysiological measures in user experience monitoring and evaluation. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems* (pp. 1-9).

Mäkelä, A., Fulton Suri, J. (2001) Supporting Users’ Creativity: Design to Induce Pleasurable Experiences. Proc. of the Int. Conf. on Affective Human Factors Design,

Mandolfo, M., Pavlovic, M., Pillan, M., & Lamberti, L. (2020, July). Ambient UX research: user experience investigation through multimodal quadrangulation. In *International Conference on Human-Computer Interaction* (pp. 305-321). Cham: Springer International Publishing.

Manoogian, J., & Benson, B. (2017). Cognitive bias codex.

Mattelmäki, T., Vaajakallio, K., & Koskinen, I. (2014). What happened to empathic design? *Design issues*, *30*(1), 67-77.

Maslow, A. H. (1954). The instinctoid nature of basic needs. *Journal of personality*

Meehl PE (1954) Clinical versus statistical prediction: a theoretical analysis and a review of the evidence

Micallef L, Dragicevic P, Fekete JD (2012) Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. IEEE Trans Visual Comput Graphics 18(12):2536–2545

Miron-Shatz, T., Stone, A., Kahneman, D.: Memories of yesterday's emotions: does the valence of experience affect the memory-experience gap? Emotion **9**, 885–891 (2009)

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information systems research*, *2*(3), 192-222.

Morris, J. D. (1995). Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research*, *35*(6), 63-68.

Nah F. & Xiao, S. (Eds.): HCIBGO 2018, LNCS 10923, pp. 132-144, 2018. https://doi.org/10.1007/978-3-319-91716-0_1

New, S., & Kimbell, L. (2013, September). Chimps, designers, consultants, and empathy: A "theory of mind" for service design. In n *2nd Cambridge academic design management conference* (pp. 4-5).

Omdahl, B. L. (1995). Cognitive appraisal, emotion, and empathy, ser. *Lecture Notes in Computer Science. New York: Psychology Press*.

Ortiz de Guinea, A. O, and Webster, j. An investigation of information systems use patterns: technological events as triggers, the effects of time, and consequences for performance. ￼ MIS Quarterly, 37, 4 (2013), 1165–1188

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

Ouellette, j.A., and Wood, W. habit and intention in everyday life: the multiple processes by which past behavior predicts future behavior. Psychological Bulletin, 124, 1 (1998), 54–74.

Perrig, S. A., Aeschbach, L. F., Scharowski, N., von Felten, N., Opwis, K., & Brühlmann, F. (2022). Measurement Practices in UX Research: A Systematic Quantitative Literature Review.

Pettersson, I., Lachner, F., Frison, A. K., Riener, A., & Butz, A. (2018, April). A Bermuda triangle? A Review of method application and triangulation in user experience evaluation. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-16).

Phan, M. H., Keebler, J. R., & Chaparro, B. S. (2016). The development and validation of the game user experience satisfaction scale (GUESS). *Human factors*, *58*(8), 1217-1247.

Poels, K., & Dewitte, S. (2006). How to capture the heart? Reviewing 20 years of emotion measurement in advertising. *Journal of Advertising Research*, *46*(1), 18-37.

Polanyi, M. 1966. The tacit dimension. Anchor Books, Garden City, NY.

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.

Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and brain sciences*, *25*(1), 1-20.

Pronin E, Lin DY, Ross L (2002) The bias blind spot: perceptions of bias in self versus others. Pers Soc Psychol Bull 28(3):369–381

Postrel, V., 2002, Positive Psychology: An Introduction. American Psychologist, 55, pp. 5 – 14.

Redelmeier, D.A., Kahneman, D.: Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. Pain **66**, 3–8 (1996)

Riedl, R., & Léger, P.M.: Fundamentals of NeuroIS Studies in Neuroscience, Psychology and Behavioral Economics. Springer, Berlin, Heidelberg (2016)

Rogers, Y. 2004. New theoretical approaches for HCI. Annual review of information science and technology, 38, 1: 87-143.

Roto V., Obrist M., Väänänen-Vainio-Mattila K. (2009) User Experience Evaluation Methods in Academic and Industrial Contexts. Proceedings of UXEM 09 workshop,

Robinson, J., Lanius, C., & Weber, R. (2018). The past, present, and future of UX empirical research. *Communication Design Quarterly Review*, *5*(3), 10-23.

Rosenbaum, S., & Kantner, L. (2008, July). Learning about users when you can't go there: Remote attended useability studies. In *2008 IEEE International Professional communication Conference* (pp. 1-6). IEEE.

Rubin, J. and Chisnell, D. Handbook of useability testing (2nd.edition). Wiley Publishing, 2008.

Saadé, r., and Bahli, B. the impact of cognitive absorption on perceived usefulness and perceived ease of use in on-line learning: An extension of the technology acceptance model. Information & Management, 42, 2 (2005), 317–327.

Scott, B. A., Colquitt, J. A., Paddock, E. L., & Judge, T. A. (2010). A daily investigation of the role of manager empathy on employee well-being. *Organizational Behavior and Human Decision Processes*, *113*(2), 127-140

Sanders, E. B. N. (2002). From user-centered to participatory design approaches. In *Design and the social sciences* (pp. 18-25). CRC Press.

Sauro, J. (2016). The challenges and opportunities of measuring the user experience. *Journal of Useability Studies*, *12*(1), 1-7.

Schooler, J. W., & Eich, E. (2000). Memory for emotional events.

Semmer, N. K., Grebner, S., & Elfering, A. (2003). Beyond self-report: Using observational, physiological, and situation-based measures in research on occupational stress. In *Emotional and physiological processes and positive intervention strategies* (pp. 205-263). Emerald Group Publishing Limited.

Sharma, r.; yetton, P.; and Crawford, j. Estimating the effect of common method variance: the method-method pair technique with an illustration from tAM research.MIS Quarterly, 33, 3 (2009), 473–490.

Simon, H. A. Newell, A., & Shaw, J. C., (1957, February). Empirical explorations of the logic theory machine: a case study in heuristic. In *Papers presented at the February 26-28, 1957, western joint computer conference: Techniques for reliability* (pp. 218-230).

Singer T, Lamm C (2009) The social neuroscience of empathy. Ann NY Acad Sci 1156:81–96

Slovák, P.; Joris Janssen, J.; and Geraldine Fitzpatrick, G. 2012. Understanding heart rate sharing: towards unpacking physiosocial space. In CHI’12. Association for Computing Machinery, New York, NY, USA, 859–868.

Stepanova, E. R., Desnoyers-Stewart, J., Kitson, A., Riecke, B. E., Antle, A. N., El Ali, A., ... & Howell, N. (2023, July). Designing with Biosignals: Challenges, Opportunities, and Future Directions for Integrating Physiological Signals in Human-Computer Interaction. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference* (pp. 101-103).

Stoyanov, S. R., Hides, L., Kavanagh, D. J., Zelenko, O., Tjondronegoro, D., & Mani, M. (2015). Mobile app rating scale: a new tool for assessing the quality of health mobile apps. *JMIR mHealth and uHealth*, *3*(1), e3422.

Straub, D.W., and Burton-jones, A. Veni, vidi, vici: Breaking the tAM logjam. Journal of the Association for Information Systems, 8, 4 (2007), 223–229.

Surma-Aho, A., & Hölttä-Otto, K. (2022). Conceptualization and operationalization of empathy in design research. *Design Studies*, *78*, 101075.

Venkatesh, V. Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. Information Systems Research, 11, 4 (2000), 342–365.

Vermeeren, A. P., Law, E. L. C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010, October). User experience evaluation methods: current state and development needs. In Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries (pp. 521-530).

Visser, F. S., Stappers, P. J., Van der Lugt, R., & Sanders, E. B. (2005). Context mapping: experiences from practice. *CoDesign*, *1*(2), 119-149

Vredenburg, K., Mao, J. Y., Smith, P. W., & Carey, T. (2002, April). A survey of user-centered design practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 471-478).

Weichert, S.; Quint, G.; Bartel, T. Quick guide UX Management: So Verankern Sie Useability und User Experience im Unternehmen; Springer: Wiesbaden, Germany, 2018.

Wiem, M. B. H., & Lachiri, Z. (2017). Emotion classification in arousal valence model using MAHNOB-HCI database. *International Journal of Advanced Computer Science and Applications*, *8*(3).

Wilke, A., & Mata, R. (2012). Cognitive Bias. In: V.S. Ramachandran (ed.) The Encyclopedia of human behavior, 1, 531-535. Academic Press.

Wimmer, B., Wöckl, B., Leitner, M., & Tscheligi, M. (2010, October). Measuring the dynamics of user experience in short interaction sequences. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (pp. 825- 828).

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological review*, *107*(1), 101.

Woo, J. & Lim, Y. 2015. User experience in do-it-yourself-style smart homes. In Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing. ACM, 779–790.

Woolrych, A., Hornbæk, K., Frøkjær, E., & Cockton, G. (2011). Ingredients and meals rather than recipes: A proposal for research that does not treat useability evaluation methods as indivisible wholes. *International Journal of Human-Computer Interaction*, *27*(10), 940-970.

Wright, P., & McCarthy, J. (2008, April). Empathy and experience in HCI. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 637-646).

Zaman, B., & Shrimpton-Smith, T. (2006, October). The FaceReader: Measuring instant fun of use. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles* (pp. 457-460).

Zingoni, M. (2019). Beyond aesthetics, empathy first. *The Design Journal*, *22*(3), 351-370.

Chapter 3

Concurrent Triangulation: Investigating the impact of providing psychophysiological emotional trends to UX practitioners moderating a user test

Pascal Snow, Pierre-Majorique Léger, Sylvain Sénécal
HEC Montréal, Montréal, Canada

Abstract: Integrating psychophysiological analysis into UX evaluation has the potential to enhance the granularity and accuracy of insights derived from a user test. However, because of the technical and time-related constraints involved in processing and interpreting implicit data, UX practitioners are restricted from effectively analyzing this information throughout the user test from which it was collected. Since they are forced to derive insights on implicit measures retrospectively, they miss the opportunity to leverage this information to enhance their understanding of the user test *as it unfolds*. The objective of this study is to evaluate how sources of data (i.e., implicit, and explicit measures) provided to UX practitioners moderating a user test impact their performance and empathy towards users. More specifically, whether equipping moderators with explicit *and* implicit data during a user test, compared to explicit measures alone, has an impact on their ability to identify useability issues, their task prioritization tendencies, and self-perceived empathy towards the user from which the implicit and explicit data came from. To do so, we performed a between-subject experimental design involving 22 participants with professional user experience backgrounds. Results from the experiment suggest that providing UX practitioners moderating a user test with the user's psychophysiological emotional trends alongside self-reported scales has a positive impact on the aforementioned outcomes compared to moderators who exclusively receive the user's self-reported response. Performance wise, moderators who received psychophysiological emotional trends, depicted as valence and arousal aggregated across intervals, were more effective at identifying useability problems and prioritizing aspects of the user test that contained inconsistencies between behaviour and self-perceived response. Furthermore, moderators who received the emotional trends reported higher levels of self-perceived empathy towards the user across two dimensions: emotional interest and emotional sensitivity. This methodological research provides compelling evidence in favor of leveraging implicit data earlier in the UX testing process, rather than relying on it to provide retrospective insights on a completed test. By doing so, this data can enhance immediate understanding throughout the useability test and promote the formation of empathy towards the user, thereby supporting the post-test interview. The limitations, implications, and future research directions are outlined.

Keywords: user testing, user interview, moderator, performance, prioritization, pain points, empathy, psychophysiological trends, emotion, valence, arousal

3.1 Introduction

UX design must find effective ways to merge scientific data-driven approaches with the holistic considerations embedded in the inherently emotional human experience (Forlizzi & Battarbee, 2004). In the continuously evolving landscape of digital product design, the scope of UX has expanded beyond simple useability to encompass affective, experiential, and self-actualization considerations that compound the complexity involved in understanding technology mediated interactions (Jain et al., 2019). Given that interaction contexts underlying user experience are constantly evolving and inherently dynamic (Hazzenzahl, 2008; Law, 2009), there is a continuous push to advance UX methods that support UX designers with understanding users across diverse contextual circumstances (Pine & Gilmore, 2013). As these interaction contexts become increasingly complex, the need for more nuanced approaches to measuring and comprehending said human experience is of critical importance for contemporary UX design teams.

To respond to such challenges, there has been a renewed interest in the methodological approach to UX evaluation, and how improving the theoretical understanding and practical application of UX methods represents a fundamental path forward towards improving UX in the first place

(Alves et al., 2014). Various meta-analyses (Bargas-Avila & Hornbaek, 2011; Vermeeren et al., 2010) have come to similar conclusions: Weak applicability of theoretical contributions to practical settings and a disproportionate tendency towards questionnaires and post-test evaluation. Most recent contributions recommend improving the understanding of data integration and structuring to improve cross-analysis and method triangulation that addresses the multidimensionality of experience (Pettersson et al., 2018). Given that combining methods has been found to contribute to a richer understanding of UX and overall higher scientific quality (Bush, 2012), there have been calls for further research into which methods work together most synergistically and improving the understanding of how various types of data from different sources work together to support UX designers conducting UX evaluations (Pettersson et al., 2018; 108).

Traditionally, the field of UX has disproportionately relied on explicit measures to evaluate experience, typically eliciting direct feedback from surveys or interviews (Ortiz de Guinea et al., 2014; Riedl & Léger, 2016). However, these approaches are acutely vulnerable to recall failures and cognitive biases, hindering their ability to accurately capture the full human experience (Actis-Grosso et al., 2021). Fortunately, new methods are increasingly being used in combination with explicit measures to support a well-rounded understanding of experience (Tams et al., 2014). More specifically, implicit measures that use psychophysiological data collection instruments provide rich representations of emotional experience that support temporally contextualized evaluations of UX that can capture the ebb and flow of emotions throughout an interaction. These measures can pinpoint moments of peak emotional intensity and act as a spotlight for pain points and useability issues (Giroux-Huppé et al., 2019). The concept of experience layering suggests that experience should be interpreted across layers. Using concurrent triangulation to support data-driven experience layering has been found to contribute to more reliable and holistic interpretations of UX phenomena, while also mitigating the biases that are inherent to various UX evaluation methods (Pettersson et al., 2018; Johnson et al., 2007). For instance, concurrently analyzing explicit self-perceived responses, implicit psychophysiological data, and behavioural observation to derive a more effective qualitative exploration throughout posttest interviews.

While extremely informative in terms of the depth and sensitivity of measurement, physiological measurement tools also come with drawbacks. They typically generate enormous raw data files that are cumbersome to process and require extensive time and expertise to effectively interpret. The challenges associated with analyzing this genre of data have given rise to various SaaS products aimed at their integration into 'traditional' UX toolkits. iMotions (Copenhagen, Denmark) and Noldus (Wageningen, Netherlands), pioneers in the behavioural and biometric UX research market, have introduced programs intended to reduce the barriers to entry involved in using this technology. Building upon existing product offerings that visually summarize user tests, these companies have recently introduced live representation across their suite of products. iMotions Lab (iMotions, Copenhagen, Denmark) has recently rolled out their Lab Streaming Layer aimed at real-time importation of external sensor data to represent the raw data being gathered at any given time. This highly specific and sensitive representation framework can be challenging to interpret given the dynamic fluctuation of psychophysiological data during a user test. In other words, while information is dense and highly accurate, it may not necessarily be the best approach to data representation given UX evaluation objectives during a user test. Alternatively, aggregating raw data across temporal intervals may provide more easily interpretable representations of experience that support ongoing inference.

In addition to the data-driven specificity that implicit measures afford UX evaluation, it also provides benefits to various other dimensions of user testing. For instance, studies have demonstrated the link between shared implicit measures and the formation of empathy and pro-social behaviour between individuals (Stepanova et al., 2023). As one of the primary drivers of design thinking, empathy represents a fundamental quality when developing UX design methodologies focused on understanding the user (Brown et al., 2010). Since empathy towards users is seen to produce more relevant design solutions (Simons et al., 2011), this may have great benefit for moderators in the context of user testing. Thus, promoting empathy may be a peripheral benefit arising from integrating psychophysiological inquiry into user testing.

To our knowledge, no study has evaluated how physiological measurement tools can be used as part of a concurrent triangulation approach to user testing to identify useability issues and more effectively prioritize tasks in an interview following a user test. This study proposes redefining the predominant framework used to implement psychophysiological measurement tools into user testing. Rather than leveraging them to produce retrospective insights from a completed user testing session, this study explores the effectiveness of a swifter implementation of live psychophysiological representation. Hence, analysis of this data is carried out throughout a user test from which they are collected, thereby providing insights that support moderators with developing evidence-based strategies for the subsequent user interview. By segmenting emotional trends by individual tasks, it provides a more clear-cut narrative structure when it comes time to interpret the respective components of the completed user test. Considering this unexplored framework and format of psychophysiological representation and implementation into a user test from which it has been collected, we have established the following central research questions:

RQ1: *To what extent does providing UX practitioners with a visual representation of the user's psychophysiological trends impact the practitioner's performance outcomes while moderating a user test?*

In addition to their outcomes while moderating a user test, another area that we seek to explore is how providing implicit measures in addition to explicit measures influences perceived empathy towards the user. Given that there is evidence that empathy can be stimulated by more distinct representations of emotion, specifically through biosignals, we seek to understand whether this phenomenon can be extended to technology-mediated representations of emotion displayed through aggregated data trends in the context of a user testing session. This peripheral exploration segues into the second research question:

RQ2: *To what extent does providing UX practitioners with a visual representation of the user's psychophysiological trends impact the practitioner's perceived empathy towards that user while moderating a user test?*

Through a one-factor within-subject experimental design that is meant to simulate the context of a typical user test, this study intends to measure how implicit measures influence a moderator's inferential performance in terms of identifying useability issues and their prioritization tendencies during a posttest interview. The following chapter will outline the fundamental principles related to this field of research while emphasizing the gaps in existing literature. Subsequently, it will provide the specifics on all relevant aspects of the experimental design which informs the finding of this research. Finally, it will discuss the results, followed by implications of this research in terms of its practical and academic contributions, limitations, and future directions for this branch of methodological research.

3.2 Research Framework and Hypotheses

This brief preliminary review will amalgamate existing research that specifically pertains to the impact of implicit measures on user testing and empathy. It aims to provide peripheral evidence that sets the stage for the inferred research hypotheses of this study. By doing so, it provides context and justification for the relevance of this research in the current UX climate across research and professional practice.

3.2.1 Performance outcomes: inference and prioritization

UX practitioners engaging in a user test must make the most of their inherently limited time with users. Given this finite access to users, especially during a single user testing session, it becomes increasingly important to make the best use of time by prioritizing high-value aspects of the user test. When moderators have a high degree of understanding of the user test, they are better equipped to identify useability issues and prioritize aspects of the product that caused difficulties to the user. Similarly, when moderators engage in concurrent triangulation, they can cross-analyze various sources of data to identify inconsistencies in user reporting that should be further prioritized in the context of the posttest interview. Thus, identifying useability issues and understanding how moderators prioritize aspects of the user test is of vital importance.

Difficulties experienced by users in the context of a user test are called pain points. They represent some point of conflict or discomfort for the user throughout the interaction with the product or service (Platzer, 2018). Given the complexity and situatedness of pain points as they relate to a particular individual experiencing them, it has been proposed that inductive and open-ended ethnographic-based approaches can be one of the most effective approaches to uncovering “their knotty ambiguity” (Platzer, 2018). This line of thinking reaffirms that generative and qualitative forms of inquiry are most effective at uncovering user pain points compared to surveys (Platzer, 2018; Wang et al., 2016). However, more recent studies have not only found that quantitatively derived implicit measures are extremely effective at identifying pain points but have also provided empirical evidence of their relative superiority (Giroux-Hubbé et al., 2019). Giroux-Hubbé’s research (2019), in combination with existing studies (Ortiz de Guinea et al, 2013), finds that retrospective explicit measures lack temporal precision and are influenced by various cognitive biases. Even if users report their experience immediately following each task, studies prove major discrepancies between the lived emotional experience and what was recalled by the user themself (Cockburn et al., 2017; Eich & Schooler 2000). For instance, peak effects cause users to disproportionately anchor themselves to isolated moments of the user test characterized by moments of high emotion (Cockburn et al., 2017; Ariely, 1998), while interactions causing negative emotions are more frequently recalled than positive ones (Baumeister et al., 2001). Leading to the conclusion that relying exclusively on explicit measures will lead to inaccurate UX insights (Ortiz de guinea et al., 2013).

Useability issue detection in real-world UX practice predominantly relies on behavioural observation that is complemented with explicit measures such as interviews and self-reported questionnaires (Giroux-Hubbé et al., 2019; Mucz & Gareau, 2018). However, considering the growing demand for data-driven recommendations (Georges et al., 2017), implicit measures propose an empirically robust way to synergistically work alongside traditional measures and the qualitative richness of interviews. Prior studies that investigate detecting useability issues through implicit measures put the locus of inquiry on the user, investigating the extent to which they can retrospectively recall or perceive their own pain points. Conversely, this experiment situates UX moderators engaging in concurrent triangulation of explicit *and* implicit measures as the central area of inquiry, a logical focus given who wields them in the first place. Considering the advantageousness of having implicit data to complement behavioural observation and explicitly derived questionnaires, it is hypothesized that:

H1: UX Practitioners provided with the user's psychophysiological emotional trends [implicit data] alongside self-perceived responses **[**explicit data] while moderating a user test **will have better performance outcomes** compared to UX practitioners provided with only the user's self-perceived response.

Performance outcomes are assessed based on the following two dimensions:

H1a: UX Practitioners provided with the user's psychophysiological emotional trends [implicit data] alongside self-perceived responses [explicit data] while moderating a user test **will be more accurate at inferring the occurrence of useability problems** compared to UX practitioners provided with only the user's self-perceived response.

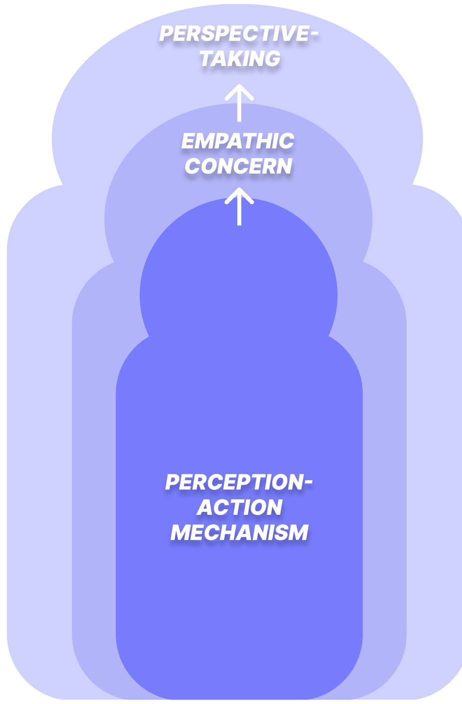
H1b: UX Practitioners provided with the user's psychophysiological emotional trends [implicit data] alongside self-perceived responses [explicit data] while moderating a user test **will more frequently prioritize aspects of the user test that contain inconsistencies between user behaviour and self-perceived response** compared to UX practitioners provided with only the user's self-perceived response.

3.2.2 Perceived empathy towards users

Empathy represents one of the most fundamental qualities of designers working on complex digital products, since it supports their ability to not only understand, but anticipate the wishes, needs, and requirements of potential users and subsequently reflect them throughout product development (Wright and McCarthy, 2008; Drouet et al., 2022; Surma-Aho and Hölttä-Otto, 2022). Empathy can be defined as an individual's capacity to understand the innate state of another (i.e., cognitive empathy), but also as the merging of affective states between two people (i.e., emotional empathy) (De Vigemont & Singer, 2006; Escalas & Stern, 2003). Empathy is an essential step in the Design Thinking methodology, stimulating the creation of innovative ideas, and subsequently ensuring their relevance and linkage to consumers (Calgren et al., 2016).

One of the most compelling streams of empathy building research is related to the positive effect that biological cues have on evoking perceived empathy (Salminen et al., 2019). The Russian-doll model of empathy, otherwise referred to as the perception-action model (PAM), provides an explanation of how higher level cognitive and emotional empathy are born out of an instinctual and biologically driven emotional contagion. Preston & De Waal (2002) suggest that empathic tendencies are born from the perception-action mechanism (PAM). At its evolutionary roots, the perception-action mechanism initiates basic expressions such as emotional contagion and motor mimicry, inducing similar emotional states between individuals interacting. In other words, when an observer is provided access to the neural or bodily representation of someone's emotional expression, the observer's neural representation of that state is automatically activated. A graphic depicting this interrelated model is presented below:

Figure 1: Perception-action model (PAM) of empathy



Simply put, being exposed to expression of emotion leads to emotional contagion and motor mimicry of heart rate, perspiration, facial expressions, and bodily posture (De Waal, 2007), which can then contribute or lead to higher levels of empathic responses. Research has shown this synchronous relationship of physiological responses between individuals in social settings (Surma-aho et al., 2019; Soto & Levenson, 2009). For instance, the synchronization of EDA between physicians and patients has been related to perceived empathy of the therapist (Marci et al., 2007). Indeed, on a biological level, empathy has been found to be associated with changes in central nervous system activity that indicate various forms of biological mirroring (Zaki, 2012). This more fundamental mechanism is said to act as the foundation to outer layers such as emotion regulation, self-other distinction, and cognition which are considered more evolved expressions of empathy, but nonetheless built upon the core socio-affective biological basis. This includes perspective taking, otherwise known as the deliberate attempt at taking another's point of view, and widely considered as a core dimension of empathic design (Surma-aho et al., 2019). Thus, empathy is simultaneously a quotidian social interaction we have with those around us, and an embedded neurophysiological capacity that is built into human evolution.

As recently as 2021, researchers have made claims that designers have no common vocabulary to describe empathy tools (Pratte et al., 2021). Pratte developed a framework for empathy tools across three dimensions: (1) the amount of agency it affords to the user using that empathy tool, (2) the user's perspective while using the tool, (3) the sensations that are experienced (i.e., using glasses that blur your vision). Prior studies have explored tools intended to support designers with the evocation of empathic responses towards users. The most common category relates to conceptual approaches such as empathy maps or personas (Marsden & Wittwer, 2022). These pseudo-replications are meant to automatically trigger cognitive empathic responses that are related to perspective-taking (Haag & Marsden, 2019). Other studies replicate conditions more tangibly; using physical items intended to replicate impairments or disability (Pratte et al., 2021). Others have tried to mediate empathic responses through *empathy machines*, which can be understood as embodying the experience of another through technology which is typically done through digital renditions or avatars that express an embodied emotion (Bollmer, 2017). While this approach to empathy building has been referred to as "the ultimate empathy machine" (Milk, 2015), this form of empathy is unapplicable to moderated user testing, where the moderator must remain relatively 'present' throughout the interaction. As pointed out by Marsden (2022), much of this research neglects the inherent social context of design teams and the user-designer relationship. Thus, there is a need for less disruptive, more subtle empathy tools that can be easily integrated into design scenarios that involve user-moderator interaction.

Existing research has provided theoretical explanations and empirically demonstrated the evolutionarily based mimicry (i.e., emotional contagion, mirroring, etc.) that occurs during human interactions, especially when mediated by visual cues that illustrate affect. Others have shown the positive effect that social biofeedback cues have in the context of interpersonal empathy processes in a dyadic technology-mediated setting (Salminen et al., 2019). While the PAM response has been proven to apply to technology-mediated displays of more graphic biosignals, it would be valuable to investigate whether psychophysiological emotional trends represented as valence and arousal are a sufficiently compelling representation format. Furthermore, there has yet to be research analyzing the effect of this mechanism on cognitive and emotional empathy in the context of the user-moderator relationship context. While many empathy scales exist, none of them had been geared towards the unique context of directed empathy in the context of product design. Fortunately, a recent scale was generated to provide insights into this very relationship. The newly developed Empathy in Design Scale is designed to measure this directed, design-oriented conception of perceived empathy (Drouet et al., 2022). This represents a more deliberate attempt to measure empathic disposition in empathic design contexts. The opportune introduction of the Empathy in Design Scale (Drouet, 2022) will support this study's investigation on these topics, providing a relevant measurement framework to assess these research objectives. Considering the pre-established research and subsequent gaps, we have developed the following replication hypotheses:

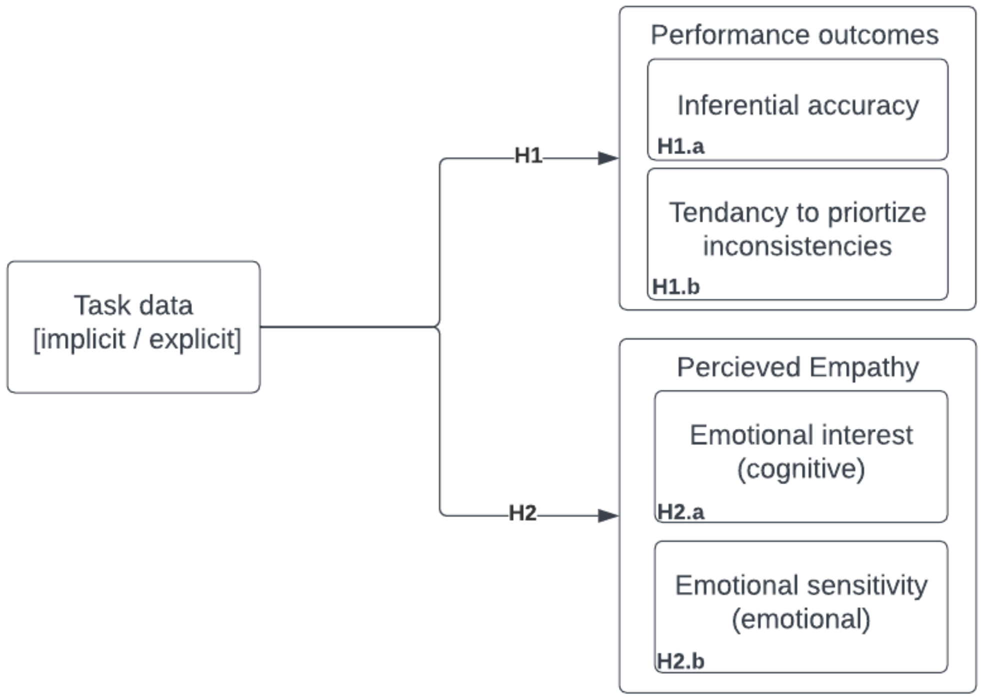
H2: UX practitioners provided with the user's psychophysiological emotional trends [implicit data] alongside self-perceived responses [explicit data] while moderating a user test will **feel more empathic towards that user** compared to UX practitioners provided with only the user's self-perceived response.

Empathy is measured across the following two dimensions:

H2a: UX practitioners provided with the user's psychophysiological emotional trends [implicit data] alongside self-perceived responses [explicit data] while moderating a user test **will feel more emotional interest [cognitive empathy] towards that user** compared to UX practitioners provided with only the user's self-perceived response.

H2b: UX practitioners provided with the user's psychophysiological emotional trends [implicit data] alongside self-perceived responses [explicit data] while moderating a user test **will feel more emotional sensitivity [emotional empathy] towards that user** compared to UX practitioners provided with only the user's self-perceived response.

Figure 2: Research model variables



3.3 Method

3.3.1 Experimental Design

A one-factor between-subject laboratory experiment was conducted to test these hypotheses. The aim is to assess how sources of data provided to moderators in the context of a user test impacts performance outcomes and perceived empathy towards the user.

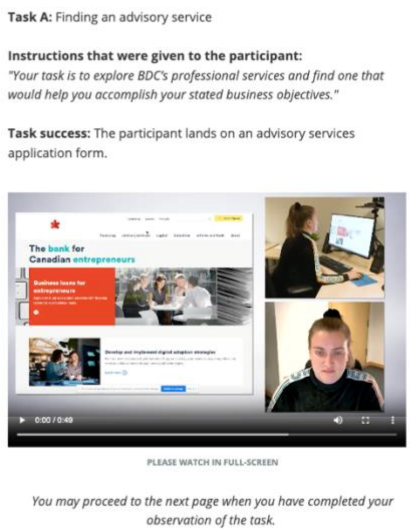
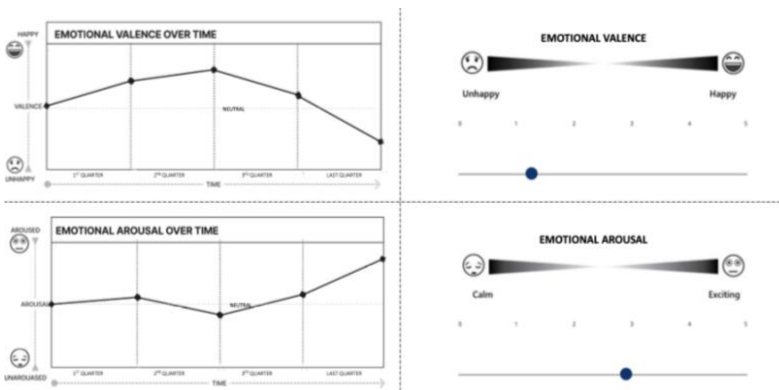
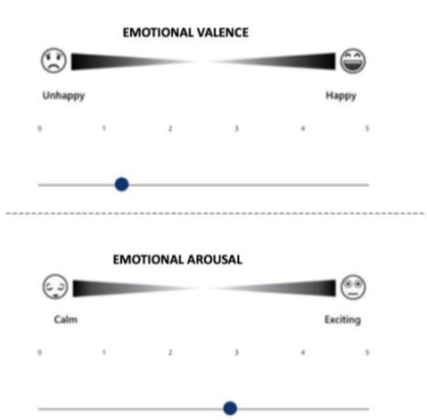
The experiment was built around a simulated user test, where the participant observes and responds to a fictitious user portrayed through pre-recorded user journeys. An actor played the role of this fictitious user by enacting 4 pre-determined user flows, where each segment of the user test was representative of a typical task: A directed objective that the user would complete in the context of a user test. The test artefact that the fictitious user was interacting with was a publicly accessible Canadian financial company. Each of the four tasks were representative of a user flow that a typical user might undergo on the test artefact in a real-world scenario. These user journeys were determined by drafting flowcharts for the website's most popular features and extracting 4 distinct flows (see Table 1 below) that varied in complexity and useability. They were edited into video collages that provided various perspectives. More specifically, the fictitious user's behavioural demeanor showed a front facial profile and side body language profile that was played synchronously with a screen recording of the digital interaction. Table 1 provides an overview of the 4 user journeys.

Table 1: Overview of user flows

Task	Description
A: Finding relevant advisory service	The user is expected to find advisory services related to their business objectives.
B: Finding relevant loan	The user is expected to search for a loan based on their business' characteristics and priorities.
C: Finding relevant learning resources	The user is expected to find pertinent information from the resource directory.
D: Finding relevant support services	The user is expected to find a community initiative that caters to their demographic profile.

Table 2 contains an example of the stimuli provided to participants. While all participants received the same behavioural observation stimuli, the user response data provided to participants differed based on condition. The image adjacent to condition A and B shows an example of the user response stimuli provided to participants after they finished watching the user flow [behavioural observation]. These are provided as a single sheet of paper contained in a cardboard folder labeled with its designated task. Images of these stimuli are provided in Table 2 below:

Table 2: Stimuli provided to participants throughout simulation

Category	Example of Stimuli provided – Task A	Description
Behavioural observation [n = 22]		***behavioural observation*** *provided to all participants as a full-screen video montage*
Condition A [n = 11]		***implicit (left) and explicit (right)*** *user response provided to participants in condition B printed on single paper and contained in folder*
Condition B [n = 11]		***explicit*** *stimuli provided to participants in condition B printed on single paper and contained in folder*

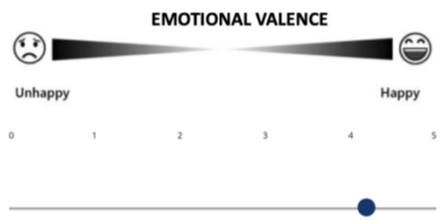
Note: See Appendix for exact copy of stimuli provided to participants

It is important to explain the framework underlying the design of these four tasks. We tried to emulate a common phenomenon in user testing when creating these four distinct emotional patterns. We organized the four tasks across two dimensions: their **emotional consistency** across the experience, and the **ending emotion** of the task (see Table 3). Therefore, we had two emotionally consistent experiences, with one being positive [Task C] and one negative [Task B], in addition to having two emotionally inconsistent experiences where one ended positively [Task D] and the other ended negatively [Task B]. Two of these tasks were considered consistent: the user's perceived experience was consistent with their behaviour. For example, they clearly executed the task with no difficulties, self-perceived their experience as positive, and their implicitly measured emotional experience was consistently positive, with the inverse also being true. However, the other two tasks were meant to emulate more inconsistent experiences. For example, one task clearly depicted the user having difficulty during the first half of the task. However, the second half of the task was carried out with relative ease. The user reported this task as positive. Inversely, one task depicted the user carrying out the task easily in the first half of the user journey, but the ending posed problems for the fictitious user. In this case, the user self-reported the experience as negative.

While there are many biases that could justify inconsistencies between implicit emotional responses and explicitly reported accounts of experience, we leaned into emulating recency bias to provide a single consistent explanatory mechanism. Recency bias is when users anchor themselves to the most recent memory of the experience, thus disproportionately attributing value to the end of the task or experience (Zhou, 2020). Thus, you can observe in Table 4 below that the user rated their entire experience based on a positive or negative ending sentiment. This is particularly notable for 'Task A' and 'Task D' because the psychophysiological emotional trend follows a divergent pattern, and therefore not fully reflected by the fictitious user's self-reported [explicit] response that only reflects the most recent part of the task.

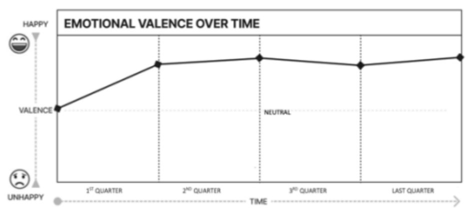
Table 3: Implicit and explicit valence structure across user flows

<table>
<tr><th></th><th>consistent
psychophysiological trend</th><th>divergent
psychophysiological trend</th></tr>
<tr><td>Psychophysiological trend ends with positive valence
∴
positive self-reported valence</td>
<td>Task C: explicit

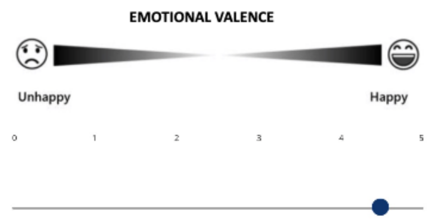


Represents positive emotion only

Task C: implicit

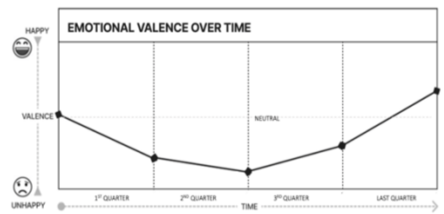


Represents consistent positive emotion</td>
<td>Task D: explicit

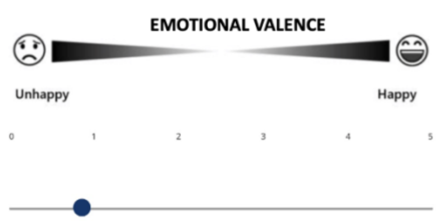


Represents positive emotion only

Task D: implicit

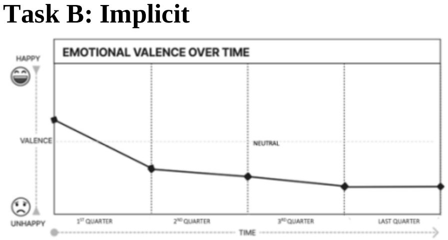


Emotion starts negative and ends positive</td></tr>
<tr><td>Psychophysiological trend ends with negative valence
∴
Negative self-reported valence</td>
<td>Task B: explicit

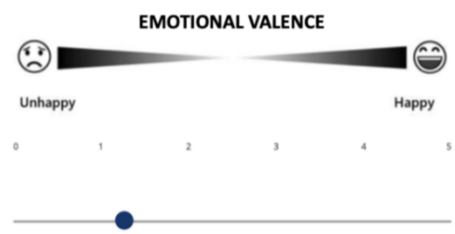


Represents negative emotion only

Task B: Implicit

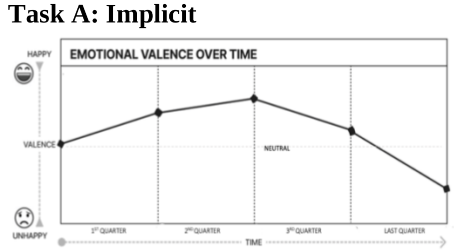


Represents consistent negative experience</td>
<td>Task A: explicit



Represents negative emotion only

Task A: Implicit



Emotion starts positive and ends negative</td></tr>
</table>

3.3.2 Participants

Participants were primarily recruited through professional networks and LinkedIn. The inclusion criteria for participation stated that they required professional working experience in the field of UX. In total, 22 participants with varying degrees of UX expertise were recruited to participate in the experiment: 11 men and 11 women (Table 4). Their professional experience ranged from less than 1 year to a maximum of 11 years of industry experience, with a mean of 3.5 years of experience and standard deviation of 2.7. The sample's mean age was about 31 years old. The mean quantity of user tests and interviews conducted by the participants throughout their past professional experience was approximately 49 sessions according to their own estimated self-assessment. Two out of the 22 total participants were not familiar with the concept of psychophysiological measurement instruments, and these two participants were allocated to condition B, which did not incorporate implicit measures during the simulation. Each participant was given a small compensation for their time, in addition to covering expenses related to transportation. The approval of the research ethics board was granted for this study (Certificate #202X-XXXX) and everyone provided their consent prior to commencing participation. Table 3 provides an overview of the participant's demographic profile and work experience.

Table 4: Demographic variables

Gender			**N=22**	**Age**	**N = 22**
	Man	50%	[11]	Mean	30.8
	Woman	50%	[11]	Median	29
	non-binary/other	N/A		Mode	30
				SD σ	5.5
				Minimum	24
				Maximum	49

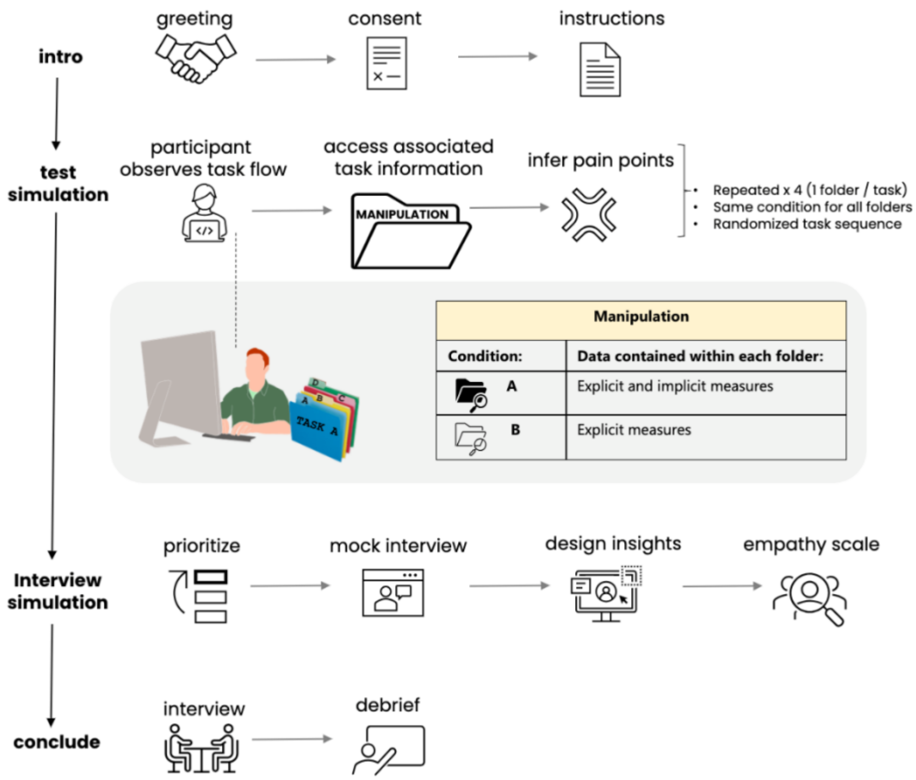
Years of Experience			
	1-2 years	40.91%	[9]
	2.01-4 years	31.82%	[7]
	4.01-6 years	13.64%	[3]
	> 6 years	13.64%	[3]

3.3.3 Experimental Procedure

The experiment design was meant to replicate the conditions of a remote user test by creating an online simulation that mimicked a real-world user test procedure. The participants of the experiment were instructed to assume the role of test moderator. In other words, they were asked to approach the simulation experiment as if they were a UX professional moderating the user test, and instructed to make decisions throughout the simulation as if they were a UX professional conducting a user test with a typical user. The only instruction that they were provided was the context of the hypothetical user test simulation and their primary objective: to gather the best UX insights from the user test consisting of four tasks and a subsequent interview.

The participant was greeted at Tech3lab and brought to the laboratory room where the experiment was being held in a laboratory setting. The participant was given a consent form which must be completed prior to participation. They are given time to read it at their discretion and ask clarifying questions. The experiment was built upon the Qualtrics platform, using its infrastructure to display instructions, media, and survey questions to the participants. The primary objective underlying the experiment was to put the participants in an engaging simulation that was meant to emulate a user testing session. Thus, the experiment followed a typical user test procedure beginning with the fictitious user completing pre-determined tasks. The duration of these videos varied between one and three minutes. After each video was played, the participant was prompted to open a physical cardboard folder that contained the user’s response to the task. This is where the manipulation occurred. For Condition A, the folders contained the fictitious user’s self-reported affective slider response [explicit measure] *and* a visual representation of the fictitious user's psychophysiological emotional trend in terms of valence and arousal over time [implicit measure]. Participants in Condition B received folders that only contained the fictitious user's self-reported affective slider response [explicit measure]. The tasks appeared in a randomized sequence through Qualtrics' randomizer function. While the order of the videos was randomized, each video had a corresponding fictitious response that was linked to that task. Therefore, no matter the order that videos were displayed, each task had a logically related set of data that was functionally related to its respective task. This was repeated four times in a row for each participant, where one folder was prompted to be opened after observing each task. Following the observation of a task and receiving the user response data from the folder, the participant is asked to infer whether the user had difficulties throughout the task and at what point in the user flow. These pain points were conveyed within the video and were consistent with the physiological information provided. Once the participant has viewed the 4 tasks, they are asked to prioritize the 4 tasks according to which one they think is most important to further explore in the context of an interview. They are then asked to produce two hypothetical open-ended questions that they might ask in an interview, followed by 2 design recommendations. For both questions, they are asked to specify which tasks they apply most to. The participant is prompted to complete the Empathy in Design Scale (Drouet, 2022). Finally, a short interview is held between the participant and primary researcher. The participant is debriefed; the independent variable is revealed, and their specific condition is explained and juxtaposed with the alternative condition. The figure below depicts the experimental procedure and highlights when exactly the experimental manipulation takes place.

Figure 3: Visual diagram of experimental procedure



3.3.4 Measurement Context

Many of the scales used in UX research and practice are geared towards users. For example, the SUS scale, which is used to measure the overall useability of a website, is conventionally filled out *by* users themselves and then interpreted *by* UX professionals. However, in this study, we were concerned with the moderator's performance, not the actual artefact of the user test. Thus, we were obligated to create our own data presented as stimuli to participants for this experiment. The fictitious user response data that was provided to participants was based on valence and arousal; two common implicit measurement dimensions used to display physiological measures taken from a user test. However, this data was contrived so that it could adhere to the structural requirements needed to delineate 4 distinct emotional trends while ensuring consistent experimental conditions. We represented valence and arousal across implicit and explicit measures: Explicit measures were represented via Affective Slider (AS) (Betella & Verschure, 2016), and implicit measures were represented through psychophysiological emotional trends that simulated psychophysiological data aggregated across four intervals. While each task varied slightly in overall duration, each of them was aggregated across relative quarterly timestamps. In theory, this data could have been collected by triangulating various physiological measures using the COBALT BlueBox device (Léger et al., 2019) to record physiological response. This device measures electrocardiography (ECG) and electrodermal activity (EDA), otherwise referred to as galvanic skin response. This combination of ECG and EDA levels provide a proxy for arousal, while FaceReader (Noldus, Wageningen, Netherlands) captures valence through facial expressions synchronized with dynamic stimulus.

For the rest of this section, we will discuss the framework in place to measure the dependent variables that were used to test the previously established hypotheses. The statistical approach for this experiment relies on the ground truth that was established as the experimental simulation was being designed. Given that establishing these parameters was necessary to assess the performance outcomes of the simulation, we created two schemas to score inferential accuracy and the value of their prioritization tendencies. These schemas will be outlined below. Finally, we adapted the Empathy in Design Scale (Drouet, 2022) to create a streamlined version to minimize the cognitive demands on the participant, while ensuring that we were focusing on the most pertinent and relevant questions that applied to the context of the simulation. An outline of the measurement framework for dependent variables is outlined below:

Table 5: Variable measurement framework

Dependent Variable	Item	Format	Source
Inferential accuracy	Which statement best describes the user's experience with [task]?	4-point nominal scale: No difficulty Difficulty at beginning Difficulty at end Difficulty throughout	Self-developed
Prioritization tendency	Drag and drop the four tasks in priority order. Produce two task-directed open-ended interview questions and specify which tasks they apply to. Produce two design recommendations and specify which tasks they apply to.	4-point Likert scale from "Top priority" to "Last priority" Open-ended section followed by drop-down menu specifying which task they apply to.	Self-developed
Emotional Interest [cognitive]	I am interested in learning about users' experiences and needs. I imagine how users think, feel, or behave in different situations. I am curious about users' experiences and needs. I want to learn about users' experiences and opinions about the service.	7-point Likert scale "Does not describe me at all" to "Completely describes me"	*Empathy in Design Scale* (Drouet et al., 2022)
Emotional Sensitivity [emotional]	I am sensitive to the experiences of users. I observe without judging how users experience the service. When thinking about the service, I take the users' point of reference. I am concerned about the experiences of users	7-point Likert scale "Does not describe me at all" to "Completely describes me"	*Empathy in Design Scale* (Drouet et al., 2022)

3.3.4.1 Inferential accuracy

We used a 4-point nominal scale to measure the participant's accuracy at inferring moments where the user had difficulties completing a task (Table 5 above). They made an inference for each task, for a total of 4 inferences. Their accuracy is based on the number of correct inferences made as a function of the ground truth that was established for the simulation. For their inference to be considered correct, they had to not only identify the occurrence of difficulty, but also its temporality. We assessed inferential accuracy through a one-tailed Wilcoxon rank sum test to derive precise p-values. These were subsequently used to assess whether there was a statistically significant difference between the rate of correct inferences across conditions.

3.3.4.2 Prioritization tendency

We measured prioritization tendencies by operationalizing it across the following three items:

Table 6: Operationalization of prioritization tendency

Construct	Operational Definition
Ranked priority	Sum of Value Index (see table 7 below) based on tasks prioritized **in any top two position**.
Open ended interview questions	Sum of Value Index based on which tasks their two open-ended interview questions applied to.
Design recommendations	Sum of Value Index based on which tasks their two design recommendations applied to.

We determined the value of prioritizing a given task based on (1) the extent to which the self-reported responses differed from the user's behaviour during that task, and (2) whether the self-reported response was representative of useability issues that were encountered during the user flow. Table 7 below illustrates the scores that were given to participants based on how they prioritized tasks across the 3 prioritization constructs in Table 6:

Table 7: Prioritization - Value Index

Task	Implicit valence structure	Explicit valence	value index
A	NEUTRAL ⟶ HIGH ⟶ LOW	LOW	1
B	NEUTRAL ⟶ LOW ⟶ LOW	LOW	0
C	NEUTRAL ⟶ HIGH ⟶ HIGH	HIGH	0
D	NEUTRAL ⟶ LOW ⟶ HIGH	HIGH	2

Table 7 provides a simplified overview of how valence is depicted across the four tasks. Implicit measures showed visual trends over time, whereas explicit measures provide a single rating that is meant to represent their self-reported assessment of the entire task (see table 2). The value index represents the score that was given to participants if they (1) rated the task in the top two highest prioritized tasks when asked to rank them in priority sequence, (2) asked open-ended interview questions, or (3) generated design recommendations that applied to the task.

We establish that task D represents the highest hidden value because it is not accurately represented by the fictitious user's self-reported [explicit] measure. As shown in table 8 above, the implicit response for task D was measured as 'NEUTRAL->LOW->HIGH', however it was self-reported as 'HIGH'. Thus, this task was considered as the most valuable to prioritize because the self-reported response was both non-representative of behaviour and it concealed useability problems encountered throughout the beginning and middle of the user flow when valence was recorded as 'LOW.' This task is meant to represent a task where the user inadvertently concealed difficulty they experienced in the middle of the task. In line with the simulation narrative, this occurred because the task ended positively, and therefore the user projects this culminating sentiment onto the entire task. Similarly, we established that Task A has the second highest value because it was non-representative of behaviour, however it did not conceal useability problems. Since it is reported as negative, and thus does not necessarily conceal a useability problem, we attribute a smaller hidden value to this task. Since this study focuses on concurrent triangulation to identify biased and inconsistent reporting, we attribute less value to consistently negative experiences, since this does not reflect critical evaluation that led to moderators identifying inconsistencies between reporting and behavioural observation.

For example, assume the participant had prioritized the 4 tasks in this order: B, A, D, C. They would have accrued a single point for the 'ranked priority' construct, since only A is in a top 2 position, and it represents a value index of 1. Open ended interview questions and design recommendations both required two answer submissions, so they would have been awarded the number of points associated with the tasks they decided to focus on. As an additional example, if they proposed design recommendations for task A and D, they would have accrued 3 points. We assessed prioritization based on mean scores accrued by item and across conditions. We used a one-tailed Wilcoxon rank sum test to derive precise p-values that were subsequently used to assess whether there was a statistically significant difference in prioritization tendencies.

3.3.4.3 Empathy

We adapted the Empathy in Design Scale to measure perceived empathy towards the fictitious user depicted during the simulated user testing session (Drouet et al., 2022). The scale is organized across four dimensions: Emotional interest (EI), sensitivity (S), personal experience, and self-awareness. We excluded the latter two dimensions because the questions were less relevant to the context of a user test, and we believed using emotional interest and sensitivity dimensions were sufficient given that they are designed to measure cognitive empathy and emotional empathy, respectively. In line with the established scale, emotional interest was measured across four individual questions (see Table 5 above). Emotional sensitivity is measured across six items, but one was omitted because it was not applicable to the simulation's context. We believed omitting this item was further justified by the fact that Drouet's publication (2022) explicitly states that this question is ranked at the bottom in terms of the understandability of scores that were established throughout the construct validity phase of scale development. The participant's responses were recorded using a 7-point Likert scale. We measured overall empathy by finding the mean empathy score across both dimensions. In terms of internal consistency in reporting, the Cronbach's Alpha for emotional interest was 0.75 and sensitivity was 0.74. This demonstrates a high degree of reliability, which is especially important for the sensitivity dimension which was minorly adjusted.

We assessed the distribution of empathy scores by using the Wilcoxon rank sum test to derive precise p-values used to assess whether the distribution between conditions was statistically significant.

3.4 Results

3.4.1 Inferential accuracy (H1a)

As a subset of performance outcomes, this study aimed to assess whether a moderator's ability to infer useability issues could be enhanced with the addition of psychophysiological trends provided during a user test. The following tables summarize the inferences made by participants:

Table 8: Condition A inference distribution – Implicit and explicit measures

	Task A	Task B	Task C	Task D
Beginning	0	2	2	5
End	6	0	1	0
None	0	0	8	1
Throughout	5	9	0	5

Note: Green cells represent the correct inference

Table 9: Condition B inference distribution – Explicit measures

	Task A	Task B	Task C	Task D
Beginning	0	7	1	7
End	1	0	0	1
None	0	0	9	2
Throughout	10	4	1	1

Table 10: Mean inferential accuracy

Condition	N	Correct inferences	Mean (1 = always correct)	Std Dev
A	44	28	0.636	0.487
B	44	21	0.477	0.505

Table 11: Inferential accuracy result

Variable	Statistic (S)	Z-score	*p* Value
Correct inference	2112	1.4887	0.0988*

One-tailed Wilcoxon rank sum test results suggest that participants in Condition A (i.e., psychophysiological trends and self-reported measures) had higher inferential accuracy of difficulties experienced during the interactions than condition B at a significance level of 0.1 (p = 0.0988). Participants who received psychophysiological trends correctly inferred the occurrence of pain points approximately 34% more often. **These results marginally support H1a**, which proposes that moderators who received psychophysiological emotional trends alongside self-perceived responses will more accurately infer the occurrence of useability problems compared to moderators provided with the user's self-perceived response only.

3.4.2 Prioritization tendency (H1b)

As a secondary dimension of performance outcomes, we wanted to better understand how moderators might adjust their approach to prioritizing aspects of the user test depending on what information they are provided. The following table outlines the constructs used to determine prioritization tendency, as well as their mean score as a function of how frequently they prioritized tasks that contained divergent valence structures.

Table 12: Mean relative value based on prioritization tendency

Condition	Observations	Construct	Mean
A	N = 11	Ranked priority (top 2)	0.455
		Open ended interview questions	1.545
		Design recommendations	1.455
B	N = 11	Ranked priority (top 2)	0
		Open ended interview questions	1.091
		Design recommendations	0.9

Table 13: Prioritization tendency result

Construct	Statistic (S)	Z-score	p value
Ranked priority (top 2)	154	2.4401	0.0175**
Open ended interview questions	145.5	1.5513	0.0986*
Design recommendations	88.5	-1.6547	0.0577*

A one-tailed Wilcoxon rank sum test results suggest that participants in Condition A (i.e., psychophysiological trends and self-reported measures) were more likely to prioritize topics that are considered to have high concealed value. Our findings suggest that when participants were asked to broadly prioritize topics for further exploration, moderators in Condition A more frequently prioritized Task D in their top two rankings ($p = 0.0175$). Participants in Condition A were more likely to ask open-ended interview questions related to the emotionally inconsistent topics ($p = 0.0986$), implying that they perceived that there were UX insights worth digging deeper to uncover in the context of the post user test interview. Furthermore, participants in Condition A were also more likely to propose design recommendations for these topics ($p = 0.0577$). **These results marginally support H1b** which proposes that moderators provided with psychophysiological emotional trends alongside self-perceived responses more frequently prioritize tasks that contain inconsistencies between user behaviour and self-perceived response compared to moderators provided with the user's self-perceived response only.

3.4.3 Empathy (H2)

This study set out to explore whether providing a UX practitioner with a user's psychophysiological trends had any impact on how moderators developed and perceived empathy towards that user. Based on the results from the experiment, participants who received these psychophysiological trends perceived themselves as significantly more empathic towards the user.

Table 14: Mean perceived empathy towards user

Condition	N	Construct	Mean	SD σ
A	11	Emotional interest	6.48	0.79
		Emotional sensitivity	6.18	0.66
		General Empathy	6.33	0.65
B	11	Emotional interest	6.14	0.48
		Emotional sensitivity	5.43	0.70
		General empathy	5.78	0.55

Table 15: Perceived empathy result

Variable	Statistic (*S*)	*Z*-score	*p* value
Emotional interest	153.5	1.7709	0.0356**
Emotional sensitivity	161.5	2.2868	0.0092**
General empathy	161.5	2.2712	0.01**

Moderators who received the user's emotional trends felt more emotional interest ($p = 0.0356$) and emotional sensitivity ($p = 0.0092$) towards the fictitious user depicted in the user test, which suggests that this form of emotional representation stimulates both cognitive and emotional empathic responses. Based on the scores across emotional interest and emotional sensitivity which were measured with a 7-point Likert scale, we derived an overall mean empathy score of 6.33 for moderators who received implicit and explicit responses compared to 5.78 for moderators who received exclusively explicit measures. A one-tailed Wilcoxon rank sum test results suggest that this represents a statistically significant difference across conditions and leads to the conclusion that **H2a and H2b are supported.**

3.5 Discussion

It has long been established that implicit measures support a wide range of UX insights; they provide unmatched contextual granularity compared to relying on explicit measures. However, this study provides compelling evidence in favor of expanding how exactly implicit measures can be used to support user testing. Rather than using them retrospectively to provide insights on a completed user test, this study provides evidence that moderators can leverage this information effectively throughout a user test to help them with decision-making and support understanding throughout the test itself. This study set out to measure the impact that implicit measures have on user testing when combined alongside traditional explicit measures. By controlling the sources of task-based user response data information provided to UX moderators throughout a user test, we were able to measure the differences in their inferential accuracy, their prioritization tendencies, and the degree of empathy they felt towards the subject of the user test (i.e., the fictitious user).

The results from this study lead to various contributions to theory and implications for practice. First and foremost, this study provides evidence that implicit measures, particularly those that are displayed through aggregated psychophysiological trends and made accessible to moderators for analysis throughout the user test as it takes place, is helpful in supporting various moderator performance outcomes. The findings lead us to believe that it is feasible to integrate this visually represented form of implicit data into user testing, leading to enhanced concurrent triangulation of implicit measures, explicit measures, and behavioural observation. Providing moderators with tools that promote immediate detection of useability issues as they occur allows them to develop a more informed approach to prioritizing topics. By implementing a concurrent triangulation approach to understanding interactions, moderators can user various sources of data to cross-analyze factual inconsistencies and develop a more deliberate methodological approach to prioritizing further exploration or clarification of specific elements of the user test. Given that interviews are an essential component of co-creating meaning between UX practitioners and users, having the tools that would support more intentional evidence-based interview strategies could support UXPs with extracting superior UX insights during their inherently finite time alongside one another.

Another vein of compelling results is related to the role of implicit measures on the formation of empathy towards users in the context of user testing. The results from the experiment suggest that providing moderators with a user's psychophysiological emotional trends has a significant effect on perceived empathy towards that user. This can be explained by the perception-action mechanism, which has been found to apply across various human-to-human interactions related to other professional practices. While operationalizing empathy is challenging in terms of performance outcomes, there is robust evidence showing its beneficial impact on creativity and relevant problem-solving in the context of UX design thinking (Dorst & Cross, 2001). Therefore, perception of empathy might theoretically lead to moderators feeling more inclined to forge connections with users and predispose cognitive and emotional empathic responses.

3.5.1 Limitations and future directions

Having a standardized user test meant that we could replicate the experimental circumstances across participants and conditions. We were able to achieve this through the creation of a simulated user test. However, this came with limitations on how closely we could emulate real-world circumstances. Therefore, the primary limitation to this research is that participants who completed the simulated user test were not faced with a live data-generating user. Another limitation to the study is that the research team was responsible for producing the ground truth. For example, it was our responsibility to determine what constituted a pain-point, and when exactly it took place. To create a cohesive emotional representation of this, we closely matched physiological patterns with the implied occurrence of them; this is representative of how implicit measures closely match implicit pain points. However, this provided the group who received implicit measures an advantage in inferring the occurrence of pain points. Nonetheless, the evidence showing the efficacy of interpreting physiological measures throughout a user test is foundational evidence for this methodological approach's feasibility. While ambitious, future research should conduct a similar experiment using paired samples: UX professionals with similar backgrounds being paired to either condition, as well as a user who self-discloses their own occurrence of pain points, having a double-blind approach to evaluating the performance of moderating participants.

Another limitation is that we attributed higher value to tasks that had behavioural and self-reported inconsistencies. Attributing highest value to the task where the self-reported measure did not reflect the behavioural observation assumes this is the most problematic task that should be pursued. However, one could make the argument that focusing on the consistently negative task is also valuable. The reason it was coded as such is because the consistently negative task represents a more “obvious” UX insight, and thus does not represent the objectives of this methodological approach: to identify inconsistencies through a concurrent triangulation approach to user testing.

Finally, there are ethical dilemmas related to interpreting physiological measures while the user is in the moderator’s direct vicinity (Stepanov, 2023). Considering the collaborative nature of user testing, it would be of vital importance to prevent the user from feeling that they are being subjected to privacy concerns. For instance, the asymmetry involved in accessing the user’s biofeedback which can highlight inconsistencies in their responses could lead to ethical issues (Moge et al., 2022). To prevent the user from feeling as though they are being subject to lie detection test, future studies should investigate how users respond to collaborative sharing of emotional trends, or any other implicit measurement that is supported through visually accessible formats. For example, organizing emotional trends into discrete categories could achieve similar informational objectives while minimizing the impression of being tracked in such a precise manner as this test. While evidence suggests that perceived empathy was enhanced when moderators were provided with the user’s emotional trends, future studies should evaluate whether this has an influence on prosocial behaviours and empathic design outcomes.

Additionally, future studies should continue generating knowledge on concurrent triangulation approaches to user testing by assessing alternative forms of data representation, and further study on how UX practitioners perceive, trust, and manage multiple forms of implicit and explicit data. Future research should attempt to replicate similar research objectives using wearable technology in a remote setting; that is, using smart wearables and laptop computers to generate the same category of psychophysiological emotional trends as was used in this study. This would enable many of the benefits of psychophysiological approaches to user testing while also leveraging the strengths of remote data collection.

Finally, further research could build upon the experimental design of this study. The use of a simulation to emulate the act of conducting a user test shows much promise. During the post-experiment interview, participants noted that this simulation could be used as an instructional, training, or even practice tool prior to jumping into interviews with real participants. It would allow UXPs to develop and practice their individual strategic approach to user testing sessions, or even as an evaluative tool whereby another professional can evaluate their open questions and design recommendations.

3.5.2 Conclusion

This research proposes a novel methodological approach to integrating implicit measures into live moderated user testing, such that it integrates live representation of psychophysiological trends into a concurrent triangulation approach to analyzing the user test and thus understanding the user experience. By doing so, it allows UX practitioners moderating the user test to extract the benefits of implicit measures while the user is still present, further compounding richness of insights. Triangulating implicit data alongside self-reported measures allows moderators to identify misreporting and take an evidence-based approach to exploring topics during the interview. Considering the positive effects that visually represented emotional trends seem to have in terms of performance outcomes and the cultivation of empathy, UX researchers and designers could benefit from integrating this approach into their existing user testing workflows. Similarly, companies developing products that focus on live representation of raw data should consider purposefully reducing sensitivity and specificity since aggregated trends portray much of the same information while supporting inferential capacity.

References

Actis-Grosso, R., Capellini, R., Ghedin, F., & Tassistro, F. (2021, July). Implicit Measures as a Useful Tool for Evaluating User Experience. In *International Conference on Human-Computer Interaction* (pp. 3-20). Cham: Springer International Publishing.

Alves, R., Valente, P., & Nunes, N. J. (2014, October). The state of user experience evaluation practice. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (pp. 93-102).

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

Bargas-Avila, J. A., & Hornbæk, K. (2011, May). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2689-2698)

Baumeister, R., Bratslavsky, E., Finkenauer, C., Vohs, K.: Bad is stronger than good. Review of General Psychology. 5, 323-370 (2001)

Betella, A., & Verschure, P. F. (2016). The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one*, *11*(2), e0148037.

Brown, T. A., Sautter, J. A., Littvay, L., Sautter, A. C., & Bearnes, B. (2010). Ethics and personality: Empathy and narcissism as moderators of ethical decision making in business students. *Journal of Education for Business, 85*(4), 203-208.

Bush, T. (2012). Authenticity in research: Reliability, validity and triangulation. *Research methods in educational leadership and management*, *6*(19), 75-89.

Carlgren, L., Rauth, I., & Elmquist, M. (2016). Framing design thinking: The concept in idea and enactment. *Creativity and innovation management*, *25*(1), 38-57.

Cockburn, A., Quinn, P., Gutwin, C.: The effects of interaction sequencing on user experience and preference. International Journal of Human-Computer Studies. 108, 89104 (2017)

Dorst, K., & Cross, N. (2001). Creativity in the design process: co-evolution of problem–solution. *Design studies*, *22*(5), 425-437.

Drouet, L., Bongard-Blanchy, K., Koenig, V., & Lallemand, C. (2022, April). Empathy in design scale: development and initial insights. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1-7).

Escalas, J. E., & Stern, B. B. (2003). Sympathy and empathy: Emotional responses to advertising dramas. *Journal of Consumer Research*, *29*(4), 566-578.

Forlizzi, J. & Battarbee, K., 2004, Understanding experience in interactive systems. In Proceedings of the 2004 conference on Designing Interactive Systems (DIS 04): processes, practices, methods, and techniques (New York: ACM), p. 261.

Georges, V., Courtemanche, F., Sénécal, S., Léger, P. M., Nacke, L., & Fredette, M. (2017). The Evaluation of a Physiological Data Visualization Toolkit for UX Practitioners: Challenges and Opportunities. In Workshop on Strategies and Best Practices for Designing, Evaluating and Sharing Technical HCI Toolkits (HCI Tools).

Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., & Léger, P. M. (2019). Identifying psychophysiological pain points in the online user journey: the case of online grocery. In *Design, User Experience, and Useability. Practice and Case Studies: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part IV 21* (pp. 459-473). Springer International Publishing.

Hassenzahl, M. 2008. User experience (UX): towards an experiential perspective on product quality. In Proc. of the 20th international Conference of the Association Francophone D'interaction Homme-Machine. IHM '08, vol. 339. (2008) ACM, New York, NY, 11-15.

Jain, P., Djamasbi, S., & Wyatt, J. (2019). Creating value with proto-research persona development. In Proceedings of HCI in Business, Government and Organizations.

Johnson, B. & Onwuegbuzie, A., and Lisa A Turner. 2007. Toward a definition of mixed methods research. Journal of mixed methods research 1, 2 (2007), 112–133.

Law, E., Roto, V., Hassenzahl, M., Vermeeren, A., and Kort, J. (2009). Understanding, Scoping and Defining User eXperience: A Survey Approach. Proc. CHI'09, ACM SIGCHI conference on Human Factors in Computing Systems.

Marci, C. D., Ham, J., Moran, E., & Orr, S. P. (2007). Physiologic correlates of perceived therapist empathy and social-emotional process during psychotherapy. *The Journal of nervous and mental disease*, *195*(2), 103-111.

Marsden, N., & Wittwer, A. (2022). Empathy and exclusion in the design process. *Frontiers in Human Dynamics*, *4*, 1050580.

Haag, M., & Marsden, N. (2019). Exploring personas as a method to foster empathy in student IT design teams. *International journal of technology and design education*, *29*, 565-582.

Michalec, B., & Hafferty, F. W. (2021). Challenging the clinically-situated emotion-deficient version of empathy within medicine and medical education research. *Social Theory & Health*, 1- 19.

Milk, C. (2015). How virtual reality can create the ultimate empathy machine. *TED talk*, *22*.

Moge, C., Wang, K., & Cho, Y. (2022, April). Shared user interfaces of physiological data: systematic review of social biofeedback systems and contexts in hci. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

Mucz, D., & Gareau-Brennan, C. (2018). "MAPPING THE JOURNEY: Engaging with customers, identifying touch points, and developing recommendations at a public library." Library Journal, 143(17), 32.

Ortiz de Guinea, A. O., Titah, R., & Léger, P. M. (2014). Explicit and implicit antecedents of users' behavioural beliefs in information systems: A neuropsychological investigation. Journal of Management Information Systems, *30*(4), 179-210.

Ortiz de Guinea, A. O, and Webster, j. An investigation of information systems use patterns: technological events as triggers, the effects of time, and consequences for performance. MIS Quarterly, 37, 4 (2013), 1165–1188

Pettersson, I., Lachner, F., Frison, A. K., Riener, A., & Butz, A. (2018, April). A Bermuda triangle? A Review of method application and triangulation in user experience e valuation. In *Proceedings of the 2018 CHI conference on human factors in computing systems* pp. 1-16).

Pine, B. J., & Gilmore, J. H. (2013). The experience economy: Past, present and future. *Handbook on the Experience Economy*. https://doi.org/10.4337/9781781004227.00007

Platzer, D. (2018, October). Regarding the pain of users: towards a genealogy of the "pain point". In *Ethnographic Praxis in Industry Conference Proceedings* (Vol. 2018, No. 1, pp. 301- 315).

Pratte, S., Tang, A., & Oehlberg, L. (2021, February). Evoking empathy: a framework for describing empathy tools. In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction* (pp. 1-15).

Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioural and brain sciences*, *25*(1), 1-20.

Riedl, R., & Léger, P.M.: Fundamentals of NeuroIS Studies in Neuroscience, Psychology and Behavioural Economics. Springer, Berlin, Heidelberg (2016)

Schooler, J. W., & Eich, E. (2000). Memory for emotional events.

Simons, T., Gupta, A., & Buchanan, M. (2011). Innovation in R&D: Using design thinking to develop new models of inventiveness, productivity and collaboration. *Journal of Commercial Biotechnology*, *17*, 301-307.

Soto, J. A., & Levenson, R. W. (2009). Emotion recognition across cultures: the influence of ethnicity on empathic accuracy and physiological linkage. *Emotion*, *9*(6), 874.

Stepanova, E. R., Desnoyers-Stewart, J., Kitson, A., Riecke, B. E., Antle, A. N., El Ali, A., ... & Howell, N. (2023, July). Designing with Biosignals: Challenges, Opportunities, and Future Directions for Integrating Physiological Signals in Human-Computer Interaction. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference* (pp. 101-103).

Surma-aho, A., Chen, C., Hölttä-Otto, K., & Yang, M. (2019, August). Antecedents and outcomes of designer empathy: A retrospective interview study. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (Vol. 59278, p. V007T06A033). American Society of Mechanical Engineers.

Vermeeren, A. P., Law, E. L. C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010, October). User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries* (pp. 521-530).

Vignemont F, Singer T (2006) The empathic brain: how, when and why? Trends in Cognitive Sciences 10:435–441.

Wang, B., Miao, Y., Zhao, H., Jin, J., & Chen, Y. (2016). “A biclustering-based method for market segmentation using customer pain points.” Engineering Applications of Artificial Intelligence, 47, 101-109.

Wright, P., & McCarthy, J. (2008, April). Empathy and experience in HCI. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 637-646).

Zaki J, The neuroscience of empathy: progress, pitfalls and promise. Nat Neurosci. 2012;15(5):675-680. doi:10.1038/nn.3085

Chapter 4. Managerial Article

Enhancing the toolkit: Equipping UX Practitioners with psychophysiological trends during user test moderation

Pascal Snow, Pierre-Majorique Léger and Sylvain Sénécal
HEC Montréal

Summary

Physiological measurement tools can assess emotional experience with unmatched precision and granularity. Historically, many of the tools used to collect and post process physiological (i.e., implicit) data have been time consuming and capital intensive. However, as this emergent technology becomes increasingly prevalent and accessible, its methodological implementation has remained stagnant. More specifically, UX practitioners (UXPs) collecting physiological data throughout a product testing cycle seem to exclusively use it to produce retrospective insights on the user test; one that has finished – the user long gone from the vicinity of the test environment. While this nonetheless affords valuable insights to be derived from implicit test data, it restricts them from effectively engaging in concurrent data triangulation. As a result, it impedes their ability to cooperate with the user in terms of clarifying inconsistent reporting and exploring topics with more deliberate evidence-based reasoning. The purpose of this article is to highlight the productive potential of redefining how implicit physiological measures are used to support UXPs throughout the user testing process, including the posttest interview, by integrating the analysis of implicit data into the user test *as it unfolds*. By reframing the methodological framework around how psychophysiological measurement approaches are integrated into user testing, this article suggests that it would have positive effects on performance outcomes and empathic design more broadly.

4.1 Introduction

Affective computing is taking the world by storm as sensors and biometrics are increasingly imbedded into everyday technology. As you unlock your phone with your retina, the watch on your wrist passively measures heart rate, and the speaker in your room listens – ready to respond to voice commands. The global affective computing market was valued at 26.17 billion in 2020 and is estimated to grow at a compound annual growth rate of 33% from 2020 to 2027 (Grand View Research, 2021). As sensors, cameras, microphones, and other hardware become increasingly affordable and integrated into quotient life, the underlying software capabilities increasingly support the overall democratization and potential usefulness.

Included in this are software products such as iMotions (Copenhagen, Denmark) and NoldusHub (Noldus, Wageningen, Netherlands) that make use of this technology and offer marketable psychophysiological research solutions designed to support UX business objectives. However, while affective computing and psychophysiological measurement tools are increasingly being used to support UX research, the paradigm around how they are integrated into existing UX workflows appears unchanged; the implicit data generated is typically used to support retrospective understanding of the user test long-after the subject has left the test environment. As a result, those who collected and are subsequently analyzing the data cannot use it to support their understanding of the test *as it unfolds.* Instead, they must generate projective understanding of what took place, since they no longer have direct access to the user's knowledge and experience to clarify understanding and engage in meaningful dialogue. This restricts the opportunity to effectively construct meaning between the implicit data and the user's firsthand qualitative contributions.

Historically, laborious post-processing and analysis of implicit measures, alongside cost-prohibitive technologies, has prevented UX moderators from effectively implementing implicit measures into typical user testing workflows. Improvement in affective computing, including everyday technologies, have led to democratization of technologies and expedited processes, reducing the time and effort needed to effectively wield implicit data. However, this is not reflected by the current state of UX research and practice across industry. This leads us to the following questions:

Is it feasible for UX practitioners to interpret psychophysiological emotional trends on the fly throughout a user testing session, and if so, how might having access to this information provide performance enhancements to the UX practitioner moderating the user test?

To respond to this question, we conducted an experiment that was intended to emulate the decision-making process involved in a typical user test. 22 UX practitioners with professional experience related to UX carried out the simulation across two conditions: either having access to the user's traditional self-reported responses, or alternatively, receiving these same self-reported responses in addition to visual representations of the user's psychophysiological emotional trends.

4.2 Integrating implicit measures into user testing

Implicit measures are notoriously challenging to manage. In their raw form, they require expert knowledge to process and render useful (Georges et al., 2017). Currently, iMotions (Copenhagen, Denmark) and NoldusHub (Noldus, Wageningen, Netherlands) represent readily deployable SaaS solutions aimed at improving the feasibly of integrating this technology into typical UX testing workflows. Both products offer programs that render raw data into live representation of physiological data. While highly precise and sensitive in terms of representing the fluctuations of biological data, its granularity may be counterproductive; highly dynamic data is erratic and can be challenging to draw inferences from. Tweaking this representation approach, one solution developed at Tech3lab allows for live emotional prediction models that aggregate raw data into intervals that visually depict emotional trends over time. This reduces the inherent volatility of this category of data while simultaneously preserving the benefits afforded by live inference. More specifically, it offers the opportunity to engage in concurrent triangulation of data sets. Not only does this allow UX designers moderating the user test to cross-analyze self-reported measures, easily digestible psychophysiological data, and behavioural observations, but it also encourages an ongoing robust understanding of the user test *as it unfolds* and prepares moderators with the most holistic understanding of the user test prior to conducting the posttest interview. The figures below depict a comparison of the conventional process compared to a recently developed data processing sequence.

Figure I: Retrospective psychophysiological analysis

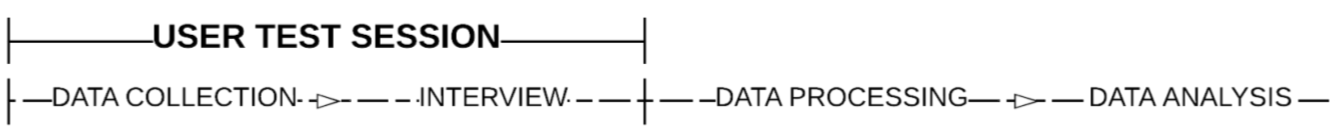
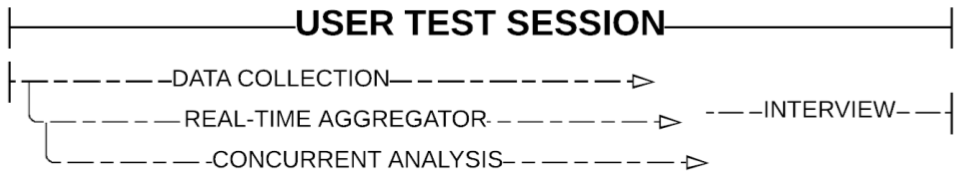


Figure II: Concurrent psychophysiological analysis



These figures juxtapose the predominant form of psychophysiological analysis with the proposed framework. In the former model, data analysis is done after the user test has completed, and the moderator is thus forced to derive insights from psychophysiological data without the opportunity to receive guidance or clarification from the user during the interview. Conversely, the later model depicts a concurrent approach to analysis, where psychophysiological data is aggregated and visually represented on an ongoing basis such that the moderator can interpret and thus enhance their understanding of the user test before the interview takes place. This allows the moderator and user to critically engage with the psychophysiological data, and thus derive higher quality insights from these implicit measures.

4.3 Concurrent triangulation to support the validity of UX insights

When conducting useability studies, roughly 75% of useability issues can be detected over the course of five user tests (Neilson Norman). Beyond this number, each additional user test is said to have diminishing returns. So how might UX practitioners more effectively extract the remaining value from these initial useability tests?

Results from the experiment suggest that providing UX practitioners with psychophysiological emotional trends throughout a user test, when combined with behavioural observations and self-reported measures, produced significantly enhanced performance outcomes compared to moderators who rely on self-reported responses and behavioural observation alone. Based on our findings, we came to the following conclusions:

1. **It improved their ability to identify useability issues by 33%.**

UX practitioners were more likely to correctly infer precisely when the user experienced useability issues throughout a user test when their behavioural observation of the task was paired with psychophysiological emotional trends. Not only does this indicate a higher degree of inferential accuracy, but it also demonstrates the ease-of-interpretation of aggregated psychophysiological emotional trends when engaging in ‘on the fly’ analysis.

2. **The post-test interview focused on tasks that contained unreported useability issues 62% more often.**

Users frequently fail to accurately perceive and report pain points throughout an interaction. Recent studies have demonstrated that less than 25% of pain points were perceived and self-disclosed by participants during an interview following a user test because they inadvertently forgot, marginalized, or failed to notice transient discomfort (Giroux-Hubbé et al., 2019). Our results suggest that when it comes time for UX practitioners to prioritize topics and generate an interview strategy, moderators provided with psychophysiological emotional trends were more likely to prioritize tasks that contained unreported useability issues. Considering the finite time that moderators have with the user during a testing session, it is important that they maximize the value of their time together. Our results suggest that a concurrent analysis framework would assist moderators in developing more evidence-based interview strategies that deliberately focus on interview topics that appear misreported or inconsistent across data sets.

3. **Their design recommendations applied to unreported useability issues 42% more often.**

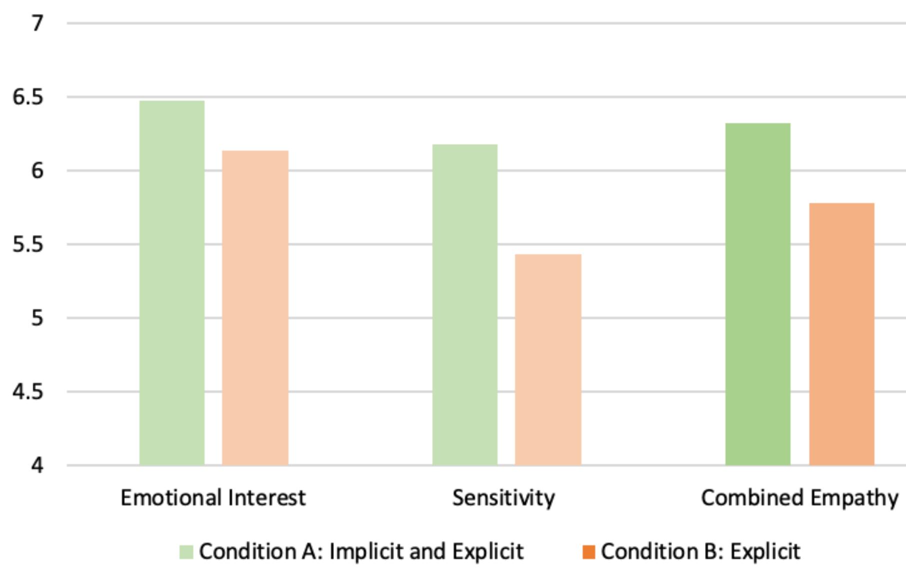
One metric to assess UX evaluation outcomes is the extent to which design objectives represent solutions to the most pertinent product shortcomings. Our study suggests that providing moderators with visual representations of psychophysiological emotions trends supports evidence-based reasoning that leads to more relevant design recommendations – a key metric through which empathic design is achieved.

By engaging in concurrent methodological triangulation throughout a user test, moderators are better equipped to effectively identify subtle pain points, prioritize topics during the posttest interview, and produce design solutions that may have otherwise remained concealed. This leads to a more deliberate and productive allocation of time and effort, and thus supports UX designers with generating the richest insights from a user test during their time alongside the user.

4.4 As a tool to stimulate empathy

Empathy has been found to strengthen social competence and prosocial behaviours across various relationship contexts (Nancy et al., 1987). It has been found to promote more creative, relevant, and satisfying UX design solutions (Carlgren, 2016; Cross, 2002; Simons et al., 2011). In the context of UX, attempts to stimulate empathy are broadly divided across two categories. The first includes conceptual approaches such as personas and user journey maps, while the second category contains tool-based approaches that mimic disability and impairment. Notably, research has proven the biological foundation of empathic behaviour, where humans are biologically disposed to respond to the emotions of others through mirror neurons and emotional contagion (De Waal, 2002). The results from this study provide evidence that visually representing psychophysiological emotional trends leads to higher perceived empathy across two core dimensions of the Empathy in Design Scale: namely, emotional interest [cognitive empathy] and emotional sensitivity [emotional empathy] (Drouet, 2022).

Figure: Perceived empathy towards the user



**Magnified axis from a 7-point Likert scale*

4.4 Conclusion

Improvements to technology and data processes have rendered affective computing into a more feasible addition to the UX toolkit. The proliferation of technologies capable of capturing heart rate, pupillometry, vocal features, and facial expressions will aid in the democratization of psychophysiological inference as they become increasingly integrated into UX research and practice. However, in addition to technological improvements, there is also a need for process enhancements. Historically, complicated psychophysiological data has led to UX practitioners deriving retrospective insights on a past user test, impeding moderators from enhancing *immediate* performance throughout the user testing session and interview. Conversely, current products that do lean into live proportional representations of raw data have introduced SaaS products that are not conducive to easily digestible analysis and on the fly interpretation. Representing raw data through aggregated emotional trends over time-based intervals at the end of each task is more compatible with the constraints of ongoing analysis during user testing, and thus supports the inferential potential for the UX practitioners wielding this valuable information.

Results from the experiment suggest that equipping moderators with psychophysiological emotional trends effectively supports a concurrent triangulation approach to user testing. By equipping them with the tools to effectively cross-analyze sources of data, it enhances their inferential precision in terms of identifying pain points and inconsistent representations of experience. In theory, this leads to a greater opportunity to clarify misunderstandings and mitigate bias. Furthermore, its effect on prioritization tendencies in the context of interview topics seems to lead to more relevant discussion and design recommendations – a fundamental aspect of empathic design. Finally, portraying visualizations of emotional trends seems to enhance both cognitive and affective empathy in the context of the user-moderator relationship, suggesting a novel approach to augmenting empathy in UX evaluation contexts. Given these positive findings, representing psychophysiological data as emotional trends and finding ways to expedite their integration into analysis of a user test appears to be an invaluable methodological approach to the UX evaluation processes.

References:

Affective computing market size: Industry report, 2020-2027. Affective Computing Market Size | Industry Report, 2020-2027. (n.d.). https://www.grandviewresearch.com/industry-analysis/affective-computing-market

Carlgren, L., Rauth, I., & Elmquist, M. (2016). Framing design thinking: The concept in idea and enactment. *Creativity and innovation management*, *25*(1), 38-57.

Georges, V., Courtemanche, F., Sénécal, S., Léger, P. M., Nacke, L., & Fredette, M. (2017). The Evaluation of a Physiological Data Visualization Toolkit for UX Practitioners: Challenges and Opportunities. In *Workshop on Strategies and Best Practices for Designing, Evaluating and Sharing Technical HCI Toolkits (HCI Tools)*.

iMotions, Pedersen, M., & Wang, R. (2023, July 31). *Imotions Lab - Human Behaviour Research Platform*. iMotions. https://imotions.com/products/imotions-lab/

Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., & Léger, P. M. (2019). Identifying psychophysiological pain points in the online user journey: the case of online grocery. In *Design, User Experience, and Useability. Practice and Case Studies: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part IV 21* (pp. 459-473). Springer International Publishing.

Nielsen Norman. World Leaders in Research-Based User Experience. (n.d.). *Why you only need to test with 5 users*. Nielsen Norman Group. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

Noldushub. NoldusHub - Multimodal platform for human behaviour studies. (n.d.). https://www.noldus.com/noldushub

Simons, T., Gupta, A., & Buchanan, M. (2011). Innovation in R&D: Using design thinking to develop new models of inventiveness, productivity and collaboration. *Journal of Commercial Biotechnology*, *17*, 301-307.

Chapter 5: Conclusion

We conducted a between-subject experimental design with 22 IT professionals with UX experience, where we had participants engage with a user testing simulation. The user response data was the manipulated variable, where participants either received implicit and explicit measures (n=11) or exclusively explicit measures (n=11), alongside the behavioural stimuli that was provided to all participants. Overall, this thesis's main aim was to assess the impact of implicit measures on inferential accuracy, prioritization tendencies, and empathy towards the fictitious user in the simulation. This closing chapter will revisit the research questions, hypotheses, and offer the research's conclusions in terms of practical and theoretical contributions to this field.

5.1 Revisiting research questions and hypotheses

This research explored the extent to which implicit measures influence user testing outcomes. More specifically, the extent to which UXPs moderating a user test are impacted by this additional source of information, and whether it supports them with 1) identifying useability issues, 2) aspects of the user test that contained inconsistencies across behaviour and self-reported explicit measures, and 3) empathic sentiment. Data obtained through Qualtrics reporting allowed us to measure decision-making and proclivities across two experimental conditions: participants who received psychophysiological emotional trends and self-reported affective sliders (n=11), or participants who exclusively received self-reported affective sliders (n=11). This data allowed us to meaningfully respond to the following question:

RQ1: *To what extent does providing UX practitioners with a visual representation of the user's psychophysiological trends impact the practitioner's performance outcomes while moderating a user test?*

Results from the experiment suggest that providing UXPs moderating a useability test with the user's psychophysiological emotional trends supports their overall understanding and performance throughout the user test. This suggests that there is a connection between the richness of information provided to UX moderators and their ability to prioritize their efforts more effectively. Results support our initial hypothesis, which assumed that participants who received implicit measures would have better performance outcomes in the user test simulation compared to participants who were exclusively provided explicit measures (**H1**). Consistent with the existing literature that demonstrates the efficacy of implicit measures as a method of enhancing moderators' capacity to retrospectively identify pain points after a completed user test (Giroux-Hubbé et al., 2019), we found that participants provided with implicit measures were similarly more effective at inferring the occurrence of useability issues (**H1.a**). Furthermore, participants who received implicit measures were more likely to prioritize aspects of the user test that contained inconsistencies between the user's behaviour and self-reported explicit response (**H1.b**). This leads us to the next research question:

RQ2: *To what extent does providing UX practitioners with a visual representation of the user's psychophysiological trends impact the practitioner's perceived empathy towards that user while moderating a user test?*

Results from the experiment suggest that providing moderators with the user's psychophysiological emotional trends has the effect of stimulating empathy towards that user (**H2**). While this has been observed across various professional contexts, this result suggests that the relationship persists across the user-moderator relationship within technology-mediated interactions. Observing the user's emotion seems to implicitly trigger an empathic response leading to increased feelings of cognitive (emotional interest) and affective empathy (sensitivity) towards the user (**H2.a; H2.b**). This suggests that incorporating psychophysiological emotional trends throughout a user testing has a peripheral effect on empathy in addition to enhanced performance outcomes.

5.2 Contributions

Theoretical contributions

This research proposes various theoretical contributions to UX evaluation methods. As it stands, implicit measures are used, to our knowledge, almost exclusively to support retrospective understanding of past data collections. While useful, this inherently limits a UXPs ability to meaningfully engage with insights, where the user is no longer present to offer contributions that would assist in clarifying misunderstood implicit indications or inconsistencies across datasets. Despite advancements in current technology that facilitate automation in data processing and representation, the methodological framework used to implement implicit measures appears stagnant. It is as if UX researchers are reluctant to explore alternative implementation schemas outside of the embedded approach. As is the case with any technological innovation; eventually it becomes necessary to facilitate change by challenging previously accepted norms.

Considering this context, we propose a shift in the framework underlying how psychophysiological measurement tools are used to support user testing, such that there is an emphasis on more immediate utilization of psychophysiological measures throughout a user test rather than analyzing them after the user testing session. This represents a fundamentally more productive extraction of value in terms of using this data to support UX evaluation. While there are programs that offer live representations of psychophysiological responses (i.e., NoldusHub and iMotions), they represent a niche in product offerings that offer a real-time display of raw psychophysiological data. These programs pride themselves in having a high degree of specificity and sensitivity. However, we propose that these highly precise representations are potentially counterproductive; they make inference more challenging due to the dynamic and highly variable fluctuations in biological responses from moment to moment. In theory, deliberately reducing specificity, such that raw data is aggregated and displayed in intervals, is a more effective way to represent psychophysiological data when moderators are in a position where they value quick analysis while deriving insights throughout an ongoing a user test. In this sense, specificity does not equate to superiority, and psychophysiological representation frameworks should take this into account when designing SaaS products for UX research and practice.

Existing empathy studies have confirmed the role of biosignals in the formation of various pro-social behaviours, including empathy (Winters et al., 2021; Liu et al., 2019; Curran et al, 2019). However, no studies have assessed the persistence of this relationship in the context of technology mediated remote user testing. This study confirms that viewing a user's emotion, even as a simple visual representation of emotional trends, contributes to perceived empathy towards that user. This can largely be explained by the PAM mechanism that biologically predisposes humans to have an empathic response to emotion, a feedback mechanism that has evolutionarily evolved in humans over time. The PAM mechanism has been observed throughout various academic and professional contexts (Hojat, 2016; Preston & De Waal, 2002), suggesting that it is a strong causal force for empathy in humans, regardless of the circumstances. However, it has not been testing in technology-mediated representations of emotion represented as valence and arousal during a user test. This research not only confirms the phenomenon in user testing, but also the presumed role of PAM as an antecedent to cognitive and affective empathy in the context of user testing based on the results from the Empathy in Design Scale that was administered to participants. It seems that representing a user's emotion through psychophysiological emotional trends, where implicit data is aggregated across intervals and visually portrayed according to linear representations of valence and arousal, is 'enough' to instigate this mechanism and produce empathic predispositions in moderators. Many research models highlight empathy as an early stage in the design thinking process and sequester it to the initial generative phases of the process. This undermines the importance of it throughout the entire process, especially the testing phase. By focusing on empathy in the context of user testing, it would theoretically lead to improved user-moderator relationship outcomes. This favorably impacts prosocial behaviours ranging from general understanding all the way to more engaged and productive dialogue during the posttest interview.

Practical contributions

Researchers have stated that many studies fail to provide research that effectively considers the environment and practical considerations around the implementation of academic UX research into practical contexts (Gray, 2016). This research intended to emulate the conditions around a typical user test while ensuring control over conditions and manipulations needed to derive sound scientific findings. Given that the simulation was intended to mimic real-world user testing processes, the practical findings of this study are more generalizable for practitioners seeking to integrate this framework into practice.

First and foremost, it provides compelling evidence in favor of the overall feasibility of a concurrent triangulation approach to interpreting visualized psychophysiological emotional trends while facing the time constraints of a typical user test. In line with Georges' (2017) framework that outlines what constitutes supportive psychophysiological analysis tools, it seems that visually represented psychophysiological emotional trends based on aggregated data across time-based intervals are well received by UX practitioners. This suggests that this format is an effective representation scheme for implicit measures and allows them to wield it in a way that supports their inferential accuracy and allows them to prioritize further exploration more effectively during posttest interviews. Despite participants being provided with numerous sources of data, it seems that providing visual representations of emotional trends alongside self-reported data can be concurrently understood and juxtaposed to identify inconsistencies and thus improve the reliability of data. While similar live representations of raw psychophysiological data have been introduced by companies such as Noldus (Wageningen, Netherlands) and iMotions (Wageningen, Netherlands), this research offers a novel approach to representing emotional experience across tasks in a way that aggregates data based on time-based intervals rather than a direct proportional representation of live data. By aggregating the data, it minimizes the intrinsic fluctuations in physiological data that can undermine a UX practitioners' ability to effectively derive inferences from the data. While counterintuitive, reducing the sensitivity and specificity of live data by distilling it across linear trends may in fact be a more effective way to represent live data to UX practitioners in a more easily digestible format that is more compatible with the circumstances underlying a user testing session. Based on the findings of the study, this form of visual representation supported moderators with identifying inconsistencies across implicit data, explicit measures, and behavioural observations during a user test which may support bias mitigation and ability to identify useability issues and pain points. This was most likely achieved due to consistency in how the data was represented: both implicit and explicit measures were represented through valence and arousal. This suggests that practical implementation of this concurrent triangulation approach should ensure that there is a high degree of consistency in how self-reported and psychophysiological data is represented, such that they represent comparable constructs rather than aspects than dimensions that are in a class of their own. For example, asking the user to complete a System Useability Scale (Brooke, 1986) as a form of explicit measure would be challenging to meaningfully match with psychophysiological emotional trends.

Finally, given that practical approaches to stimulating empathy in UX are broadly divided across conceptual models (i.e., personas, customer journey maps, empathy mapping) and more literal tools that forcibly impose manufactured simulations (i.e., vision impairment glasses, restrictive gloves), this approach to empathy building represents a novel technique of stimulating empathy within the user-moderator relationship developed during a user testing session. Considering that empathy is a highly desirable trait in UX design, these findings suggest that physiological responses should not just be used for their informational value, but also as a technique for passively stimulating empathy towards users.

To summarize, this research provides compelling preliminary evidence in favor of concurrent triangulation of self-reported affective self-assessments (i.e., explicit), psychophysiological emotional trends (i.e., implicit), and behavioural observations as a method of supporting the moderator's performance outcomes during user testing. This immediate integration of implicit measures offers a more productive approach to leveraging implicit data such that it supports a user test *as it unfolds,* rather than deriving retrospective insights on a completed user test. By doing so, it allows moderators to derive insights that can support the immediate user test and gives them an opportunity to leverage an overall enhanced understanding while engaging in meaningful dialogue with the user while they are still available to provide their own qualitative feedback on the experience. This gives moderators the chance to build upon implicit measures by having a more deliberate and evidence-based approach to exploring topics and clarifying misunderstandings in the context of the interview. Given that moderators conducting user tests are faced with inherently finite time with users, having more informed prioritization of time and human resources is especially important to various stakeholders involved in organizing, funding, and carrying out product testing. Finally, in addition to the various performance metrics discussed, findings suggest that psychophysiological emotional trends may unconsciously stimulate an empathic response towards users during a user testing session. These practical contributions offer actionable guidelines that may help render more effective and efficient UX evaluation processes.

Bibliography

Affective computing market size: Industry report, 2020-2027. Affective Computing Market Size | Industry Report, 2020-2027. (n.d.). https://www.grandviewresearch.com/industry-analysis/affective- computing-market

Agarwal, R., & Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly*, 665-694.

Agourram, H., Alvarez, J., Sénécal, S., Lachize, S., Gagné, J., & Léger, P. M. (2019). The relationship between technology self-efficacy beliefs and user satisfaction–user experience perspective. In *Human-Computer Interaction. Design Practice in Contemporary Societies: Thematic Area, HCI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part III 21* (pp. 389-397). Springer International Publishing.

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International journal of industrial ergonomics, 36*(7), 623-636.

Arhippainen, L., Pakanen, M., & Hickey, S. (2013). Mixed UX methods can help to achieve triumphs. In *Proceedings of CHI 2013 Workshop "Made for Sharing: HCI Stories for Transfer, Triumph and Tragedy* (pp. 83-88).

Al-Azzawi, A. Experience with Technology; Springer: London,UK, 2014. Dordrecht, A2003; Volume 3, pp. 31–42. 5.

Albert, J., Diéguez-Risco, T., Aguado, L., & Hinojosa, J. A. (2013). Faces in context: Modulation of expression processing by situational information. *Social neuroscience, 8*(6), 601-620.

Alves, R., Valente, P., & Nunes, N. J. (2014, October). The state of user experience evaluation practice. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* pp. 93- 102).

Ariely, D., & Carmon, Z. (2003). *Summary assessment of experiences: The whole is different from the sum of its parts*. Russell Sage Foundation.

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

Bailey, B. P., Adamczyk, P. D., Chang, T. Y., Chilson, N. A.: A framework for specifying and monitoring user tasks. Computers in Human Behaviour, 22(4), 709-732 (2006)

Bailey, B. P., & Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behaviour*, *22*(4), 685- 708.

Bargas-Avila, J. A., & Hornbæk, K. (2011, May). Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2689-2698)

Barki, h.; Paré, G.; and Sicotte, C. Linking It implementation and acceptance via the construct of psychological ownership of information technology. Journal of Information Technology, 23, 4 2008), 269–280.

Bartoszek, G., & Cervone, D. (2017). Toward an implicit measure of emotions: Ratings of abstract images reveal distinct emotional states. *Cognition and Emotion*, *31*(7), 1377 -1391.

Batson, D.; Fultz, J; & Schoenrade, P. "Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences", *J. Pers.*, vol. 55, no. 1, pp. 19-39, 1987.

Bauer, H. H., Falk, T., & Hammerschmidt, M. (2006). eTransQual: A transaction process-based approach for capturing service quality in online shopping. *Journal of business research*, *59*(7), 866-875.

Baumeister, R., Bratslavsky, E., Finkenauer, C., Vohs, K.: Bad is stronger than good. Review of General Psychology. 5, 323-370 (2001)

Bazerman, M. H., & Moore, D. A. (2012). *Judgment in managerial decision making*. John Wiley & Sons.

Benson, B. (2016). Cognitive bias cheat sheet. better humans. https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18.

Betella, A., & Verschure, P. F. (2016). The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one*, *11*(2), e0148037.

Bernhaupt, R., & Pirker, M. (2013). Evaluating user experience for interactive television: towards the development of a domain-specific user experience questionnaire. In *Human-Computer Interaction– INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part II 14* (pp. 642-659). Springer Berlin Heidelberg.

Bevan, N. (2009, August). What is the difference between the purpose of useability and user experience evaluation methods. In *Proceedings of the Workshop UXEM* (Vol. 9, No. 1, pp. 1-4).

Biggest companies in the world by market Cap 2023. Statista. (2023, August 8). https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-capitalization/

Bigné, J. E., Andreu, L., & Gnoth, J. (2005). The theme park experience: An analysis of pleasure, arousal and satisfaction. *Tourism management*, *26*(6), 833-844.

Bogers, M., & Horst, W. (2014). Collaborative prototyping: Cross-fertilization of knowledge in prototype-driven problem solving. *Journal of Product Innovation Management*, *31*(4), 744-764.

Brooke, J. (1996). Sus: a "quick and dirty'useability. *Useability evaluation in industry*, *189*(3), 189-194.

Brown. T. 2009. Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation. New York: Harper Collins.

Brown, T. A., Sautter, J. A., Littvay, L., Sautter, A. C., & Bearnes, B. (2010). Ethics and personality: Empathy and narcissism as moderators of ethical decision making in business students. *Journal of Education for Business*, *85*(4), 203-208.

Brunel, F. F., Ruth, J. A., & Otnes, C. C. (2002). Linking thoughts to feelings: Investigating cognitive appraisals and consumption emotions in a mixed-emotions context. *Journal of the Academy of Marketing Science*, *30*, 44-58.

Bruun, A. (2018, September). It's not complicated: A study of non-specialists analyzing GSR sensor data to detect UX related events. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (pp. 170-183).

Bruun, A., & Ahm, S. (2015). Mind the gap! Comparing retrospective and concurrent ratings of emotion in user experience evaluation. In *Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg,*

Buie, E., Dray, S., Instone, K., Jain, J., Lindgaard, G., & Lund, A. (2010). How to bring HCI research and practice closer together. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 3181-3184).

Johnson, B. & Onwuegbuzie, A., and Lisa A Turner. 2007. Toward a definition of mixed methods research. Journal of mixed methods research 1, 2 (2007), 112–133.

Bush, T. (2012). Authenticity in research: Reliability, validity, and triangulation. *Research methods in educational leadership and management*, *6*(19), 75-89.

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2007). Psychophysiological science: Interdisciplinary approaches to classic questions about the mind. *Handbook of psychophysiology*, *3*, 1-16.

Cajander, Å., Larusdottir, M., & Geiser, J. L. (2022). UX professionals' learning and usage of UX methods in agile. *Information and Software Technology*, *151*, 107005

Carlgren, L., Rauth, I., & Elmquist, M. (2016). Framing design thinking: The concept in idea and enactment. *Creativity and innovation management*, *25*(1), 38-57.

Chang, Y., Lim, K., and Stolterman, E. 2008. Personas: From theory to practices. In Proceedings of the 5th Nordic Conference on Humancomputer Interaction: Building Bridges, 439-442.

Chang-Arana, Á. M., Piispanen, M., Himberg, T., Surma-aho, A., Alho, J., Sams, M., & Hölttä-Otto, K. (2020). Empathic accuracy in design: Exploring design outcomes through empathic performance and physiology. *Design Science, 6*, e16.

Charles, R., Nixon, J.: Measuring mental workload using physiological measures: A systematic review. Applied Ergonomics. 74, 221-232 (2019)

Cipolla, C., & Bartholo, R. (2014). Empathy or inclusion: A dialogical approach to socially responsible design. *International Journal of Design*, *8*(2).

Cockburn, A., Quinn, P., Gutwin, C.: The effects of interaction sequencing on user experience and preference. International Journal of Human-Computer Studies. 108, 89104 (2017)

Cockburn A, Quinn P, Gutwin C.: Examining the Peak-End Effects of Subjective Experience. In: Proceedings of the 33rd annual ACM conference on human factors in computing systems, pp. 357–366 (2015)

Cooper, J. R. (2005). *Curing analytic pathologies: Pathways to improved intelligence analysis* (p. 6). Washington, DC: Center for the Study of Intelligence.

Curran, M. T., Gordon, J. R., Lin, L., Sridhar, P. K., & Chuang, J. (2019, May). Understanding digitally-mediated empathy: An exploration of visual, narrative, and biosensory informational cues. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13)

d.school, *An Introduction to Design Thinking PROCESS GUIDE*, 2010.

Decety, J., & Batson, C. D. (2009). Empathy and morality: Integrating social and neuroscience approaches. *The moral brain: Essays on the evolutionary and neuroscientific aspects of morality*, 109-127.

De Waal, F. B., & Preston, S. D. (2017). Mammalian empathy: behavioural manifestations and neural basis. *Nature Reviews Neuroscience, 18*(8), 498-509.

De Waal, F. B. (2007). The Russian doll model of empathy and imitation. *On Being Moved: From Mirror Neurons to Empathy.* Advances in consciousness research

Dorst, K., & Cross, N. (2001). Creativity in the design process: co-evolution of problem–solution. *Design studies*, *22*(5), 425-437.

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

Dirican, A. C., & Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science*, *3*, 1361-1367.

Djamasbi, S., & Strong, D. (2019). User experience-driven innovation in smart and connected worlds. *AIS Transactions on Human-Computer Interaction*, *11*(4), 215-231.

Dorst, K., & Cross, N. (2001). Creativity in the design process: co-evolution of problem–solution. *Design studies*, *22*(5), 425-437.

Drouet, L., Bongard-Blanchy, K., Koenig, V., & Lallemand, C. (2022, April). Empathy in design scale: development and initial insights. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1-7).

Dwivedi, M. S. K. D., Upadhyay, M. S., & Tripathi, A. (2012). A working framework for the user-centered design approach and a survey of the available methods. *International Journal of Scientific and Research Publications*, *2*(4), 12-19.

Ellis, G. (2018). So, what are cognitive biases? *Cognitive biases in visualizations*, 1-10.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1978). Emotion in the human face: guide-lines for research and an integration of findings. *(No Title)*.

Ekman, P. (1984). Expression and the nature of emotion. *Approaches to emotion*, *3*(19), 344.

Ekman, P. (1997). Expression or communication about emotion.

Escalas, J. E., & Stern, B. B. (2003). Sympathy and empathy: Emotional responses to advertising dramas. *Journal of Consumer Research*, *29*(4), 566-578.

Fang, Yulin, Israr Qureshi, Heshan Sun, Patrick McCole, Elaine Ramsey et Kai H Lim (2014). « Trust, satisfaction, and online repurchase intention: The moderating role of perceived effectiveness of e-commerce institutional mechanisms », Mis Quarterly, vol. 38, no 2.

Feijt, M. A., Westerink, J. H., De Kort, Y. A., & IJsselsteijn, W. A. (2023). Sharing biosignals: An analysis of the experiential and communication properties of interpersonal psychophysiology. *Human–Computer Interaction*, *38*(1), 49-78.

Følstad, A., Law, E., & Hornbæk, K. (2012, May). Analysis in practical useability evaluation: a survey study. In *proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2127-2136).

Forlizzi, J. & Battarbee, K., 2004, Understanding experience in interactive systems. In Proceedings of the 2004 conference on Designing Interactive Systems (DIS 04): processes, practices, methods, and techniques (New York: ACM), p. 261.

Fuentes, C., Gerea, C., Herskovic, V., Marques, M., Rodríguez, I., & Rossel, P. O. (2015). User interfaces for self- reporting emotions: a systematic literature review. In *Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information: 9th International Conference, UCAmI 2015, Puerto Varas, Chile, December 1-4, 2015, Proceedings 9* (pp. 321-333). Springer International Publishing.

Furniss, D. (2008). *Beyond Problem Identification: Valuing methods in a'system of useability practice'*. University of London, University College London (United Kingdom).

Gasparini, A. A. (2015). Perspective and Use of Empathy in Design Thinking. Advancements in Computer-Human Interaction, ACHI, Lisbon.

Georges, V., Courtemanche, F., Sénécal, S., Léger, P. M., Nacke, L., & Fredette, M. (2017). The Evaluation of a Physiological Data Visualization Toolkit for UX Practitioners: Challenges and Opportunities. In *Workshop on Strategies and Best Practices for Designing, Evaluating and Sharing Technical HCI Toolkits (HCI Tools)*.

Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., & Léger, P. M. (2019). Identifying psychophysiological pain points in the online user journey: the case of online grocery. In *Design, User Experience, and Useability. Practice and Case Studies: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part IV 21* (pp. 459-473). Springer International Publishing.

Goldman, A. (2011). Two routes to empathy. *Empathy: Philosophical and psychological perspectives*, 31-44.

Goodman, E. S. (2013). *Delivering design: Performance and materiality in professional interaction design*. University of California, Berkeley

Gray, C. M. (2016, May). " It's More of a Mindset Than a Method" UX Practitioners' Conception of Design Methods. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 4044-4055).

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4.

Hackman, J. R. (2011). Collaborative intelligence: Using teams to solve hard problems San Francisco, CA: Berrett-Koehler Publishers.

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

Hassenzahl, M. 2008. User experience (UX): towards an experiential perspective on product quality. In Proc. of the 20th international Conference of the Association Francophone D'interaction Homme-Machine. IHM '08, vol. 339. (2008) ACM, New York, NY, 11-15.

Hassenzahl, M. The Thing and I: Understanding the Relationship between User and Product. In Funology: From Useability to Enjoyment; Kluwer Academic Publishers: Dordrecht, Germany, 2003; Volume 3, pp.

Hassenzahl, M.; Burmester, M.; Koller, F. Der User Experience (UX) auf der Spur: Zum Einsatz von. Available online: www.attrakdiff (accessed on 12 February 2020).

Hayashi, E. & Hong, J. 2015. Knock x knock: the design and evaluation of a unified authentication management system. In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, 379–389.

Heylighen, A., & Dong, A. (2019). To empathise or not to empathise? Empathy and its limits in design. *Design Studies*, *65*, 107-124.

Heylighen, A., & Devlieger, P. (2007). In dialogue with (dis-) Ability. *Urban Trialogues. Coproductive ways to relate visioning and strategic urban projects*, 8.

Ho, D. K. L., Ma, J., & Lee, Y. (2011). Empathy@ design research: a phenomenological study on young people experiencing participatory design for social inclusion. *CoDesign*, *7*(2), 95-106.

Hojat, M. (2007). *Empathy in patient care: antecedents, development, measurement, and outcomes* (Vol. 77). New York: springer.

Houwer, J. (2006). What are implicit measures and why are we using them? *The handbook of implicit cognition and addiction*, 11-28.

Hoffman, M. L. (2000). Empathy, its arousal, and prosocial functioning. *Empathy and Moral Development*, *2*, 29-62.

Howell, N., Devendorf, L., Tian, R., Vega Galvez, T., Gong, N. W., Poupyrev, I., ... & Ryokai, K. (2016, June). Biosignals as social cues: Ambiguity and emotional interpretation in social displays of skin conductance. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (pp. 865-870).

Hussein, I., Mahmud, M., & Tap, A. O. M. (2014, September). A survey of user experience practice: a point of meet between academic and industry. In *2014 3rd International Conference on User Science and Engineering (i-USEr)* (pp. 62-67). IEEE.

iMotions, Pedersen, M., & Wang, R. (2023, July 31). *Imotions Lab - Human Behaviour Research Platform*. iMotions. https://imotions.com/products/imotions-lab/

Interaction Design Foundation. (2022, July 12). *What is design thinking?* The Interaction Design Foundation. https://www.interaction-design.org/literature/topics/design-thinking

ISO 9241-11:2018. ISO. (2023, April 15). https://www.iso.org/standard/63500.html

ISO DIS 9241--210. Ergonomics of human system interaction - part 210: Human-centred design for interactive systems. Tech. rep., International Organization for Standardization, Switzerland, 2010

Jain, P., Djamasbi, S., & Wyatt, J. (2019). Creating value with proto-research persona development. In Proceedings of HCI in Business, Government and Organizations.

Johnson, R. & Onwuegbuzie, A., and Lisa A Turner. 2007. Toward a definition of mixed methods research. Journal of mixed methods research 1, 2 (2007), 112–133.

Jordan, P. W. (2000). Contemporary Trends and Product Design. *Contemporary Ergonomics*.

Kaasinen, E., Roto, V., Hakulinen, J., Heimonen, T., Jokinen, J. P., Karvonen, H., ... & Turunen, M. (2015). Defining user experience goals to guide the design of industrial systems. *Behaviour & Information Technology*, *34*(10), 976-991.

Karahanna, E., and Straub, D.W. the psychological origins of perceived usefulness and ease-of-use. Information & Management, 35, 4 (1999), 237–250.

Kaye, J. Evaluating experience-focused HCI. In CHI '07 Extended Abstracts on Human Factors in Computing Systems CHI '07. (2007) ACM, New York, NY, 1661-1664. 19.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions.

Kahneman, D., & Knetsch, J. (1993). Strong influences and shallow inferences: An analysis of some anchoring effects. *Unpublished manuscript, University of California, Berkeley*.

Ketola, P., Roto, V. (2008) Exploring User Experience Measurement Needs. 5th COST294-MAUSE Open Workshop on Valid Useful User Experience Measurement (VUUM). Reykjavik, Iceland.

Khan A, Breslav S, Glueck M, Hornbæk K (2015) Benefits of visualization in the mammography problem. Int J Hum-Comput Stud 83:94–113

Kim, S. S., Kaplowitz, S., & Johnston, M. V. (2004). The effects of physician empathy on patient satisfaction and compliance. *Evaluation & the health professions*, *27*(3), 237-251.

Kouprie, M., & Visser, F. S. (2009). A framework for empathy in design: stepping into and out of the user's life. Journal of Engineering Design, 20(5), 437-448. DOI: 10.1080/09544820902875033. https://doi.org/10.1080/09544820902875033

Kretz, D. R. (2018). Experimentally evaluating bias-reducing visual analytics techniques in intelligence analysis. *Cognitive Biases in Visualizations*, 111-135.

Kress, G. L. (2012). *The effects of team member intrinsic differences on emergent team dynamics and long-term innovative performance in engineering design teams*. Stanford University.

Kretz DR (2015) Strategies to reduce cognitive bias in intelligence analysis: can mild interventions improve analytic judgment? The University of Texas at Dallas

Kumar, A., & Oliver, R. L. (1997). Special session summary cognitive appraisals, consumer emotions, and consumer response. *ACR North American Advances*.

Lallemand, C., Gronier, G., & Koenig, V. (2015). User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in human behaviour*, *43*, 35-48.

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

Law, E. L. C., Van Schaik, P., & Roto, V. (2014). Attitudes towards user experience (UX) measurement. International Journal of Human-Computer Studies, *72*(6), 526-541.

Law, E., Roto, V., Hassenzahl, M., Vermeeren, A., and Kort, J. (2009). Understanding, Scoping and Defining User eXperience: A Survey Approach. Proc. CHI'09, ACM SIGCHI conference on Human Factors in Computing Systems.

Lederman, R., Wadley, G., Gleeson, J., Bendall, S., and Álvarez-Jiménez, M., 2014. Moderated online social therapy: Designing and evaluating technology for mental health. ACM Transactions on Computer- Human Interaction (TOCHI) 21, 1 (2014), 5.

Lawson, B. 2006. How designers think: the design process demystified. Architectural Press, Oxford, UK.

Léger, P. M., Courtemanche, F., Fredette, M., & Sénécal, S. (2019). A cloud-based lab management and analytics software for triangulated human-centered research. In Information Systems and Neuroscience: NeuroIS Retreat 2018 (pp. 93-99). Springer International Publishing.

Léger, P. M., Karran, A. J., Courtemanche, F., Fredette, M., Tazi, S., Dupuis, M., ... & Sénécal, S. (2022). Caption and observation based on the algorithm for triangulation (COBALT): Preliminary results from a beta trial. In *NeuroIS Retreat* (pp. 229-235). Cham: Springer International Publishing.

Leonard, D., & Rayport, J. F. (1997). Spark innovation through empathic design. Harvard Business Review, 75(6), 102–113.

Leong, T.W., Vetere, F., & Howard, S. 2012. Experiencing coincidence during digital music listening. ACM Transactions on Computer-Human Interaction (TOCHI) 19, 1 2012), 6.

Lewinski, P., Fransen, M. L., & Tan, E. S. (2014). Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli. *Journal of Neuroscience, Psychology, and Economics*, *7*(1), 1.

Liu, F., Kaufman, G., & Dabbish, L. (2019). The effect of expressive biosignals on empathy and closeness for a stigmatized group member. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1-17.

Loiacono, E. T., Watson, R. T., & Goodhue, D. L. (2002). WebQual: A measure of website quality. *Marketing theory and applications*, *13*(3), 432-438.

Lourties, S., Léger, P. M., Sénécal, S., Fredette, M., & Chen, S. L. (2018). Testing the convergent validity of continuous self-perceived measurement systems: an exploratory study. In *HCI in Business, Government, and Organizations: 5th International Conference, HCIBGO 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings 5* (pp. 132-144). Springer International Publishing.

Maia, C. L. B., & Furtado, E. S. (2016, October). A study about psychophysiological measures in user experience monitoring and evaluation. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems* (pp. 1-9).

Mäkelä, A., Fulton Suri, J. Supporting Users' Creativity: Design to Induce Pleasurable Experiences. Proc. of the Int. Conf. on Affective Human Factors Design, (2001) pp. 387394.

Makki, A. H. (2020). *Design Method to Enhance Empathy for User-Centered Design: Improving the Imagination of the User Experience* (Doctoral dissertation, Carleton U niversity).

Mandryk, R. L., Inkpen, K. M., & Calvert, T. W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & information technology*, *25*(2), 141-158.

Manoogian, J., & Benson, B. (2017). Cognitive bias codex.

Marci, C. D., Ham, J., Moran, E., & Orr, S. P. (2007). Physiologic correlates of perceived therapist empathy and social-emotional process during psychotherapy. *The Journal of nervous and mental disease*, *195*(2), 103-111.

Martin, R. 2009. The Design of Business: Why Design Thinking Is the Next Competitive Advantage. Cambridge MA: Harvard Business Press.

Marsden, N., & Wittwer, A. (2022). Empathy and exclusion in the design process. *Frontiers in Human Dynamics*, *4*, 1050580.

Haag, M., & Marsden, N. (2019). Exploring personas as a method to foster empathy in student IT design teams. *International journal of technology and design education*, *29,* 565-582.

Mandolfo, M., Pavlovic, M., Pillan, M., & Lamberti, L. (2020, July). Ambient UX research: user experience investigation through multimodal quadrangulation. In *International Conference on Human-Computer Interaction* (pp. 305-321). Cham: Springer International Publishing

Marshall, C. & Rossman, G.B., 1999, Designing Qualitative Research, (Thousand Oaks, CA: Sage)

Maslow, A. H. (1954). The instinctoid nature of basic needs. *Journal of personality*.

Mattelmäki, T., Vaajakallio, K., & Koskinen, I. (2014). What happened to empathic design? *Design issues*, *30*(1), 67-77.

Mattelmäki, T., & Battarbee, K. (2002). "Empathy Probes." In PDC 02 Proceedings of the Participatory Design Conference, edited by T. Binder, J. Gregory, and Wagner, 266–271. Retrieved from http://ojs.ruc.dk/index.php/pdc/article/viewFile/265/257

Meehl PE (1954) Clinical versus statistical prediction: a theoretical analysis and a review of the evidence

Meyer, M. W., & Norman, D. (2020). Changing design education for the 21st century. *She Ji: The Journal of Design, Economics, and Innovation*, *6*(1), 13-49.

Micallef L, Dragicevic P, Fekete JD (2012) Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. IEEE Trans Visual Comput Graphics 18(12):2536–2545

Michalec, B., & Hafferty, F. W. (2021). Challenging the clinically-situated emotion-deficient version of empathy within medicine and medical education research. *Social Theory & Health*, 1-19.

Milk, C. (2015). How virtual reality can create the ultimate empathy machine. *TED talk*, *22*.

Mirhoseini S.M.M., Léger PM., Sénécal S.: The Influence of Task Characteristics on Multiple Objective and Subjective Cognitive Load Measures. In: Information Systems and Neuroscience. Lecture Notes in Information Systems and Organisation, vol 16. Springer, Cham (2017)

Miron-Shatz, T., Stone, A., Kahneman, D.: Memories of yesterday's emotions: does the valence of experience affect the memory-experience gap? Emotion **9**, 885–891 (2009)

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information systems research*, *2*(3), 192-222.

Morris, J. D. (1995). Observations: SAM: the Self-Assessment Manikin; an efficient cross- cultural measurement of emotional response. *Journal of advertising research*, *35*(6), 63-68.

Morrow, S. L., & Smith, M. L. (2000). Qualitative research for counseling psychology. *Handbook of counseling psychology, 3*, 199-230.

Nah F. & Xiao, S. (Eds.): HCIBGO 2018, LNCS 10923, pp. 132-144, 2018. https://doi.org/10.1007/978-3-319-91716-0_1

New, S., & Kimbell, L. (2013, September). Chimps, designers, consultants and empathy: A "theory of mind" for service design. In n *2nd Cambridge academic design management conference* (p. 4-5).

Nielsen Norman. World Leaders in Research-Based User Experience. (n.d.). *Why you only need to test with 5 users*. Nielsen Norman Group. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

Nielson, J. (2017). *A 100-year view of user experience.* Nielsen Norman Group. https://www.nngroup.com/articles/100-years-ux/

NoldusHub. (n.d.). *Noldus | Advance your behavioural research*. Retrieved from https://www.noldus.com/noldushub.

Omdahl, B. L. (1995). Cognitive appraisal, emotion, and empathy, ser. *Lecture Notes in Computer Science. New York: Psychology Press*.

Ortiz de Guinea, A. O., Titah, R., & Léger, P. M. (2014). Explicit and implicit antecedents of users' behavioural beliefs in information systems: A neuropsychological investigation. Journal of Management Information Systems, *30*(4), 179-210.

Ortiz de Guinea, A. O, and Webster, j. An investigation of information systems use patterns: technological events as triggers, the effects of time, and consequences for performance. ￼ MIS Quarterly, 37, 4 (2013), 1165–1188

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

Ouellette, j.A., and Wood, W. habit and intention in everyday life: the multiple processes by which past behaviour predicts future behaviour. Psychological Bulletin, 124, 1 (1998), 54– 74.

Perrig, S. A., Aeschbach, L. F., Scharowski, N., von Felten, N., Opwis, K., & Brühlmann, F. 2022). Measurement Practices in UX Research: A Systematic Quantitative Literature Review.

Pettersson, I., Lachner, F., Frison, A. K., Riener, A., & Butz, A. (2018, April). A Bermuda triangle? A Review of method application and triangulation in user experience evaluation. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-16).

Phan, M. H., Keebler, J. R., & Chaparro, B. S. (2016). The development and validation of the game user experience satisfaction scale (GUESS). *Human factors, 58*(8), 1217-1247.

Pine, B. J., & Gilmore, J. H. (2013). The experience economy: Past, present and future. *Handbook on the Experience Economy*. https://doi.org/10.4337/9781781004227.00007

Platzer, D. (2018, October). Regarding the pain of users: towards a genealogy of the "pain point." In *Ethnographic Praxis in Industry Conference Proceedings* (Vol. 2018, No. 1, pp. 301-315).

Poels, K., & Dewitte, S. (2006). How to capture the heart? Reviewing 20 years of emotion measurement in advertising. *Journal of Advertising Research*, *46*(1), 18-37.

Polanyi, M. 1966. The tacit dimension. Anchor Books, Garden City, NY.

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.

Postrel, V., 2002, Positive Psychology: An Introduction. American Psychologist, 55, pp. 5 – 14.

Pratte, S., Tang, A., & Oehlberg, L. (2021, February). Evoking empathy: a framework for describing empathy tools. In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction* (pp. 1-15).

Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioural and brain sciences*, *25*(1), 1-20.

Pronin E, Lin DY, Ross L (2002) The bias blind spot: perceptions of bias in self versus others. Pers Soc Psychol Bull 28(3):369–381

Purdy, C. (2021). *Bias in research for design: Considerations for designers when conducting user experience research* (Master's thesis, NTNU).

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.: Fundamentals of NeuroIS Studies in Neuroscience, Psychology and Behavioural Economics. Springer, Berlin, Heidelberg (2016)

Robinson, J., Lanius, C., & Weber, R. (2017). The past, present, and future of UX empirical research. *Communication Design Quarterly Review*, *5*(3), 10-23.

Roedl, D., and Stolterman, E. 2013. Design research at CHI and its applicability to design practice. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13), 1951-1954. http://doi.acm.org/10.1145/2470654.2466257

Rogers, Y. 2004. New theoretical approaches for HCI. Annual review of information science and technology, 38, 1: 87-143.

Rosenbaum, S., & Kantner, L. (2008, July). Learning about users when you can't go there: Remote attended useability studies. In *2008 IEEE International Professional Communication Conference* (pp. 1-6). IEEE.

Roto V., Obrist M., Väänänen-Vainio-Mattila K. (2009) User Experience Evaluation Methods in Academic and Industrial Contexts. Proceedings of UXEM 09 workshop,

Rubin, J. and Chisnell, D. Handbook of useability testing (2nd.edition). Wiley Publishing, 2008.

Saadé, r., and Bahli, B. the impact of cognitive absorption on perceived usefulness and perceived ease of use in on-line learning: An extension of the technology acceptance model. Information & Management, 42, 2 (2005), 317–327.

Sanders, E. B. N. (2002). From user-centered to participatory design approaches. In *Design and the social sciences* (pp. 18-25). CRC Press.

Sauro, J. (2016). The challenges and opportunities of measuring the user experience. *Journal of Useability Studies*, *12*(1), 1-7.

Scapin, D. L., Senach, B., Trousse, B., & Pallot, M. (2012). User experience: Buzzword or new paradigm? In Proceedings of the ACHI Fifth International Conference on Advances in Computer-Human Interactions.

Schooler, J. W., & Eich, E. (2000). Memory for emotional events.

Scott, B. A., Colquitt, J. A., Paddock, E. L., & Judge, T. A. (2010). A daily investigation of the role of manager empathy on employee well-being. *Organizational Behaviour and Human Decision Processes*, *113*(2), 127-140

Semmer, N. K., Grebner, S., & Elfering, A. (2003). Beyond self-report: Using observational, physiological, and situation-based measures in research on occupational stress. In *Emotional and physiological processes and positive intervention strategies* (pp. 205-263). Emerald Group Publishing Limited.

Sharma, r.; yetton, P.; and Crawford, j. Estimating the effect of common method variance: the method-method pair technique with an illustration from tAM research.MIS Quarterly, 33, 3 (2009), 473–490.

Sheppard, B., Sarrazin, H., Kouyoumjian, G., & Dore, F. (2018, October 25). *The Business Value of Design*. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-design/our-insights/the-business-value-of-design

Simon, H. A. Newell, A., & Shaw, J. C., (1957, February). Empirical explorations of the logic theory machine: a case study in heuristic. In *Papers presented at the February 26-28, 1957, western joint computer conference: Techniques for reliability* (pp. 218-230).

Singer T, Lamm C (2009) The social neuroscience of empathy. Ann NY Acad Sci 1156:81–96

Slovák, P.; Joris Janssen, J.; and Geraldine Fitzpatrick, G. 2012. Understanding heart rate sharing: towards unpacking physiosocial space. In CHI'12. Association for Computing Machinery, New York, NY, USA, 859–868.

Soto, J. A., & Levenson, R. W. (2009). Emotion recognition across cultures: the influence of ethnicity on empathic accuracy and physiological linkage. *Emotion*, *9*(6), 874.

Statista Research Department, & 8, A. (2023, August 8). *Biggest companies in the world by market Cap 2023*. Statista. https://www.statista.com/statistics/263264/top-companies-in-the-world-by-market-capitalization/

Stepanova, E. R., Desnoyers-Stewart, J., Kitson, A., Riecke, B. E., Antle, A. N., El Ali, A., ... & Howell, N. (2023, July). Designing with Biosignals: Challenges, Opportunities, and Future Directions for Integrating Physiological Signals in Human-Computer Interaction. In *Companion Publication of the 2023 ACM Designing Interactive Systems Conference* pp. 101-103).

Stoyanov, S. R., Hides, L., Kavanagh, D. J., Zelenko, O., Tjondronegoro, D., & Mani, M. (2015). Mobile app rating scale: a new tool for assessing the quality of health mobile e apps. *JMIR mHealth and uHealth*, *3*(1), e3422.

Straub, D.W., and Burton-jones, A. Veni, vidi, vici: Breaking the tAM logjam. Journal of the Association for Information Systems, 8, 4 (2007), 223–229.

Suri, F. (2003). The Experience of Evolution: Developments in Design Practice, The Design Journal, 6(2), 39-48. DOI:10.2752/146069203789355471. Retrieved from: https://doi.org/10.2752/146069203789355471

Surma-Aho, A., & Hölttä-Otto, K. (2022). Conceptualization and operationalization of empathy in design research. *Design Studies, 78*, 101075.

Temkin, B. D. (2010). Mapping the customer journey. *Forrester Research*, *3*, 20.

Venkatesh, V. Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. Information Systems Research, 11, 4 (2000), 342–365.5.

Vermeeren, A. P., Law, E. L. C., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010, October). User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries* (pp. 521-530).

Vignemont F, Singer T (2006) The empathic brain: how, when and why? Trends in Cognitive Sciences 10:435–441.

Visser, F. S., Stappers, P. J., Van der Lugt, R., & Sanders, E. B. (2005). Context mapping: experiences from practice. *CoDesign*, *1*(2), 119-149.

Vredenburg, K., Mao, J. Y., Smith, P. W., & Carey, T. (2002, April). A survey of user-centered design practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 471-478).

Wang, B., Miao, Y., Zhao, H., Jin, J., & Chen, Y. (2016). "A biclustering-based method for market segmentation using customer pain points." Engineering Applications of Artificial Intelligence, 47, 101-109.

Westcott, M. (2014, March 10). *Design-driven companies outperform S&P by 228% over ten years*. https://www.dmi.org/blogpost/1093220/182956/Design-Driven-Companies-Outperform-S-P-by-228-Over-Ten-Years--The-DMI-Design-Value-Index.

Weichert, S.; Quint, G.; Bartel, T. Quick guide UX Management: So Verankern Sie Useability und User Experience im Unternehmen; Springer: Wiesbaden, Germany, 2018.

Wiem, M. B. H., & Lachiri, Z. (2017). Emotion classification in arousal valence model using MAHNOB-HCI database. *International Journal of Advanced Computer Science and Applications*, *8*(3).

Wilke, A., & Mata, R. (2012). Cognitive Bias. In: V.S. Ramachandran (ed.) The Encyclopedia of human behaviour, 1, 531-535. Academic Press.

Wilson, G.M. & Sasse, M.A., 2000a, Do Users Always Know What's Good For Them? Utilizing Physiological Responses to Assess Media Quality. InHCI 2000: People and Computers XIV – Useability or Else,pp. 327 – 339 (Sunderland, UK: Springer).

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological review*, *107*(1), 101.

Wimmer, B., Wöckl, B., Leitner, M., & Tscheligi, M. (2010, October). Measuring the dynamics of user experience in short interaction sequences. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries* (pp. 825-828).

Winters, R. M., Walker, B. N., & Leslie, G. (2021, May). Can you hear my heartbeat?: hearing an expressive biosignal elicits empathy. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-11).

Woolrych, A., Hornbæk, K., Frøkjær, E., & Cockton, G. (2011). Ingredients and meals rather than recipes: A proposal for research that does not treat useability evaluation methods as indivisible wholes. *International Journal of Human-Computer Interaction*, *27*(10), 940-970.

Woo, J. & Lim, Y. 2015. User experience in do-it-yourself-style smart homes. In Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing. ACM, 779–790.

Wright, P., & McCarthy, J. (2008, April). Empathy and experience in HCI. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 637-646).

Zaman, B., & Shrimpton-Smith, T. (2006, October). The FaceReader: Measuring instant fun of use. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles* (pp. 457-460).

Zaki J, The neuroscience of empathy: progress, pitfalls and promise. Nat Neurosci. 2012;15(5):675-680. doi:10.1038/nn.3085

Zingoni, M. (2019). Beyond aesthetics, empathy first. *The Design Journal*, *22*(3), 351-370.

Appendix 1

Test artefact context provided to participants

The Product

The session is being conducted on behalf of **Business Development Bank of Canada** (BDC). You can think of BDC as a financial institution - a bank - that is devoted to supporting Canadian entrepreneurs across the country. They provide loans, professional services, learning resources, and various community initiatives that are intended to help new businesses.

BDC is in the process of redesigning their website and is carrying out user tests to better understand how users perceive the current design - that's where you come in!

Test subject outline provided to participants

Meet the Participant

NAME: Barb Robson
OCCUPATION: Entrepreneur - Founder & CEO
INDUSTRY: Sustainable clothing manufacturing

Barb is a Canadian entrepreneur facing real constraints that hinder the execution of her business objectives. She represents a typical BDC user who would benefit from the array of services that they offer. Leading up to this user test, Barb declared her priorities for 2023:

1. Purchase more technologically advanced equipment to support their niche manufacturing processes.

2. Expand the brand by entering into international markets beyond North America.

Survey questions used to assess priority tendencies

a) Prioritizing interview topics

Assume that you have limited time to conduct your interview with Barb. Consequently, imagine that you must prioritize which tasks you would like to further investigate during the upcoming interview.

BDC has emphasized that **all 4 tasks are of equal importance**. Therefore, you should prioritize exploration based on aspects of their particular experience that are ambiguous and/or misunderstood.

Drag and drop the different tasks in order of priority
(1= highest priority)

1 TASK D - COMMUNITY SUPPORT SERVICES
2 TASK B - FINDING A LOAN
3 TASK A - FINDING AN ADVISORY SERVICE
4 TASK C - SELF-DIRECTED LEARNING RESOURCES

b) Two open-ended interview questions

If you could only ask **2** questions based on the tasks that you just observed **and** the data provided to you, what questions would you ask the user?

Interview question #1

Which task does this question pertain to?

✓
TASK A - FINDING AN ADVISORY SERVICE
TASK B - FINDING A LOAN
TASK D - COMMUNITY SUPPORT SERVICES
TASK C - SELF-DIRECTED LEARNING RESOURCES

c) Two design recommendations

To conclude, please suggest 2 UX design recommendations that BDC could implement based on insights that you derived from observing the user-test and interview responses.

Design recommendation #1

Which task does this pertain to?

✓ A) FINDING ADVISORY SERVICE
B) FINDING A LOAN
D) COMMUNITY SUPPORT SERVICES
C) SELF-DIRECTED LEARNING RESOURCES

Survey question used to assess inferential accuracy

Task A: Difficulty

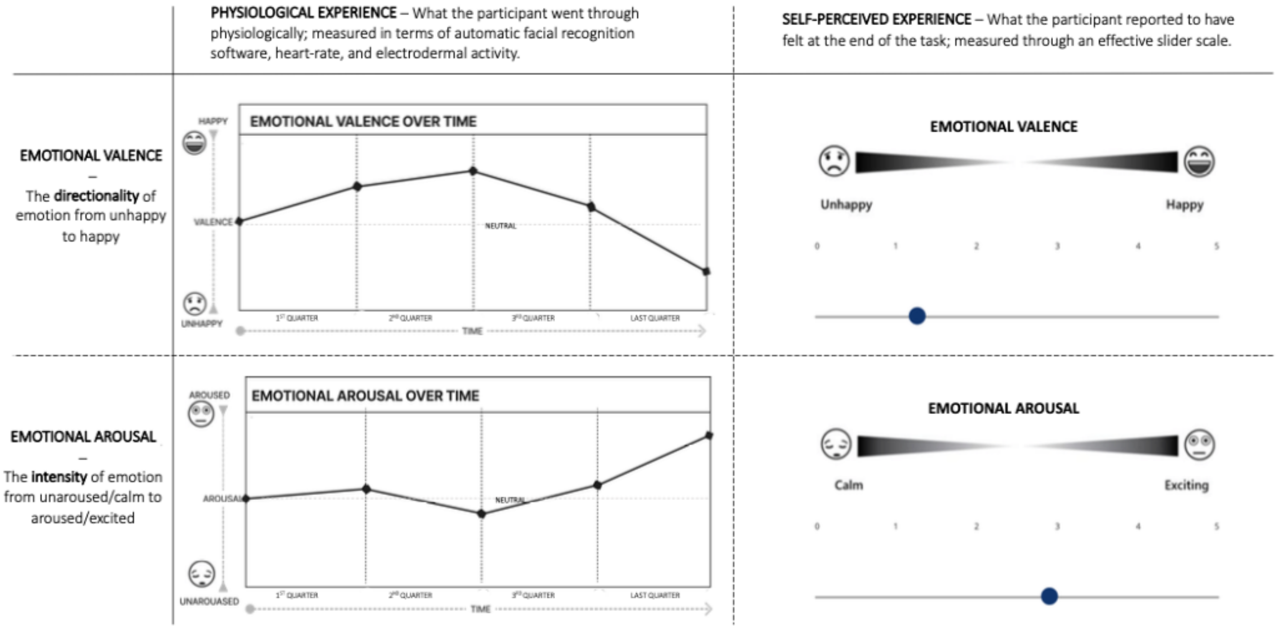
Which statement best describes the user's experience with **finding an advisory service?**

- ○ The participant did not have any difficulty completing the task.
- ○ The participant had difficulty at the beginning of this task.
- ○ The participant had difficulty at the end of this task.
- ○ The participant had difficulties throughout this task.

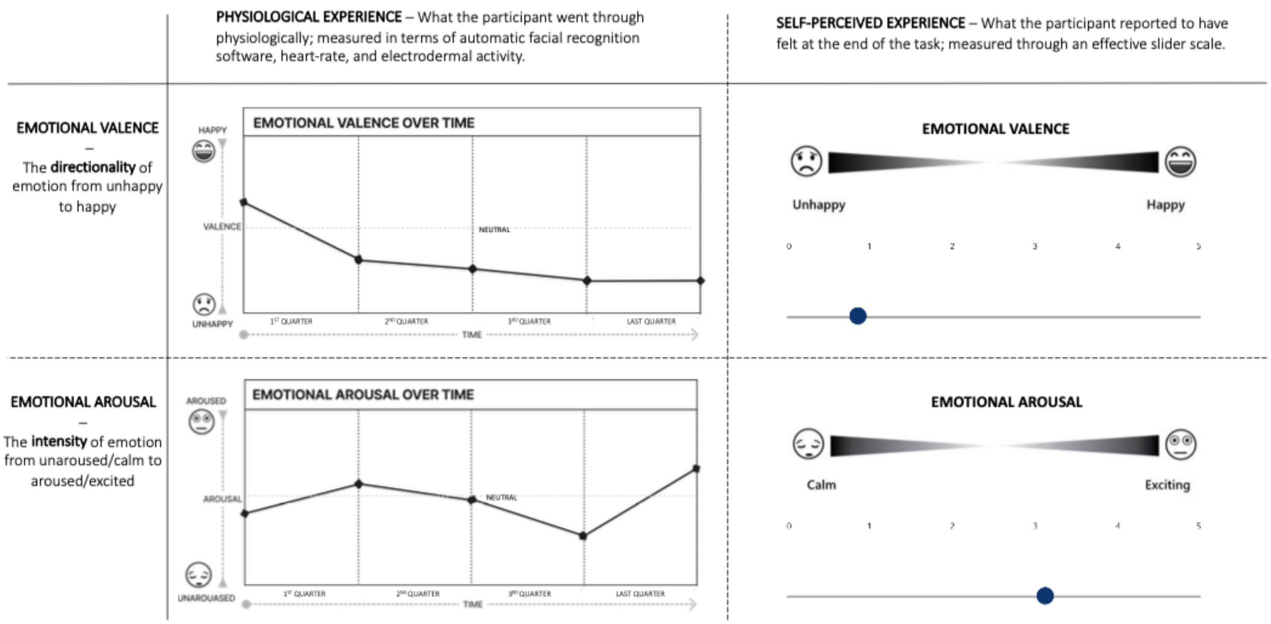
Appendix 2

Fictitious user implicit and explicit response stimuli – Condition A

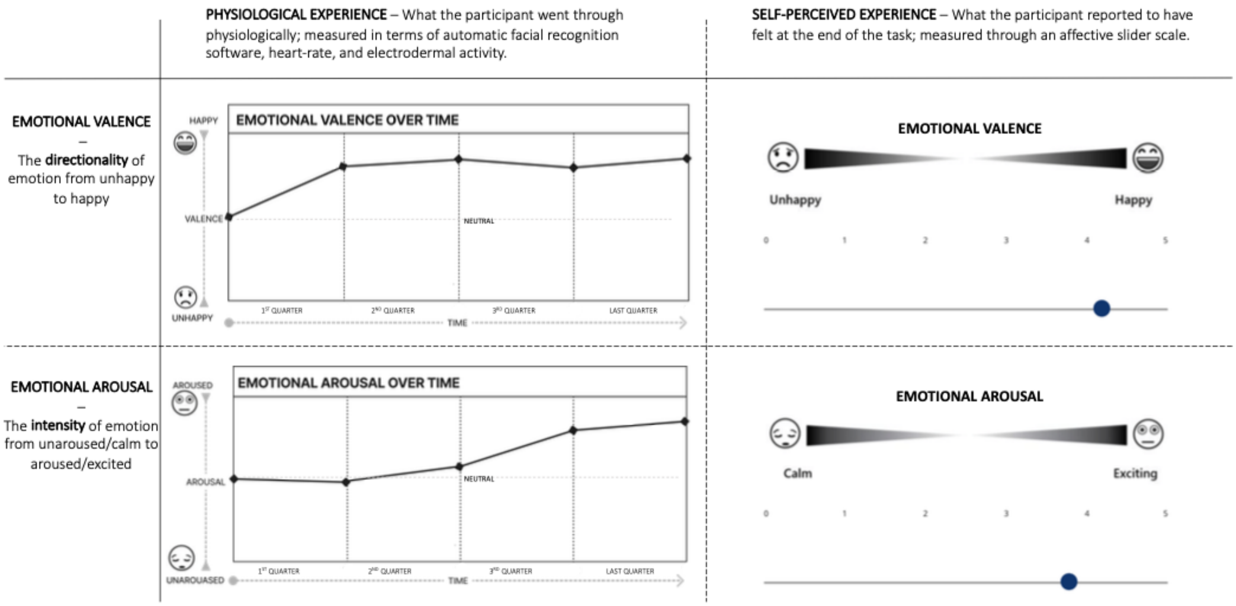
Task A – Finding an advisory service



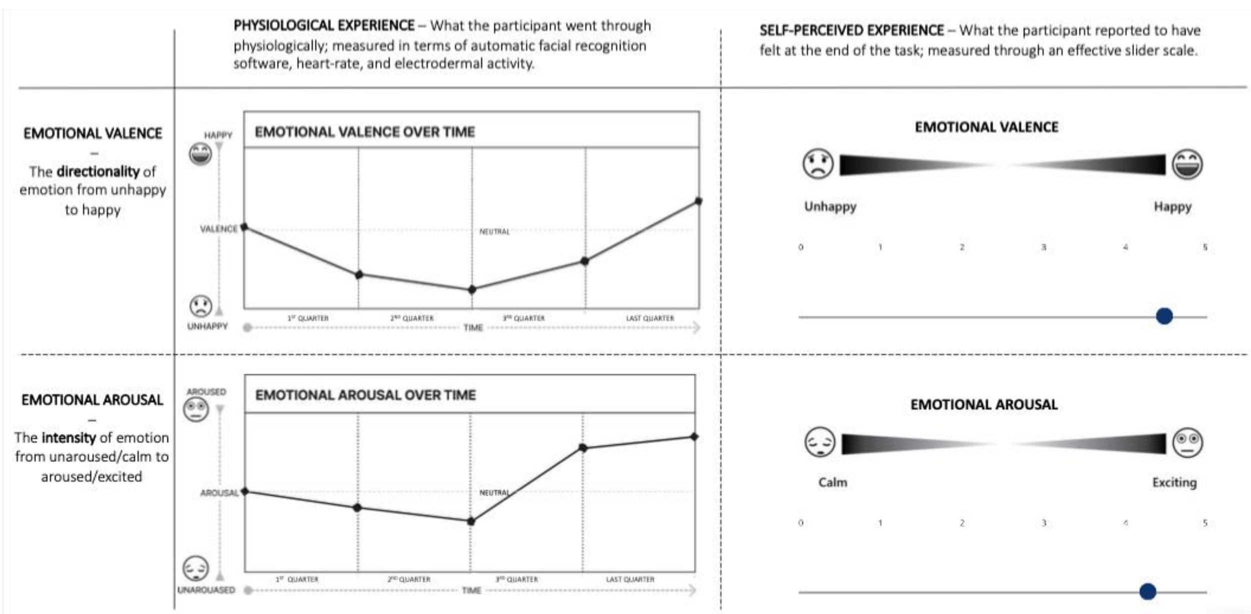
Task B – Finding a loan



Task C – Finding a learning resource

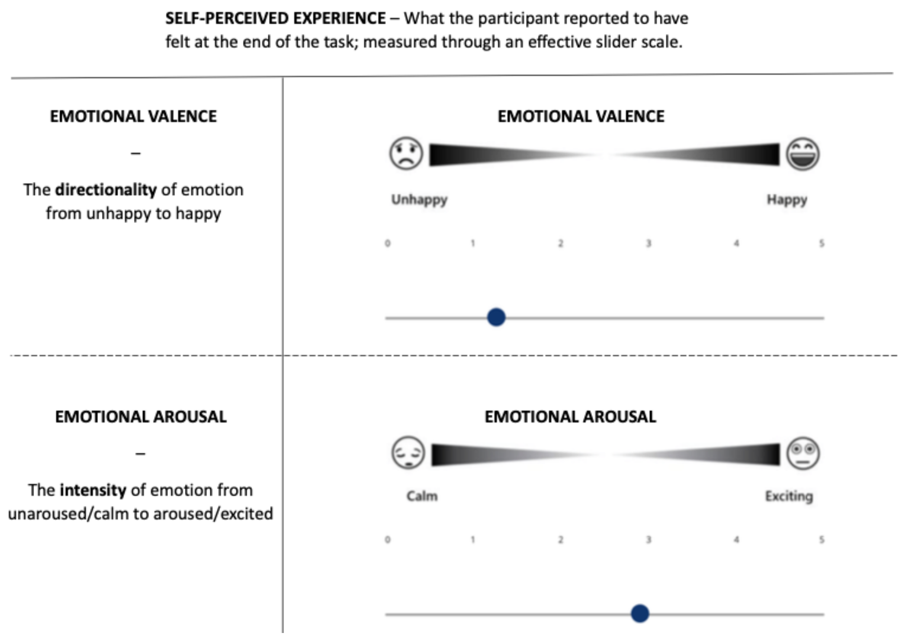


Task D – Finding community support

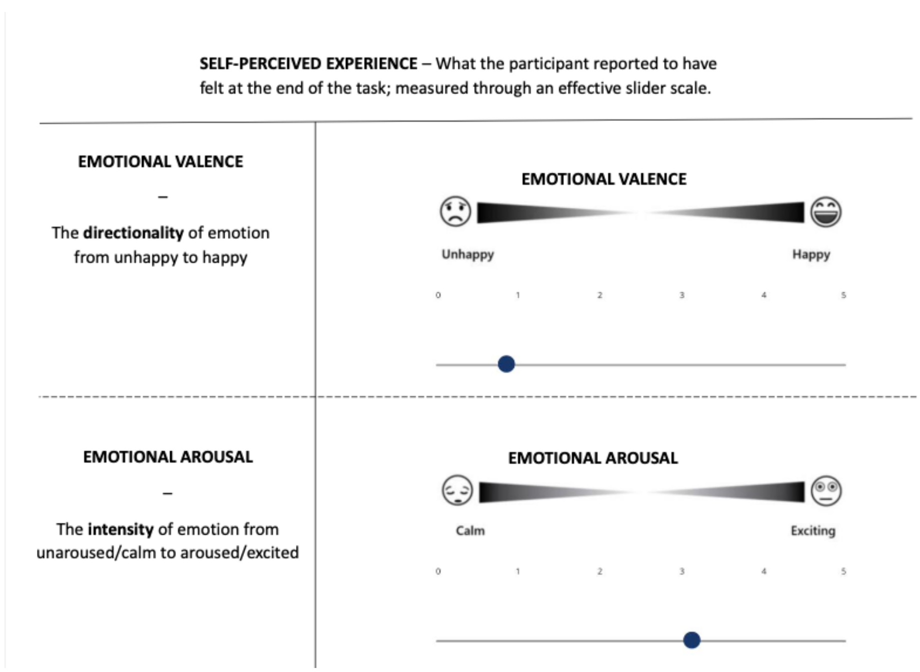


Fictitious user explicit response stimuli – Condition B

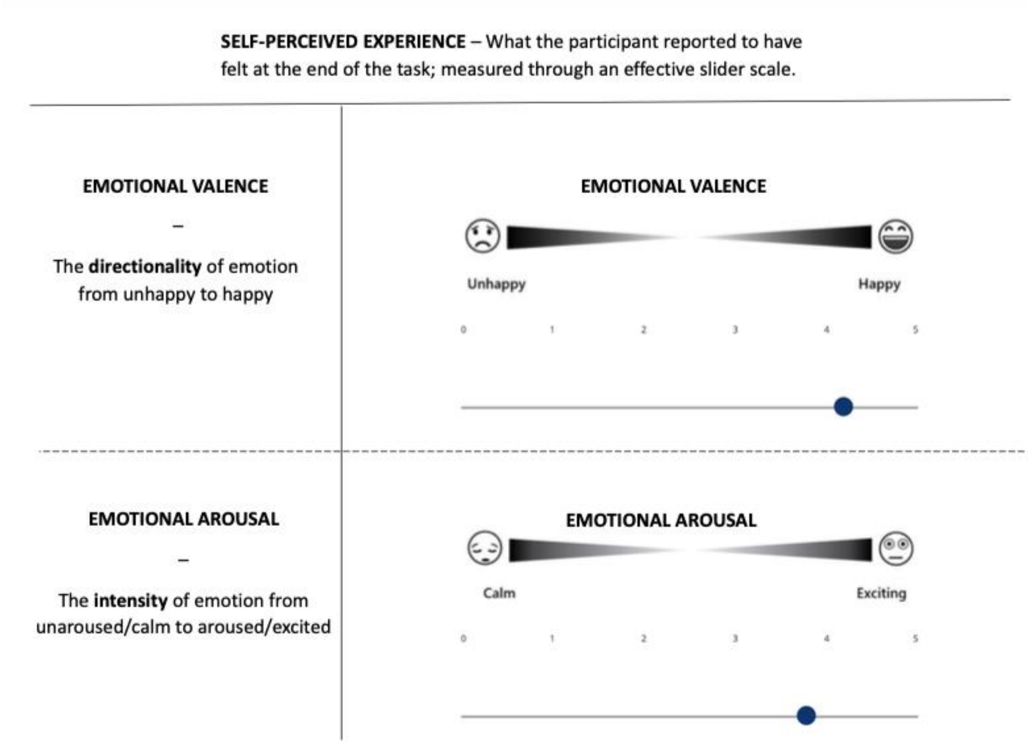
Task A – Finding an advisory service



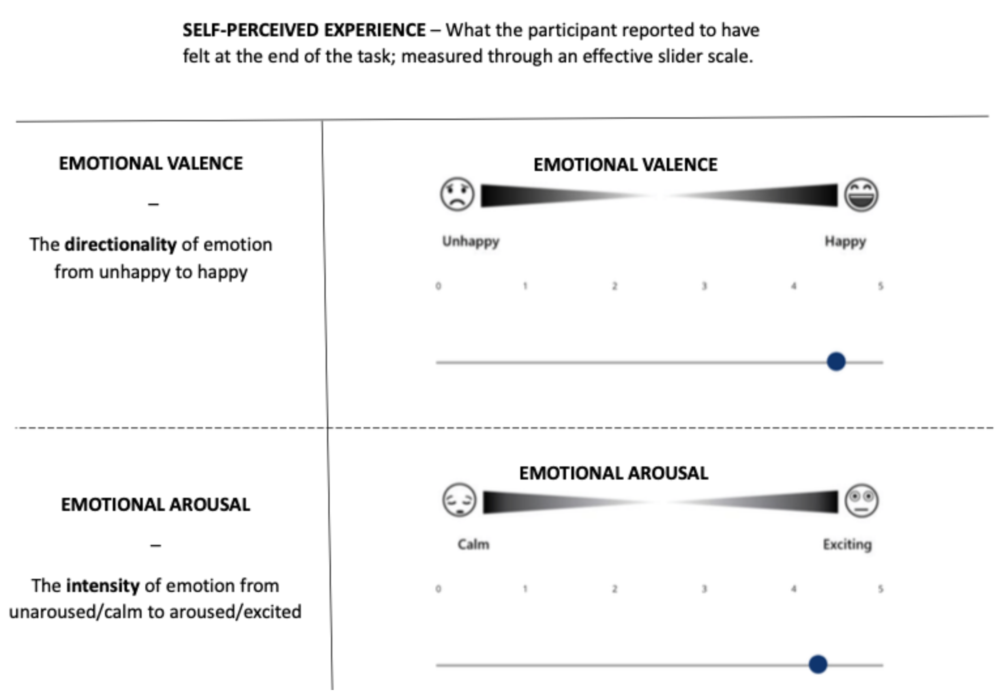
Task B – Finding a loan



Task C – Finding a learning resource



Task D – Finding a community support service



Behavioural observation stimuli

Task A video: https://hecmontreal.eu.qualtrics.com/jfe/form/SV_6R9v4luLri0E0iG

Task B video: https://hecmontreal.eu.qualtrics.com/jfe/form/SV_09waSQZKfmyB6dg

Task C video: https://hecmontreal.eu.qualtrics.com/jfe/form/SV_0AHe1QQbVhQbavk/

Task D video: https://hecmontreal.eu.qualtrics.com/jfe/form/SV_1FvfCrtJmmxiQjs