[Page de garde]

HEC MONTRÉAL

The Development and Validation of a Supervised Machine Learning Model for Automated Assessment of Online Banking User Experiences

By:

Jasmine Labelle

Sylvain Sénécal
HEC Montréal
Directeur de recherche

Pierre-Majorique Léger
HEC Montréal
Co-Directeur de recherche

Sciences de la gestion
(M.Sc. User Experience)

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

December, 2024

HEC MONTRÉAL

Comité d'éthique de la recherche

May 05, 2023

To the attention of: Pierre-Majorique Léger, HEC Montréal

Re: Ethics approval of your research project

Project No.: 2023-5392

Title of research project: Assessing the validity of rankings in the context of digital experiences

Funding source : NSERC/HEC Institutional Chair/Contract-Deloitte (R2882B/NSERC; 32-153-300-31-R2882 / Alliance industrial partners)

Title of the grant : ALLRP 575332 — 22 — NSERC Alliance Grants (ALLRP) — Improving industrial UX measures and Methods

---

Dear Mr. Léger:

Your research project has been evaluated in accordance with ethical conduct for research involving human subjects by the Research Ethics Board (REB) of HEC Montréal.

A Certificate of Ethics Approval attesting that your research complies with HEC Montréal's *Policy on Ethical Conduct for Research Involving Humans* has been issued, effective May 05, 2023. This certificate is **valid until May 01, 2024.**

Please note that you are nonetheless required to renew your ethics approval before your certificate expires using Form *F7 – Annual Renewal*. You will receive an automatic reminder by email a few weeks before your certificate expires.

If any major changes are made to your project before the certificate expires, you must complete Form *F8 – Project Modification*.

When your project is completed, you must complete Form *F9 – Termination of Project*. (or *F9a – Termination of Student Project if certification is under the supervisor's name).* **All students must complete an F9 form to obtain the "Attestation d'approbation complétée" that is required to submit their thesis/master's thesis/supervised project.**

Under the *Policy on Ethical Conduct for Research Involving Humans*, researchers are responsible for ensuring that their research projects maintain ethics approval for the entire duration of the research work, and for informing the REB of its completion. In addition, any significant changes to the project must be submitted to the REB for approval before they are implemented.

You may now begin the data collection for which you obtained this certificate.

We wish you every success in your research work.

REB of HEC Montréal

HEC MONTRÉAL

Comité d'éthique de la recherche

CERTIFICAT D'APPROBATION ÉTHIQUE

La présente atteste que le projet de recherche décrit ci-dessous a fait l'objet d'une évaluation en matière d'éthique de la recherche avec des êtres humains et qu'il satisfait aux exigences de notre politique en cette matière.

Projet # : 2023-5392

Titre du projet de recherche : Assessing the validity of rankings in the context of digital experiences

Chercheur principal :

Pierre-Majorique Léger
Professeur titulaire
Technologies de l'information
HEC Montréal

Cochercheurs :
Jasmine Labelle; Salima Tazi; Sylvain Sénécal; Constantinos K. Coursaris; Luis Carlos Castiblanco Reyes; Xavier Côté; Alexander John Karran; Brendan Scully; Jared Boasen; David Brieugne; Frédérique Bouvier

Date d'approbation du projet : May 05, 2023

Date d'entrée en vigueur du certificat : May 05, 2023

Date d'échéance du certificat : May 01, 2024

Maurice Lemelin
Président
CER de HEC Montréal

Signé le 2023-05-05 à 16:03

Abstract

This paper presents the development of a supervised machine learning model, referred to as the Viral Success Model (VSM), designed to enhance the evaluation and prediction of interface success within the banking sector. Traditional user experience (UX) research predominantly relies on explicit, self-reported measures, which come with significant limitations. To address these shortcomings, we advocate for the integration of implicit measures (IMs) in measuring lived experience and predicting UX success. In this study, we introduce a formative construct called Viral Success, a combination of task success and intention to recommend, which serves as the final measure of success for the Viral Success Model (VSM) and its ranking. The study utilizes a quantitative research design, featuring a within-subject experiment where participants engage with nine different financial institution websites to complete a search related task. This research then leverages machine learning (ML) techniques to predict the Viral Success of an interface with high accuracy. To gather the data required for the development of the VSM, participants in the study were tasked with completing a search task on nine different financial institution websites. During the search task their lived experience was captured to create different supervised machine learning models (MLM's). Based on performance, a final supervised MLM was selected and validated via click testing. Our findings demonstrate that lived experiences can be utilized to predict Viral Success during search-based tasks in a banking context, achieving an accuracy rate of 83.48%.

Keywords: user experience, information technology, machine learning, prediction, rankings, IT rankings, implicit measures, explicit measures, viral success, intention to recommend, task success, lived experience.

This thesis is structured into seven chapters. It begins with an introduction, followed by a comprehensive literature review that underscores the significance of rankings and IM's in Information Technology (IT). The third chapter establishes a theoretical framework for the VSM. The fourth chapter outlines the research methodology, leading into the development of the supervised MLM and the presentation of results. The discussion of findings is then provided,

with the thesis concluding with a thorough summary that integrates the study's contributions, practical applications, and recommendations for future research.

Résumé

Cet article présente le développement d'un modèle d'apprentissage automatique supervisé, appelé modèle VSI, conçu pour améliorer l'évaluation et la prédiction de la réussite de l'interface dans le secteur bancaire. La recherche traditionnelle sur l'expérience utilisateur (UX) s'appuie principalement sur des mesures explicites et autodéclarées, qui présentent des limites importantes. Pour remédier à ces lacunes, nous préconisons l'intégration de mesures implicites dans la mesure de l'expérience vécue et la prédiction du succès de l'interface utilisateur. Dans cette étude, nous introduisons un concept formatif appelé le Succès Viral, une combinaison de la réussite de la tâche et de l'intention de recommander, qui sert de mesure finale du succès pour le modèle VSI et son classement.L'étude utilise un modèle de recherche quantitative, avec une expérience intra-sujet dans laquelle les participants s'engagent sur neuf sites web d'institutions financières différentes pour effectuer une tâche liée à la recherche.Cette recherche s'appuie ensuite sur des techniques d'apprentissage automatique pour prédire le succès viral d'une interface avec une grande précision. Pour recueillir les données nécessaires à l'élaboration du modèle VSI, les participants à l'étude ont été chargés d'effectuer une recherche sur huit web sites d'institutions financières différentes. Au cours de la tâche de recherche, leur expérience vécue a été enregistrée afin de créer différents modèles d'apprentissage automatique supervisé (MAM). Sur la base des performances, un modèle supervisé final a été sélectionné et validé à l'aide de diverses techniques, notamment des tests de clics et un examen qualitatif. Nos résultats démontrent que les expériences vécues peuvent être utilisées pour prédire le « succès viral » pendant les tâches de recherche dans un contexte bancaire, atteignant un taux de précision de 83,48%.

Mots-clés: expérience utilisateur, technologie de l'information, apprentissage automatique, prédiction, classements, classements informatiques, mesures implicites, mesures explicites, succès viral, intention de recommander, succès de la tâche, expérience vécue.

Cette thèse est structurée en sept chapitres. Elle commence par une introduction, suivie d'une analyse de littérature complète qui souligne l'importance des classements et des mesures implicites dans les technologies de l'information (TI). Le troisième chapitre établit un cadre théorique pour le modèle VSI. Le quatrième chapitre décrit la méthodologie de recherche, qui a

conduit à l'élaboration d’un modèle d'apprentissage automatique supervisé et à la présentation des résultats. La thèse se termine par un résumé complet qui intègre les contributions de l'étude, les applications pratiques et les recommandations pour les recherches futures.

Table of contents

List of tables and figures

List of Figures

List of Tables

List of abbreviations and acronyms

- IT — Information Technology
- UX — User Experience
- IS — Information Systems
- NPS — Net Promoter Score
- VSM — Viral Success Model
- VSI — Viral Success Index
- VSR — Viral Success Ranking
- TVSR — True Viral Success Ranking
- ML — Machine Learning
- MLM — Machine Learning Model
- REB — Research Ethics Board
- TTFC — Time to First Click
- EDA — Electrodermal Activity
- EKG — Electrocardiogram
- STD — Standard Deviation

Acknowledgements

This thesis would not have been possible without various people whom I would like to acknowledge. First and foremost, I am very grateful to have been able to collaborate with Deloitte for my thesis project, the opportunity helped me grow immensely professionally. I am very grateful to the NSERC and Prompt for the master's scholarship. Their financial support was invaluable in facilitating my studies.

I would first like to thank my directors, Sylvain Sénécal and Pierre Marjorique Léger for offering me the opportunity to lead the project with Tech3Lab. Throughout my entire thesis writing process, they supported me and motivated me. I would also like to thank Shang Lin Chen and the whole Tech3Lab staff for the support provided to me during the data collection process and throughout my thesis writing. I would also like to thank Chenyi Huang, Alexis Perrault, and Erika Neveau who led the entire data collection process with me over the span of three months. Large scale data collection would not have been possible without the entire team.

I would also like to thank my whole family, most importantly Michael Labelle who was by my side from day 1 through this very challenging path. I could not have asked for a better person to help me navigate this journey and all the hard obstacles that came with it. A special thank you to Pierre-Michel & Katherine for their support as well, especially though it was the end of my master's journey.

I would also like to thank all my close friends, Kathryn Pantemis, Stefi Lague, Will Ashley, Alessia Durante, Vasiliki Kapoglis, Rachel Cosby, Laura Ciamarra, Samuel ElHarrar, Michelle Partidas and Maya Lehata who motivated me, pushed me and always listened to me throughout the ups and downs of this path. Thank you all for the invaluable support and your belief in me.

Chapter 1: Introduction

Like players in a high-stakes game, businesses today operate in a competitive arena where their decisions are often driven by the pursuit of a top spot on the leaderboard. Published rankings, which have grown in importance over the past decade, play a pivotal role in shaping this environment. The reason being, buyers and decision-makers, overwhelmed by an abundance of choices, rely on rankings as essential tools to streamline their decision-making processes, especially in markets where evaluating the quality of entities is challenging and subjective (Rindova, Martins, Srinivas & Chandler, 2017). In such contexts, rankings become valuable for reducing information asymmetries, with decision-makers favoring top-ranked options (Rindova, et al., 2017).

Researchers across various fields have extensively examined the impact of rankings, particularly their influence on stakeholder responses (Chun & Larrick, 2022). The effects of rankings are profound, affecting multiple aspects, such as strategy development, reputation, and performance (Rindova et al., 2017). This raises key questions: Who decides how a firm is ranked, and what factors contribute to these rankings? Leading organizations, known as "information intermediaries," such as Gartner, Surviscor, Forrester, and Leger Marketing, play a pivotal role in creating these rankings. They are responsible for gathering, analyzing, and disseminating the comprehensive data required to develop and present rankings. However, the substantial influence and potential consequences of rankings brings concerns about the methodologies and choices used by these information intermediaries throughout rank development (Rindova et al., 2017).

In the 2017 study (Rindova et al., 2017) developed a framework called the "Integrated Model of Research on Rankings". The model was built to illustrate how current research interprets the connections “between audiences and ranked organizations as shaped by rankings”. A section of their model particularly emphasizes that rankings emerge from the informational demands of consumers as well as the reputation and status of organizations (Rindova et al., 2017).

Additionally, it highlights the significant impact that the design features within rankings have on their quality and usefulness, and additionally an entity's relative position in the ranking (Rindova et al., 2017). This thesis will contribute to the following model and the field of Information Systems (IS) research by exploring the current state of ranking methodologies in Information Technology (IT), identifying their limitations and proposing a novel framework for IT rankings.

Much of the existing literature on rankings (Rindova et al., 2017; Chun & Larrick, 2022; Ursu, 2018) focuses on the process of rank production, emphasizing the influence of information intermediaries and the ethical concerns associated with their role in a wide range of industries. Researchers highlight a key point: that the practice of measurement, the statistical techniques employed, and the selection of indicators are critical for ranking methodologies to accurately assess final outcomes (Rindova et al., 2017).

However, there is a lack of research when it comes to contextualizing and developing strong ranking methodologies in the context of IT. When it comes to the field of IT, the overall user experience (UX) is a vital component to be considered. According to the ISO 9241-110:2010 definition, UX can be defined as a "person’s perceptions and their responses resulting from the use or anticipated use of a product, system, or service”. It can be measured through various constructs related to human emotional reactions, usability (e.g., efficiency), and user perception (e.g., satisfaction) (Hussain, Khan, Hur, Bilal, Bang, Hassan, Afzal, & Lee, 2018). Thus, it encapsulates the holistic experiences users have with technological systems (Kruger, Rendani & Gelderblom, Helene & Beukes, Wynand, 2016; Koonsanit & Nishiuchi, 2021).

Researchers such as (Hussain et al., 2018) and (Cuviller, Léger & Sénécal, 2021) have identified that the subjective component of UX can make assessment challenging. The conventional methods of UX assessment typically depend on self-reported measures, usability studies and observational techniques. However, these methods fail to capture the true emotional experience of users (Hussain et al., 2018). Their main limitation lies in their inability to continuously capture a user's state as they utilize a given system (Guinea, Titah, & Léger, 2014). Rather, subjective methods are highly dependent on how a user decides to recall an event, making them subject to bias (Hussain et al., 2018).

Recognizing the importance of capturing brief and momentary emotional states, researchers such as (Allam, Hussin & Dahlan, 2013) have suggested prioritizing the evaluation of lived experience, which encompasses how users interact with a system in real-time. With the advent of implicit measures (IM's), such as psychophysiological tools in the field of IS, there is a growing emphasis from researchers on utilizing these tools to measure lived experience (Guinea et al., 2014; Maisto, Slaby & Actis-Grosso, 2023). These tools can capture user reactions in real-time, allowing for the detection of automatic and unconscious responses that happen outside of a user's conscious awareness (Guinea et al., 2014). This thesis is therefore expected to make a significant impact by addressing a gap in IT ranking methodologies. It introduces a novel ranking approach that incorporates IM's rather than relying solely on self-reported subjective data, thereby tackling a critical issue in IS measurement (Pavlou & Dimoka, 2010).

Now that we have highlighted the importance of proper IS measurement techniques, this thesis aims to additionally introduce a novel, formative construct called "Viral Success". The novel construct, composed of Intention to Recommend and Task Success, serves as the success measure and foundation of the ranking model proposed, the Viral Success Model (VSM). We aim to introduce this construct as a success metric in order to address the limitations present in solely relying on metrics such as NPS, as a success metric. In essence, we stipulate that introducing the novel formative construct of Viral Success, will provide stakeholders with a comprehensive understanding of how users feel (via their intention to recommend) and how they act (task success) when interacting with a digital artifact. Thus, in this master thesis, we aim to explore how the novel construct of Viral Success can be more insightful than either metric alone. We stipulate that this combination will ensure that stakeholder decision making is balanced, considering both word of mouth intention and immediate user experience via task success. This brings a first research question:

> **Research Question 1 (RQ1):** Can a unidimensional index effectively capture both task success and the intention to share this success in the context of user experience?

Despite the advantages of introducing the novel construct of Viral Success, it does have a handful of limitations which we are aware of. There are limiting factors that come from adding two

distinct variables into a single construct which then serves as the outcome variable in a ranking. First, formative constructs combine multiple variables that can obscure the distinct contributions and interactions of each variable (Petter, Straub, Rai, 2007). In addition, there is context sensitivity, meaning the relevance and impact of the combined variables may vary across different contexts of user groups and industries. What constitutes viral success in one scenario — e.g., banking interfaces — might not hold true in another, limiting the generalizability of the construct.

Building on the following, different studies (Chromik, Lachner, Butz, 2020; Koonsanit & Nishiushi, 2021) showcase the benefits of machine learning in the UX design process and in the prediction of user experience metrics like user satisfaction. Given the increasing popularity of ML techniques, our study is also based on this axis: Exploring how the combination of lived user experience and machine learning can predict the proposed novel construct of Viral Success.

We therefore raise a second research question:

> **Research Question 2 (RQ2):** To what extent can lived experience data be utilized to train a supervised machine learning model for accurately predicting Viral Success in the banking sector?

This thesis is organized as follows: Chapter 2 begins with a review of the literature on rankings, focusing on their efficacy, production processes, and the involvement of various stakeholders, as well as the consequences of the frameworks employed. We then contextualize this literature within the field of IT, addressing the limitations specific to IT rankings, particularly in the context of ranking digital experiences. Next, we emphasize the importance of incorporating IM's, to enhance the assessment of user experience outcomes and explore how new technologies can be integrated into IT ranking methodologies. We also introduce Viral Success as a formative construct that should be assessed. In Chapter 3, we discuss the benefits and challenges of introducing this novel construct. We delved into the study's conceptual framework, examining how different IM's can be utilized to measure and predict Viral Success. In Chapter 4, we outline the methodology used to collect data for our supervised machine learning model (MLM). In

Chapter 5 and 6 we showcase the results and discussion, which includes the process of building and selecting the final version of the Viral Success Model (VSM).

Student's Contribution and Responsibilities in the Completion of this Thesis

Table 1 - Student's Personal Contribution Table

Steps	Contribution
Research Question Definition	Definition of research questions and issues - **75%** ● Contextualization of the problem developed in collaboration with an industrial partner. ● Translation of the industrial partner needs into a research question and definition of the problem.
Literature Review	Conducting the literature review - **80%** ● Identification of the existing literature on the subject. ● Help of co-authors identify research topics ● Definition of scales and measurements to be used in the study ● Laboratory assistance with physiological tools and use of established resources. Drafting of the literature review - **100%**
Application for research ethics	Drafting of REB application and subsequent - **90%** ● Research laboratory team reviewed application prior to submission
Experimental Design	Creation n of experimental design and test protocols - 80% of total costs ● Experimental protocol design - 80% ● The research laboratory team recommended a protocol for using the physiological tool. ● Organizing the data collection room at the partner's site - 0%

	• Set up equipment for data collection
Participant Recruitment	Development of the recruitment questionnaire **- 75%** • The recruitment questionnaire was drawn up in collaboration with the laboratory's research team. Recruitment and participant management **- 90%** • Recruitment carried out by the institution's laboratory panel and personal connections. • Participant data were anonymized by the research laboratory. • Potential participants were filtered by the panel according to the inclusion and exclusion criteria. • Selected participants were contacted by e-mail with the help of a research assistant.
Pre-test and data collection	**Responsible for operations during pre-test - 100%** • All pretests were led by the researcher. **Responsible for operations during the data collection - 80%** • Present during most of the data collection process.
Extraction and transformation of the data	**Data extraction and formatting in preparation for analysis - 100%**
Data Analysis	**Statistical Analysis - 75%** • Support from the research laboratory team and statistician in data processing
Copywriting	**Writing the articles for the dissertation - 100%** • Autonomous writing with corrections and improvements by co-authors

Chapter 2 : Literature Review

Chapter 2 consists of a literature review that will explore the various roles of information intermediaries, both generally and specifically within IT. It will examine the rationale behind rankings and evaluate their effectiveness. The review will then categorize and provide examples of various rankings, shifting focus to specific IT rankings. This includes researching whether any current rankings utilize NPS as a success measure. Lastly, the literature review will address the limitations and biases present in current ranking methodologies, highlighting the gaps present.

2.1 The Purpose of Rankings

For many consumers and decision-makers, the surge in available options poses a challenge known as "choice overload," which essentially complicates the decision-making process (Scheibehenne, Greifeneder & Todd, 2010). Recognizing the complexity of our information-rich environment, decision-makers actively seek tools to alleviate this uncertainty and help them make optimal choices (Chun & Larrick, 2022). One possible tool that consumers and stakeholders utilize to help ease feelings of choice overload is rank information (Quaschning, Vermeir, & Pandelaere, 2011). For this master's thesis, I will build off of the definition established by (Chun & Larrick, 2022), wherein a ranking is described as a simple tool to classify entities (such as individuals, products, services, or organizations) according to a specific attribute or metric, facilitating comparative assessment. Researchers including (Chun & Larrick, 2022) and (Zitek & Tiedens, 2012) agree that rankings encapsulate hierarchical relationships by explicitly conveying the relative positioning of each option in contrast to one another. Furthermore, (Chun & Larrick, 2022) highlight that the effectiveness of rankings lies in their ability to streamline decision-making, which makes them popular in numerous contexts, whether that be for products, businesses or organizations.

2.2 The Efficacy of Rankings

Researchers including (Rindova et al., 2017) and (Van Vught & Westerheijden, 2010) both emphasize that rankings are pivotal in decision-making, particularly when they function as

comprehensive tools that consider multiple dimensions of interests to stakeholders. This section of the literature review will therefore highlight the impact of rank information on both consumer and producer decision making.

2.2.1 Rank information and Consumer Decision-Making

From the Nielsen Ratings, Brand Finance Global 500, and Fortune 500 to Billboard Charts, rank information is present in everyday life for consumers (Quaschning et al., 2011). Researchers, (Qasching et al., 2011) and (Ursu, 2018) who have explored the ranking industry and its subsequent consequences have established that a top-ranked option also known as "the winner", and the bottom-ranked option, "the loser", assist consumers in determining which direction to pursue thereby reducing the risk and likelihood of a suboptimal outcome. Researchers (Qasching, et al., 2011) and (Chun & Larrick, 2022) suggest, the highest-ranked option in a ranking naturally instills a sense of confidence within consumers, heightening their likelihood of associating themselves with said company, in comparison to those positioned lower. Consumers who feel confident with their choices are inclined to act upon this. A study done by (Gartner, 2019) highlights that confident consumers can "spend up to 2.6 times more". In addition to this, (Rindova et al., 2017) suggest that the proliferation of social media and user-generated content has exponentially streamlined the reach of rankings and therefore they have become direct influencers to consumer purchasing and their adoption decision (Haans & Rietveld, 2024).

2.2.2 Rank Information and Stakeholder Decision-Making

Researchers such as (Doshi, Kelley, & Simmons, 2019), (Koski, Xie & Olson, 2015), and (Zitek et al., 2012) emphasize that social pressure is conveyed through information; rankings function as a means of transmitting that information. Consequently, the importance of rankings extends beyond consumers to encompass firms and investors engaged in business itself. The standing of what we will refer to as an 'entity' (e.g. a product, service, or business) within a ranking, whether it be first or last, significantly influences how stakeholders perceive the firm. That being, rank information displays whether an entity is legitimate and deserving of 'esteem' (Sharkey, Kovács & Hsu, 2022). Scholars who have assessed the effects of rankings have emphasized that rankings cause decision-makers to set goals and adapt their structures in response to the indicators being

used and can even cause poorly ranked entities to emulate the strengths of the highest-ranked option (Doshi et al., 2019; Chun & Larrick, 2022). As an example of this scenario: In a ranking system that evaluates states using a set of indicators, governments are likely to care about the opinions of voters, business organizations, international investors and their own reputation (Doshi, et al., 2019). Given the strong desire to enhance long-term growth, this drives leaders to "compete by becoming versed in effective ranking" (Doshi et al., 2019).

2.3 The Influence of Information Intermediaries

Having set the efficacy of rankings and their impact, the next section dives into understanding the role of the individuals responsible for crafting such rankings. Although the literature remains minimal when it comes to "Information Intermediaries" (e.g. the individuals involved in crafting rankings), the next part of this literature review will establish a clear definition of an information intermediary. In the literature reviewed by (Sharkey et al., 2022) and (Rindova, 2005) it is established that an information intermediary was first considered an expert with "extensive subject matter expertise". However, an information intermediary has grown to be broader in scope and function. For this master thesis, I will build off of the definition shared by (Sharkey et al., 2022) where an information intermediary is a third-party evaluator that consumers turn to choose their products and services, or more specifically it can be defined as "an entity that occupies the interface between consumers and producers". In sum, information intermediaries, the source, are responsible for creating rankings, the outcome. Given the strong impact of their evaluations, on both consumer and producer decision making, (Haans et al., 2024) highlight that it is critical to understand the different types of information intermediaries and what drives their evaluations. Building off the definition of the three ideal intermediary types classified by (Sharkey et al., 2022), I have highlighted different examples of each and contextualized it to suit the context of UX.

2.3.1 Expert Critics

The first type of information intermediary highlighted by (Sharkey et al., 2022) is the expert critic. This type of information intermediary is considered to be a highly influential professional

employed by an organization that is tasked with evaluating entities. The legitimacy of these expert critics is established through technical training, education, or accumulated experience and reputation (Sharkey et al., 2022). In the context of UX, this could be a seasoned UX researcher, designer, or specialist who has received formal training or who has gained enough experience in the field to build the knowledge needed to conduct an expert review. There are many benefits to the expert-based evaluation, being that it is affordable, quick and easy to conduct (Lallemand, Koenig, & Gronier, 2014). However, there are challenges inherent in this form of evaluation. The most common type of expert-based evaluation used by HCI practitioners is the heuristic evaluation or the cognitive walkthrough. Despite their adherence to usability principles, industry standards, and protocols, research studies in the field of HCI have indicated that these types of expert evaluations are highly susceptible to subjectivity (Lallemand et al., 2014; Arhippainenm 2013). First, researchers (Lallemand et al., 2014) have shown that UX experts will primarily approach UX from a positive perspective rather than a negative one. Due to this bias and the subjectivity of the evaluations, many UX designers tend to have issues with evaluator recommendations (Kruger, 2016). Furthermore, research done by (Lallemand et al., 2014) on trends and changes in UX highlighted that an expert is usually, not a real user. The effectiveness of their evaluations can subsequently vary depending on the experts' experience and knowledge of the area being evaluated. An expert conducting an evaluation does not fully adopt the perspective of an actual user when conducting evaluations and their inspections tend to have high variability and limited reliability (Lallemand et al., 2014). Therefore, although an expert evaluation may be cost-effective and efficient, there is room for cognitive biases that affect evaluations (Negro & Leung, 2013).

2.3.2 Organizations

The second type of information intermediary highlighted by (Sharkey et al., 2022) is the organization. Organizations are tasked with building rankings, ratings, and certifications that showcase the quality and performance of entities in comparison to each other, based on set criteria (Sharkey et al., 2022). These organizations often employ a standard methodology, they collect and analyze data, which is then transformed into easily comprehensible formats for

audiences to view (e.g. a ranking or rating) (Sharkey et al., 2022). In organizational assessment, the use of “statistical analysis, formulae with weighted components, and tabular presentations” is a key feature of evaluations (Sharkey et al, 2022). These elements combined show the scientific rigor that follows the assessments led by organizations.

If we contextualize this example to IT, an organization can be one such as Leger, a Canadian-owned polling and marketing research firm that regularly creates rankings that highlight digital offerings which surpass industry standards. For example, Leger creates annual “WOW Digital Rankings'. Similarly, Surviscor, a prominent leader in North America, specializes in the analysis and ranking of digital customer experiences offered by Canadian service firms. Surviscor will regularly provide precise assessments of both individual firms and industry offerings based on a proprietary scoreCard methodology (Surviscor, 2024). Although there is greater standardization present in evaluations done by organizations, there are two main drawbacks that are important to highlight. It is stated by (Chatteri & Toffel, 2010; Sharkey & Bromley, 2015; Sharkey et al., 2022) that the determinants of a ranking or rating provided by an organization can come to be dependent on the analyst’s interpretation that comes from both objective and subjective sources of data. Therefore, (Sharkey et al., 2022) highlight that the greater standardization of the methods used for these evaluations does not necessarily lead to an increased accuracy. Additionally, there are concerns about potential conflicts of interest with the close relationships between the ranked entities and the organizations leading rankings. The misalignment of motivations, combined with a need for revenue, may bias the evaluative outcomes provided by organizations (Sharkey et al., 2022).

2.3.3 Online Review Aggregators

Lastly comes online review aggregators, which have grown in popularity. Online review aggregators serve as effective platforms for sharing perspectives and experiences (Sharkey et al., 2022). Utilizing a pre-specified framework, ratings are curated by the intermediary, which determines the metrics to emphasize and will then synthesize these ratings into a comprehensive summary for internet users to view. Spanning diverse sectors, from travel (e.g., TripAdvisor) to books (e.g., Goodreads), entertainment (e.g., IMDb), and even finance (e.g., MoneyGenius).

Research done by (Yin et al., 2021) highlights that the richness of reviews, encompassing the detail and variety of the information provided, has been shown to correlate positively with consumer purchasing. Essentially, online review aggregators have been supported in their ability to serve as a cognitive shortcut for consumers, who may not have the capacity to process extensive individual reviews (Shen, Shan, & Luan, 2018).

The drawbacks of online review aggregators as highlighted by (Sharkey et al, 2022; Chua & Banerjee, 2014; Shen et al., 2018) lies in the inherently unstructured and informal nature of its reviews, which can vary significantly depending on the individual providing the review. Put simply, a customer's perception of their experience may be influenced by factors such as their mood, previous ratings, or external circumstances that are simply beyond the control of the service provider. This variability leads to divergence among reviews, sometimes even resulting in contradictions (Sharkey et al., 2022). Furthermore, researchers (Huang, Boas, Zhao, 2023) and (Shen et al., 2018) highlight that ratings provided by online review aggregators can prove to be undependable due to the influx of fake reviews, which can essentially tarnish the perceived quality of an entity and damage the overall reliability of the review.

2.4 Examples of Rankings

From universities, athletes, and artists to hospitals, businesses, hotels, restaurants, and nations (Ringel & Werron, 2020) who explored the history of rankings stated that just about anything can be ranked. In support of this (Chun et al., 2022) highlights that rankings are even prevalent in sectors where individual preferences can vary extensively, such as political views or music tastes.

In academia alone, (Rindova et al., 2017) has identified over 50 different university rankings. For example, (Times Higher Education, 2023) created the World University Ranking, which encourages universities to compete by ranking institutions that showcase their quality based on indicators like teaching, research environment, research quality, international outlook, and industry. Similarly, the research executed by (Ringel et al., 2020) identified that the Human Development Index highlights nations' achievements in human development. The indicators present in the ranking can be motivating for these nations to improve their “GDP, healthcare, and

education systems” to enhance their standings, often at the expense of others (Ringel et al., 2020).

Business rankings, such as Fortune's 500 Most Admired Companies, evaluate firms based on indicators like quality of management and talent attraction, significantly impacting firm reputation (Fortune, 2023; Rindova et al., 2017). Product and service rankings, exemplified by consumer reports, assess consumer goods or services based on quality, customer feedback, and other indicators (e.g., Runner’s World Best Running Shoes of 2024) (Rindova et al., 2017). Additionally, online review aggregators like “TripAdvisor” rank entities based on consumer satisfaction metrics or purchase rates (Rindova et al., 2017; Ursu, 2018). While the examples we’ve provided only scratch the surface of the various types of rankings available to the public, they highlight the significance of indicators and raise questions about the quantification methods used to compare ranked entities and build rankings.

2.5 Rankings in IT

Broadly speaking, we’ve established that rankings serve as a means of evaluating the performance of entities based on specific indicators. As (Ringel et al., 2020) highlight, this process not only compares performance but also incentivizes entities to improve and strive for better outcomes. However, a significant gap remains in the literature regarding the ranking methodologies applied in IT. In the next part of this literature review, we will attempt to explore for ourselves the different types of IT rankings found and identify the role of information intermediaries involved in the methodology employed in such rankings.

2.5.1 IT & Online Customer Experience Rankings

One prominent category of ranking within IT is online customer experience rankings. These types of indexes and their rankings are created to measure the overall quality of online customer experiences. We’ve found that information intermediaries (e.g. organizations) such as “Leger” and “J.D. Power” annually release digital customer experience rankings, each providing valuable insights into how online platforms perform based on set indicators. For instance, each year Leger publishes the "WOW Digital: The Best Online Customer Experience in Canada" report, which

evaluates the online shopping experience from browsing to delivery, using key performance indicators (Leger, 2024). This study is anchored by the “WOW Digital Index,” a ranking index developed by Leger that assesses 23 dimensions of online customer experience, including visitor profile, performance metrics, online irritants, and more. The broad categories covered by WOW Digital Index include visual appeal, product offerings, customer assistance, overall experience, transaction process, delivery, and post-purchase services (Leger, 2024). After a review of the methodology used by the organization, the index is derived using explicit measures, where an online survey is sent to Canadians who have visited each website within the past 12 months. The WOW Index is then used to create a final ranking that is published for the public.

Similarly, “J.D. Power”, a consumer insights and data analytics firm, releases sector-specific online customer experience rankings, such as the "2024 U.S. Banking Mobile App Satisfaction Study." Their study measures customer satisfaction with online experiences based on indicators like navigation, speed, visual appeal, and the quality of information or content provided (J.D Power, 2024). The methodology and index used to craft this ranking is collected from 17,843 customers within the targeted sector over one month (J.D. Power, 2024). The outcome is an Overall Customer Satisfaction Index Ranking, where firms are evaluated on a 1,000-point scale.

2.5.2 Vendor Assessment Rankings

Another significant category of rankings in IT is vendor assessments. These assessments are most commonly led by market research firms, where the positioning of ranked firms is demonstrated graphically. Research done by (Taherdoost & Brard, 2019) has established that these types of assessments and their subsequent rankings are critical in the success of an organization. They reduce purchase risk and maximize overall value to a purchaser (Taherdoost, Brard, 2019). Prominent information intermediaries, including Gartner and Forrester, have developed their own methodologies that guide vendor assessments. For example, Forrester employs the “Forrester Wave Methodology”, while Gartner uses the Magic Quadrant. These reports provide invaluable insights into the top providers within specific industries, enabling stakeholders to grasp key market drivers, their impacts, and competitive positioning (Forrester, 2024). For instance, the Forrester Wave Methodology, visually represents how firms are positioned within the market

based on industry-specific indicators. As an example, in the Forrester Augmented BI Ranking, companies are evaluated on broad categories such as current offerings, strategy, and market presence, with specific indicators like conversational BI, innovation, and performance playing a crucial role. Similarly, Gartner's Magic Quadrant offers a graphical depiction of how technology providers align with a company's business goals, needs, and priorities. The Gartner Magic Quadrant Research Methodology Categorizes vendors into four segments—Leaders, Visionaries, Niche Players, and Challengers—based on their ability to execute and the completeness of their vision (Gartner, 2024). These assessments are crucial in guiding decision-makers in choosing their business partners.

2.5.3 UX Awards in IT

The next kind of IT ranking we've identified are rankings in the form of awards. Essentially, awards are a notable form of ranking given their ability to recognize the exceptional work of firms and IT service providers across various industries based on predefined categories. Information Intermediaries (e.g. organizations) such as the Nielsen Norman Group, the International Academy of Digital Arts and Sciences, or the International Design Center Berlin (IDZ), are examples of information intermediaries that issue these awards. The awards given are based on judging criteria that assess the quality and performance of digital products and the user experiences they deliver. For example: The UX Design Awards awarded by IDZ are based on judging criteria including "relevance, empowerment, innovation, outcome and business value, holistic thinking and user-centric approach, design and experience quality" (UX Design Awards, 2024). These awards are typically determined by a jury that consists of independent experts (e.g. expert critics) who have many years of experience within their field. As another example, the Nielsen Norman Group annually publishes the Intranet Design Annual, a detailed report that highlights the Top Intranet Designs and explains the selection criteria for each winner (Pernice, Caya, Rosala, Kaley, 2020). Winning this kind of award can have significant positive implications for a company, including enhancing its reputation, attracting more users, and demonstrating its commitment to providing high-quality digital experiences.

2.5.4 User-Generated Review Rankings in IT

The final type of ranking we've observed in IT is user-generated review rankings. Unlike rankings conducted by market research firms, analysts, or expert critics, these rankings are based on feedback and ratings provided by actual users of software and IT products. This approach offers a direct perspective on aspects such as product performance and usability, given they are completely grounded in real-world experience. A prime example of user-generated review rankings is G2 Crowd, which compiles rankings based on user feedback and ratings (G2, 2024). For instance, G2 Crowd features a ranking of the best customer experience software, where users rate products on a scale of 1 to 5 stars and can leave detailed commentary on their experiences. Another example is TrustPilot, which aggregates customer reviews and ratings for a wide range of businesses and services across various industries. These user-generated rankings provide valuable insights directly from those who interact with the products daily.

2.6 Rankings Utilizing NPS

In a survey done amongst 700 consumer experience professionals, the Net Promoter Score (NPS) was identified as one of the most frequently used metrics for understanding users (Cuvillier et al., 2021). Furthermore, in a research report led by Customer Gauge on NPS and CX Benchmarks, (Dorrell, Woerner, 2020) stated that firms with mature NPS programs often make it a habit to compare themselves to the top-performers within their industry. Given the importance of the metric across industries, and the use of the metric for comparison, here are some applications of NPS in rankings within the field of IT.

2.6.1 Gartner Peer Insights

Gartner Peer Insights is a well-known platform where enterprise and service decision-makers can find peer reviews and ratings (Gartner Peer Insights, 2023). This platform leverages Gartner's Voice of the Customer (VoC) methodology, which systematically gathers and analyzes customer feedback, including preferences, expectations, and dislikes. The VoC methodology is often used alongside Gartner's expert-driven research, such as Magic Quadrants and Market Guides. Despite being complementary, the VoC methodology is crucial in the decision-making process, as it

focuses on insights derived from the real-world experiences of peers in buying, implementing, and operating various solutions (Gartner Peer Insights, 2023). In an effort to understand user interest and adoption, one of the three critical factors used in the VoC methodology (each with equal weight) that determines the vendor score for the X axis includes the user willingness to recommend the vendor.

2.6.2 Temkin Group

Temkin is an information intermediary who utilizes the NPS metric in their vendor assessment reports, they have a metric specific report called the "Tech Vendor NPS Loyalty Benchmark B2B" which is released on an annual basis. In this case, the Net Promoter Score of over 60 technology vendors is assessed and analyzed. To gather this data, the method they employ is to explicitly survey IT decision-makers from North American firms about their relationships with their technology providers. Through this research, Temkin provides a hierarchical listing of how each tech vendor ranks based on their NPS (BusinessWire, 2017).

2.7 Limitations in IT Rankings

Now that we have explored the role of information intermediaries in the creation of ranking methodologies, examined the different types of rankings and those in the field of IT, we will highlight the limitations present within these methodologies.

2.7.1 The Reliance of Self-Reported Measures

The literature reveals significant limitations in current ranking methodologies, particularly regarding the quality of the information used. Researchers have consistently emphasized that information quality is crucial for effective ranking systems (Rindova et al., 2017; Ringel et al., 2020). Specifically, (Rindova et al., 2017) critiques these practices for relying on "arbitrary choices in normalization, weighting, and aggregation, which undermine transparency and accuracy". There is a call for robust measurement practices, rigorous statistical methods,a transparent data collection and aggregation process to enhance the credibility of rankings (Rindova et al., 2017). Many IT rankings or awards today rely heavily on explicit measures, such as online questionnaires to gauge user perspectives or expert reviews to rate user experiences.

Researchers in the field of IS have pointed out that these subjective measures are often unreliable (Nima, Cloninger, Persson, Sikström & Garcia, 2020) poorly representative of constructs (Cuvillier et al., 2021), and can hinder an unbiased data collection (Hossain, 2017). The current approach used in IT ranking methodologies overlooks the potential of IM's, which offers a more objective and nuanced understanding of user experiences. By integrating IM's, we aim to bridge the information quality gap that exists in current methodologies, capturing deeper insights into users' emotional and physiological responses. This not only complements traditional metrics but also addresses significant IS measurement problems (Pavlou & Dimoka, 2010), advancing towards more rigorous IT rankings.

2.7.2 NPS as a Growth Metric

In investigating the methodologies used in rankings, the NPS metric came out as a growth and performance metric valuable to firms. Initially introduced as a transaction-based customer loyalty metric, NPS is now utilized by many well-known companies, including Apple, as a central marketing tool that informs decision-making and is additionally communicated within earnings reports to investors (Baehre, O'Dwyer, O'Malley & Lee, 2021; Safdar & Pacheco, 2019). Since Fred Reichheld introduced it in the Harvard Business Review in 2003, NPS has gained significant popularity and value as a primary method for measuring customer experience (Keiningham, Cooil, Andreassen, & Aksoy, 2007). The powerful management tool optimizes customer loyalty by asking a straightforward question: "How likely is it that you would recommend this company to a friend or a colleague?" (Reichheld, 2006a). In essence, (Reichheld, 2006a) classifies clientele into three groups using the NPS Methodology:

- *Detractors:* Often called critics, customers within this category score less than or equal to 6. This group is more likely to leave, complain more often and generate negative word-of-mouth (WOM) (Raassens, Haans, 2017).
- *Passives:* These are indifferent customers. They score between 7 or 8.
- *Promoters:* These are those who really support the company's products and rate them 9 or 10. This group tends to be more loyal and will generate more positive WOM (Rassens, Haans, 2017).

The idea that NPS can predict growth in firms is logical: when a customer is loyal and spreads positive WOM with their friends or colleagues, it can lead to new customers for the brand, ultimately boosting sales (Baehre et al., 2021). Similarly, NPS has been successfully associated with word-of-mouth behavior (Keiningham, Cooil, Aksoy, Andreassen, & Weiner, 2007), increased consumer spending (Mecredy, Philip & Wright, Malcolm & Feetham, Pamela, 2018), retention intent (Pollack, Birgit & Alexandrov, Aliosha, 2013), and actual customer retention (De Haan, Verhoef, Wiesel, 2015). These factors form critical links to sales growth (Baehre et al., 2021). However, as more products and services transition online, there are new challenges and opportunities for measuring customer experience, including limitations that come with utilizing NPS on a broad scale or in this case, rankings.

2.7.3 Methodological Issues with NPS in an Online Context

The NPS Methodology previously defined is highly criticized in academia where many issues have been identified when it is applied more specifically in an online context. A key aspect of the NPS Methodology is in converting the *Detractors* into *Promoters.* However, theoretical and empirical research has raised concerns about the NPS Methodology and Reichheld's claims that it is the singularly most reliable predictor of growth. Critics such as (Fisher, 2019) argue that relying solely on NPS as a growth metric can be problematic because it fails to provide detailed insights into specific customer behaviors and needs. For example, research conducted by (Stoop, 2009) on a prominent health and wellbeing company, revealed shortcomings in how NPS was implemented within an online context. In the (Adams, Walpola, Scembri & Harrison, 2022) study on the use of NPS within patient experience, the NPS metric was identified as an insufficient measure due to its lack of specificity. To address concerns based on NPS feedback, researchers (Adams et al., 2022) advocate for the use of "multi-item instruments". It is suggested that this may provide a more detailed assessment of the experience and help pinpoint areas that need improvement.

Researchers (Grisaffe, 2007; Baehre et al, 2021; Fisher, 2019) have identified that NPS is shown to be a clear indicator of whether users will be loyal, buy again and recommend, but it fails to provide information regarding “why” they have indicated that they will be disloyal or loyal. In

the context of IT and more specifically in designing intuitive experiences, stakeholders need to understand root causes to derive deeper insights and modify design decisions. In this case, there is a clear necessity for feedback mechanisms that will provide deeper insights into converting Detractors into Promoters.

Furthermore, in the research we have reviewed (van Door, Leeflang & Tijs, 2013; Morgan et Rego, 2006) highlights that the single scale format in which the NPS metric is presented reduces its predictive capability for future sales growth. The methodological limitations that come with using NPS as the sole success measure, specifically in an online context, has brought us to propose a new success variable and novel construct that will serve to address some of these limitations within an online context.

Chapter 3 : The Proposed VSI Model

The following section is based on the findings we've made in the literature review. First, we introduce the novel formative construct, Viral Success, which serves as the outcome variable for our proposed ranking, Viral Success Ranking (VSR). We then present the Viral Success Model (VSI), subsequently, we delve into how IM's that can be used to predict the formative construct of Viral Success and subsequently serve as inputs to a supervised MLM designed to predict the construct. The aim of this section is therefore twofold to introduce the novel formative construct of Viral Success and to identify the potential IM's that can be used to predict the construct and its potential interaction with Viral Success.

3.1 Building a Novel Outcome Variable to Model Usability (*Viral Success*)

To address the limitations of the NPS methodology, tailor it to an online context, and in turn establish a more comprehensive variable that represents a successful online experience, we are proposing the formative construct of "Viral Success" as the success variable for the VSM. Viral Success would provide a unidimensional score, which (Segars, 1997) refers to a measurement that captures a single underlying construct or trait, ensuring that all items or elements included in the score are focused on measuring one specific aspect. Unidimensionality is essential in ensuring that the items in a scale or measurement tool are all aligned towards a common objective, which enhances the interpretability of the scores derived (Segars, 1997). For example: the System Usability Scale (SUS) generates a unidimensional score that reflects overall usability of a system, providing a straightforward metric for evaluation (Harper & Dorton, 2021). This example showcases how unidimensional scores can simplify complex constructs into a single score that is easier to interpret and apply in practice.

As argued in research by (Rodden, Hutchinson, Fu, 2010) any metric, including NPS, must explicitly be tied to specific goals and effectively monitored to gauge progress. Given user experience is multifaceted, encompassing both objective and emotional dimensions, we suggest introducing a new outcome variable that incorporates the success metric of NPS but builds upon

this by providing an objective metric (e.g task success). Thus, while we acknowledge the challenges of combining these two variables into a formative construct that becomes the single main metric of assessment, we believe that the benefits of using this metric in an online context outweigh the drawbacks. In the following section we will explore both the limitations and the advantages of combining NPS and Task Success into a single formative construct.

3.1.1 Understanding a Formative Construct

Constructs are utilized to characterize phenomena, whether observable (e.g., task success) or internal (e.g., customer attitude), encompassing various aspects such as outcomes, structures, behaviors, and cognitive or psychological dimensions relevant to the phenomenon under investigation (Petter, Straub, & Rai, 2007). The nature of formative constructs requires that the selection of indicators is essential to the definition of the construct, as each indicator represents a unique facet of the construct and contributes to its overall meaning (Franke, Preacher, Rigdon, 2008; Roberts & Thatcher, 2009). The modeling of formative constructs involves specific statistical considerations, as the relationships between indicators and the construct are not necessarily interchangeable or proportional. This means that changes in the indicators directly influence the construct's overall value, highlighting the importance of carefully selecting and validating these indicators (Cheah, Sarstedt, Ringle, Ramayah, & Ting, 2018; Devinney, Coltman, Midgley, & Venaik, 2008). As a general example of a formative construct, (Hall & Shackman, 2020) highlight job satisfaction as a formative construct with indicators including work, pay, social, supervision and growth as indicators of the construct. In this case, (Hall & Shackman, 2020) highlights that all indicators (work, pay, social, supervision and growth) cause the construct of job satisfaction.

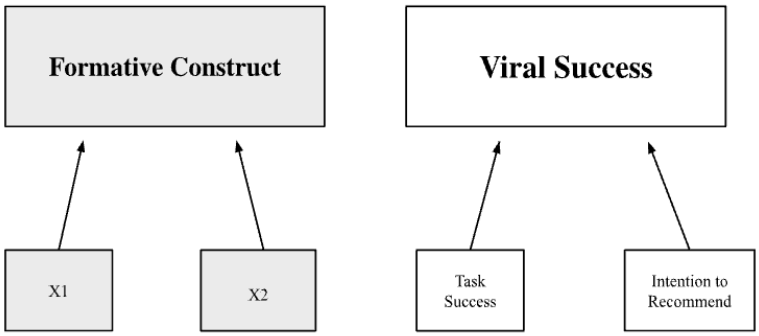
3.1.2 Benefits of Using “Viral Success” as a Novel Construct

For the VSM, we are proposing combining metrics of Task Success (to measure effectiveness) and NPS (to measure intention to recommend) into a formative construct (Figure 1). We suggest that this would offer several benefits that can enhance the understanding and evaluation of UX. Research indicates that users tend to display a higher degree of satisfaction when systems have a higher effectiveness, however it is not solely based on a system's effectiveness, it is a

combination of factors (Al-Maskari & Sanderson, 2010). For instance, when users perceive that their needs are met effectively, they are more likely to express a willingness to recommend the service (Hamilton, Lane, Gaston, Patton, Macdonald & Howie., 2014). This relationship is further supported by findings that highlight the importance of utilitarian benefits influencing intentions to recommend. Research done by (Anggraini & Bernarto, 2022) suggests that products perceived as effective directly correlates with higher NPS scores (Anggraini & Bernarto, 2022).

Research done by (Sauro, Kindlund, 2005) highlights that usability analysts in a wide range of industries are encouraging business leaders to track usability in combination with other company performance metrics. In this sense, integrating both indicators into one construct would allow for a more robust evaluation of user experiences across contexts. As an example, in healthcare settings, NPS has been adopted as a key metric for assessing patient satisfaction, which is closely associated with the effectiveness of the care received (Hamilton et al., 2014; Doyle et al., 2013). In addition, in research done by (Kurz, Brüggermeier & Breiter, 2021) it is noted that task success alone does not fully explain user experience. Therefore, we stipulate that combining task success with the traditional growth metric of NPS can facilitate a deeper understanding of the factors influencing user experience and enable firms to develop strategies that enhance both the effectiveness of their interfaces and user loyalty.

Figure 1: Formative Construct of Viral Success



3.1.3 Limitations of Using "Viral Success" as an Outcome Variable

Now that we have showcased the benefits of combining both Task Success and Intention to Recommend as indicators for the novel formative construct of Viral Success, it is essential that we highlight the many limitations that come with combining both variables. Based on the study done by (Bollen, Kenneth & Diamantopolous, 2017), we will highlight some of the key limitations of combining these variables.

As claimed by (Bollen et al., 2017), formative indicators are causes rather than measures. In this sense, introducing the novel construct introduces complexity and ambiguity in interpreting and validating the construct. As highlighted by (Edwards, 2010; Khatri & Gupta, 2019) one concern that arises from formative constructs is related to measurement error and statistical robustness which can undermine the reliability of the measurements derived from them. Formative constructs are often treated as causal indicators, which can lead to a misinterpreted relationship between the constructs and their indicators. In contrast, researchers, including (Edwards, 2010) have suggested that while formative constructs capture multidimensional aspects of a phenomenon, researchers should prioritize reflective measures, which are more straightforward in terms of interpretations, thereby avoiding the pitfalls with formative constructs. In addition, (Edwards, 2010) highlights that formative constructs often suffer from problems related to construct validity. The aggregation of indicators into a single construct can obscure the individual

contributions of each indicator, which can cause potential misinterpretations of the underlying phenomena (Edwards, 2010).

To conceptualize the limitations presented, Task Success and NPS are fundamentally different measures/indicators. Task Success is measuring the effectiveness with which users can complete a specific task, while NPS is capturing the user's willingness to recommend their experience. Since both are formative indicators, they each contribute to the overall construct of Viral Success, but in distinct ways. Given these indicators are causes, understanding how each one impacts Viral Success may become challenging. The integration of objective task performance metrics with subjective measures like intention to recommend can lead to a loss of granularity in the data. Our objective measure proposed of task success, (e.g click data) will provide quantifiable insights into user interactions, while intention to recommend (e.g NPS) will capture a users' emotional experience. However, by merging these distinct types of data into a single unidimensional variable, important nuances may be overlooked. For instance, the user may complete the task successfully but still feel dissatisfied due to other factors, such as interface design or emotional engagement, which would not be reflected in the measure (Davidson, McFarland & Glisky, 2006).

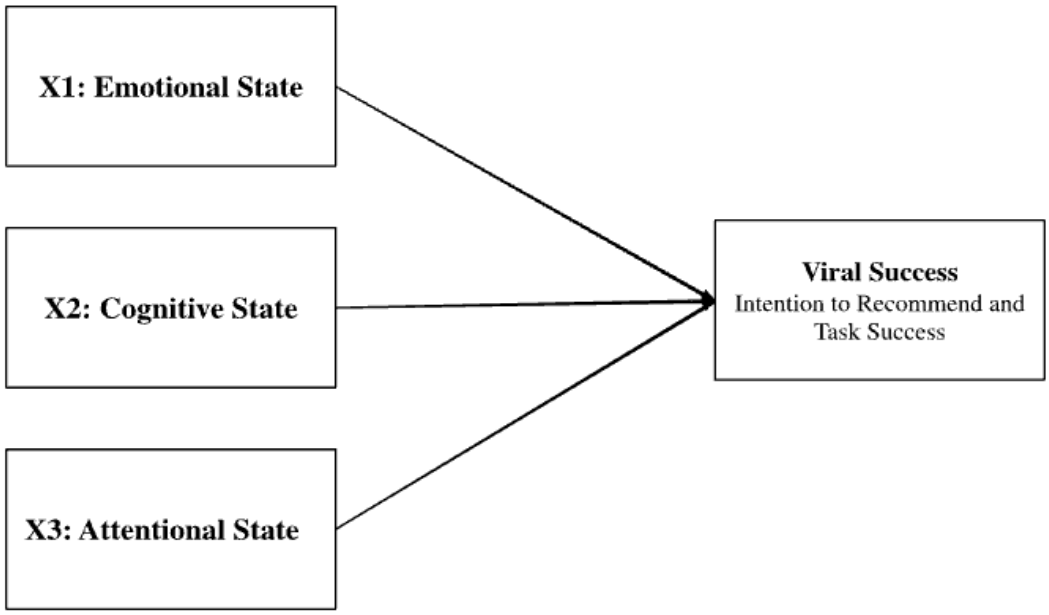
3.2 Conceptual Framework for the VSI Model

3.2.1 Predicting Viral Success

When it comes to the assessment of the novel formative construct of Viral Success, several key questions emerge: How can this construct be accurately measured and predicted? What actions signify that Viral Success has been achieved? What emotions and perceptions are associated with both achieving a Viral Success and failing to achieve one in this context? At this point, the concept of signals and the sources of data for these signals becomes relevant. As highlighted in the literature, one of the most essential points in choosing the correct signals to measure a construct, is to consider the signals that are sensitive and specific to the goal itself (Rodden, Hutchinson, Fu, 2010). The following section of the literature review dives into the different

IM's that can be used to measure and therefore predict the formative construct of Viral Success, as well as their importance when it comes to predicting the construct. More specifically, we are highlighting the relationship of the emotional, cognitive and attentional features with Viral Success. The Viral Success Model is shown in Figure 2.

Figure 2: The Viral Success Model



3.2.2 Cognitive State and Viral Success

As highlighted by (Paas, Van Merriënboer & Adam, 1994), cognitive load refers to the demands placed on an individual's cognitive system while performing a specific task. The relationship between cognitive states, as indexed by pupil size, is a well-established physiological marker of cognitive load, which we stipulate is relevant in understanding and predicting the construct of Viral Success.

3.2.2.1 Pupillometry and Intention to Recommend

Research indicates that pupil size increases with cognitive load, suggesting that as individuals engage more deeply with a task or stimulus, their cognitive effort is reflected in pupil dilation (Zénon, Sidibé, Olivier, 2014). For instance, (Wahn, Ferris, Haristron & König, 2016)

demonstrate that pupil sizes scales with attentional load, indicating that higher cognitive demands lead to larger pupil diameters. Conversely, a manageable cognitive load, indicated by stable pupil size, can enhance user satisfaction (Qu, Guo, Wang & Dang, 2022). Furthermore, in a study by (Wals & Wichary, 2023) on cognitive effort during website-based task performance, findings demonstrated that cognitive effort was linked to an increased intention to recommend a website to others. More specifically, (Wals et al., 2023) highlights that decreased cognitive effort is associated with a greater intention to recommend a website. In support of this, researchers (Hu, Hu & Fang, 2017) highlights that when it comes to performance outcomes a higher level of cognitive processing can have a negative impact on levels of satisfaction. The relationship between satisfaction and intention to recommend has been well established in the field of IS, where satisfied users are more likely to become promoters, contributing to an increased likelihood in their intention to recommend a product or service. Findings from a study done by (Hu et al., 2017) on the mediating roles of cognitive load on performance outcomes and satisfaction suggest that when a navigation structure does not reduce cognitive load, the effects on user satisfaction may weaken. Therefore, we suggest that pupil size serves as a valuable metric for assessing and predicting cognitive load during digital interactions. Understanding how cognitive load influences user experience and intention to recommend can guide the design of more effective digital interfaces that enhance user satisfaction and encourage positive word-of-mouth.

3.2.2.2 Pupillometry and Task Success

Researchers such as (Buettner, Maier, Sauer, and Eckhardt, 2018) and (Longo, 2018) have noted that mental workload is strongly correlated with user performance. More specifically, (Xie and Salvendy, 2000) and (Longo, 2018) state that both mental underload and overload can negatively impact task success. Successful task completion often requires sustained attention and cognitive effort. Studies indicate that pupil dilation can predict performance metrics in various tasks, including those requiring visual search and memory load (Stolte, Gollan & Ansorge, 2020). For instance, a study done by (Wessel et al., 2011) demonstrated that significant differences in pupil diameter can occur immediately after stimulus presentation, suggesting that pupil responses can reflect cognitive processing even before a behavioral response is made. This finding is supported

by the work of (Gilzenrat, Nieuwenhuis & Jepma, 2010) who found that larger baseline pupil diameters are associated with slower reaction times and less accurate performance, indicating a disengagement from the task.

The takeaways from the literature regarding cognitive state brings about a first research proposition (RP1):

Research Proposition 1 (RP1): Cognitive state, measured by pupillometry, is predictive of Viral Success.

3.2.3 Attentional State and Viral Success

Based on insights derived from the research led by (Krejtz, Duchowski, Kretjz, Szarkowska & Kopacz, 2016), (Qu, Guo, Wang & Dang, 2022), (Krukar, Mavros & Höelscher, 2020) and (Carrasco, 2011) we believe that the coefficient K can be a significant predictor of Viral Success. When it comes to users successfully completing a search task, users need to efficiently allocate and manage their attention. As highlighted by (Carrasco, 2011) attention gives users the capacity to selectively filter and process the vast array of information in their visual environment, emphasizing certain details by concentrating on specific areas. Coefficient K is an appealing measure of attention given its ability to be tracked over time. It offers insights into how attention modes dynamically shift as individuals navigate through different spaces (Krukar et al., 2020). Thus, the coefficient K has been shown to be an interesting measure when it comes to providing insights into cognitive strategies that are employed in task solving.

Representing the ratio of attention dispersion, coefficient K originates from studies of visual attention dynamics, especially in differentiating between ambient and focal attention. It was introduced as a parametric measure to quantify the variations between the two visual behavior types using eye-tracking data, more specifically fixations and saccades (Krejtz et al., 2016). It is based on the difference between fixation durations and the amplitude of subsequent saccades, measured in terms of standard deviations, enabling parametric statistical analysis of attention states over time (Krejtz et al., 2016). In Table 2, we have provided information regarding the interpretation of coefficient K (Negi & Mitra, 2020).

Table 2: Interpretation of Coefficient K

	Fixations and Saccades	Interpretation
K > 0	Long fixations followed by Short saccade amplitudes	Focal attention
K < 0	Short fixations followed by Long saccade amplitudes	Ambient attention
K = 0	Long (short) fixations followed by long (short) saccades	Interpretation remains ambiguous

Source: Negi, S., & Mitra, R. (2020). Fixation duration and the learning process: An eye tracking study with subtitled videos. *Journal of Eye Movement Research, 13*(6), 10.16910/jemr.13.6.1. https://doi.org/10.16910/jemr.13.6.1

3.2.3.1 Coefficient K and Intention to Recommend

As highlighted by (Kretjz, Duchowski, Krejtz, Szarkowska, 2016) and in Table 2, coefficient K serves as a valuable metric for understanding how individuals allocate their attention during tasks where short fixations followed by long saccades (K< 0) indicate ambient attention, on the contrary long fixations followed by short saccades indicate focal attention (K > 0). Although the area of attention and intention to recommend has not been vastly studied, current research suggests that the manner in which individuals allocate their attention may have a direct relationship with their intentions. For example, a study done by (Riswanto, Ha, Lee & Kwon, 2024) which showcases the importance of eye tracking technology in understanding consumer behaviour and purchasing intentions highlighted that advertisements that featured aspects including products or models were able to gain more visual attention from consumers, while the advertisements that included promotional content were able to significantly affect decision making and purchase intentions. In addition, (Behe, Bae, Huddleston, Sage, 2015) details that within brick-and-mortar stores, consumers who take more time looking at POP display elements and more cognitive effort processing the product related information, are likely to purchase the

product on display. In addition to this, when users can effectively manage their cognitive resources, they are more likely to experience higher satisfaction levels (Simsekler, Alhashmi, Azar & Osi, 2021). The following studies suggest that both attention can contribute to satisfaction and customer intentions but in different ways. As an example, research done by (Wals et al., 2023) on user experience for website-based tasks, it was demonstrated that in a situation where a there is an additional time pressure applied during visual search, a user's attention becomes disrupted due to an increased cognitive effort, and a more "superficial visual scanning behavior" is likely to become the approach to be taken (Wals et al., 2023). The study by (Wals et al., 2023) emphasized a key finding: in website-based tasks, a decreased cognitive effort (measured by coefficient K) was associated with an increased user intention to recommend.

3.2.3.2 Coefficient K and Task Success

Second, we will explore the relationship between visual attention and task success. Task success often hinges on the ability to balance focal attention— directed at specific targets—and ambient attention, which maintains awareness of the surrounding context. The coefficient K captures this balance, making it a valuable metric for assessing task success. Research has shown that focal attention plays a role in enhancing task success, particularly in tasks that require detailed visual discrimination. For instance, (Daini, Albonico, Primativo, Malaspina, Corbo & Arduino, 2021) found that focal attention can reduce reaction times in tasks involving foveal vision, emphasizing its importance in scenarios requiring quick and accurate responses. Similarly, (Daini et al., 2021) noted that the attentional window can be adjusted based on task demands, further influencing performance in visual tasks.

Conversely, ambient attention plays a complementary role by providing contextual information that enhances focal processing. For example, (Guo et al., 2022) found that eye movement patterns during complex tasks reflect the relationship between ambient and focal attention, suggesting that ambient attention can help maintain situational awareness while focal attention is directed toward specific task elements. Moreover, the importance of ambient attention in performance is evident in scenarios where situational awareness is crucial. For instance, in a study done by (Lenneman, Lenneman, Cassavaugh & Backs, 2009) on driving tasks, ambient

vision helps maintain vehicle control, while focal vision is essential for responding to immediate hazards. This distinction underscores the need for both attentional modes to ensure optimal success across various tasks. In summary, the interaction between focal and ambient attention is crucial in shaping task success. Focal attention enhances the ability to process specific stimuli quickly and accurately, while ambient attention provides the essential contextual information that supports overall cognitive functioning.

The takeaways from the literature regarding attentional state measured by coefficient K, a combination of fixation and saccades, brings about a first research proposition (RP2):

Research Proposition 2 (RP2): Attentional State, measured by coefficient K, is predictive of Viral Success.

3.2.4 Emotional State and Viral Success

An emotional state is the psychological and physiological condition in which emotions and behaviors are interconnected and evaluated within a specific context, encompassing emotional dimensions like valence and arousal (Kim, Kim & Kim, 2013). Russell's (1980) Circumplex Model of Affect stands as a pivotal framework in the exploration of emotional and affective states, particularly within the field of Information Systems IS. This model posits that emotional states originate from two core neurophysiological systems: one associated with valence, representing a continuum from pleasure to displeasure, and the other, linked to arousal or alertness (Posner, Russell, & Peterson, 2005). It was proposed that the dimensions of valence and arousal are independent of one another, since how someone feels is not directly associated to how calm or activated they are (Posner et al., 2005). Given the dimensions of emotional arousal and valence are treated independently, we will highlight the relationship of each construct separately based on its predicted association to Viral Success. We suggest that understanding the link between a user's experienced emotional state during momentary interaction can generate exciting avenues for understanding and predicting how user experience design can evoke Viral Success.

3.2.4.1 Arousal and Intention to Recommend

We first dive into what the literature says about the relationship between emotional arousal and the first component of viral success, intention to recommend. Research by (Cuviller et al., 2021) states that emotional arousal refers to the strength of an emotion felt. Research by (Wang, Zheng, Tang, & Luo, 2023) shows that arousing a positive emotion like joy and trust can enhance a consumer's intention to recommend and strengthen their attachment to a brand. Moreover, studies by (Nawjin & Biran, 2018) and (Wang et al., 2023) suggest that on the flip side negative emotions, such as sadness, can also significantly impact a user's intention to recommend, often leading to complaints or switching behavior due to their stronger cognitive and behavioral effects. High arousal, has also been said to enhance the memorability of experiences, as indicated by (Costanzi, Cianfanelli, Saraulli, Lasaponara, Doricchi, Cestari & Rossi-Arnaud, 2019), who found that arousal levels can improve spatial memory performance, particularly for emotionally charged stimuli. This suggests that experiences that evoke strong emotional responses, whether positive or negative, can leave a lasting impression on consumers,

3.2.4.2 Arousal and Task Success

Arousal, as a dimension of emotion, can also significantly impact users' ability to perform tasks effectively. Arousal is known to enhance cognitive performance under certain conditions. For instance, in a study done by (Demanet, Liefooghe & Verbruggen, 2011) found that higher arousal levels can help individuals avoid interference from irrelevant task-sets, thereby strengthening the focus on the currently relevant task. This suggests that when users are in a heightened state of arousal, they may be better equipped to manage distractions and maintain task success, which is crucial in environments requiring sustained attention. Conversely, the effects of emotional stimuli on task success can be complex. Research done by (Houwer & Tibboel, 2010) demonstrated that emotional pictures could interfere with task success, primarily driven by their arousal value. This indicates that while moderate arousal can enhance focus, excessive arousal—especially from negative emotional stimuli—can lead to cognitive overload, impairing task performance. Research by (Costanzi et al., 2019) further supports the notion that arousal significantly influences memory and cognitive tasks. Their findings indicated that high arousal, particularly

from negative stimuli, can enhance task success in spatial working memory tasks (Costanzi et al., 2019). This suggests that arousal can facilitate certain cognitive processes, making it a valuable factor in designing tasks that require memory recall and spatial awareness.

3.2.4.3 Valence and Intention to Recommend

Valence (pleasure) encapsulates six emotional states that are universally accepted. It includes joy, sadness, surprise, fear, anger and disgust (Cuvillier et al., 2021). Valence has a significant impact on the dimension of Viral Success, notably intention to recommend. Researchers such as (Markus, Makkonen, Riekkinen, Frank & Jussila, 2018) and (Sbai, 2013) have highlighted that positive emotions (e.g joy) have a favorable effects on recommendation intention, meaning that if a consumer feels good during their shopping experience, it may enhance the desirability of the product and lead to a higher intention to recommend. For instance, (Hosany, Prayag, Huang & Dessilatham, 2016) found that positive emotions significantly mediate the relationship between user satisfaction and the intention to recommend tourism experiences, highlighting the importance of emotional engagement in fostering positive word-of-mouth. Similarly, (Gomes et al., 2013) demonstrated that emotional valence affects recall and retrieval processes, suggesting that positive emotional experiences enhance memory and subsequently influence recommendation behaviors. In this sense, we predict that user valence is strongly associated with their intention to recommend.

3.2.4.4 Valence and Task Success

We also propose that valence is associated with task success. First, (Zsido, Bernáth, Labadi & Deak, 2020) highlight that negative valence can decrease task success by diverting attention away from the task, but arousal can compensate for this by increasing attentional capacity. Further, research indicates that emotional valence can significantly affect memory and attentional processes, which are crucial for task success. Research done by (Schnitzspahn, Horn, Bayern & Kliegel, 2012) found that emotional valence influences not only memory but also the controlled attentional processes necessary for successfully performing prospective memory tasks. This suggests that positive or negative emotional cues can modulate how effectively users engage with tasks, impacting their overall performance. Similarly, (Kopf, Dresler, Reicherts, Herrmann

& Reif, 2013) demonstrated that emotional content affects brain activation during cognitive tasks, particularly in working memory scenarios. Their findings indicate that in more challenging tasks, such as a 3-back working memory task, the valence of the stimuli becomes critical for performance. This highlights the importance of considering emotional valence in UX design, as it can influence cognitive load and the ability to manage task demands effectively.

The takeaways from the literature regarding emotional state, characterized by valence and arousal, brings about a third research proposition (RP3):

Research Proposition 3 (RP3): Emotional state, measured by Valence and Arousal, is predictive of Viral Success.

Chapter 4 : Methodology

The proposed VSM was built using data collected during an experiment conducted at the institution's laboratory that specializes in user experience (UX) evaluation from June — October 2023. The study and all of its procedures adhered to the ethical guidelines set by the institution's Research Ethics Board (REB). The study was approved by the Research Ethics Board (REB) with ethical approval ID: 2023-5392.

4.1 Experimental Design

The experimental design employed during the data collection was a within-subjects design to ensure counter balanced exposure. Participants were asked to perform the same three (3) experimental tasks across nine (9) different financial institutions. This was a 2-factor experiment, where the first factor was the banking interface shown to the user, and the second factor was the task at hand. Users needed to complete 3 tasks on each of the 9 interfaces, for a total of 27 different tasks 9 (interfaces) x 3 (tasks). Nine (9) banks were chosen based on a need for a representative sample of the Canadian banking industry. The three tasks were designed to replicate a typical credit card shopping experience, where the user lands on a page, reviews the information presented, and then makes a decision based on their analysis. The three tasks included having users (1) find and click on the annual fee of the card, (2) find and click on the grocery rewards of the card and (3) find and click on all other purchases rewards. All users were given ninety (90) seconds to complete all three tasks. A time frame of 90 seconds was chosen for two reasons. First, the pretests we conducted before the experiment demonstrated that 90 seconds was the ideal time frame in order to have users conduct the tasks and respond to questions after. Second, given the user needed to complete 3 tasks, 90 seconds showed to be enough for most users to engage with the content on the page in a meaningful way without being rushed. The stimuli were counterbalanced, ensuring that most participants experienced distinct sequences of bank interfaces throughout the experiment. This strategic approach was crucial for mitigating sequence effects and maintaining control in the experimental design. Within the cohort of nine

banks, four (4) banks had critical credit card information positioned at the top of the page, while the remaining five (5) banks had critical information placed at a mid to bottom section of the page. This deliberate variation in how credit card details were presented aimed to offer diversified insights into the different strategies of information placement.

4.2 Sample

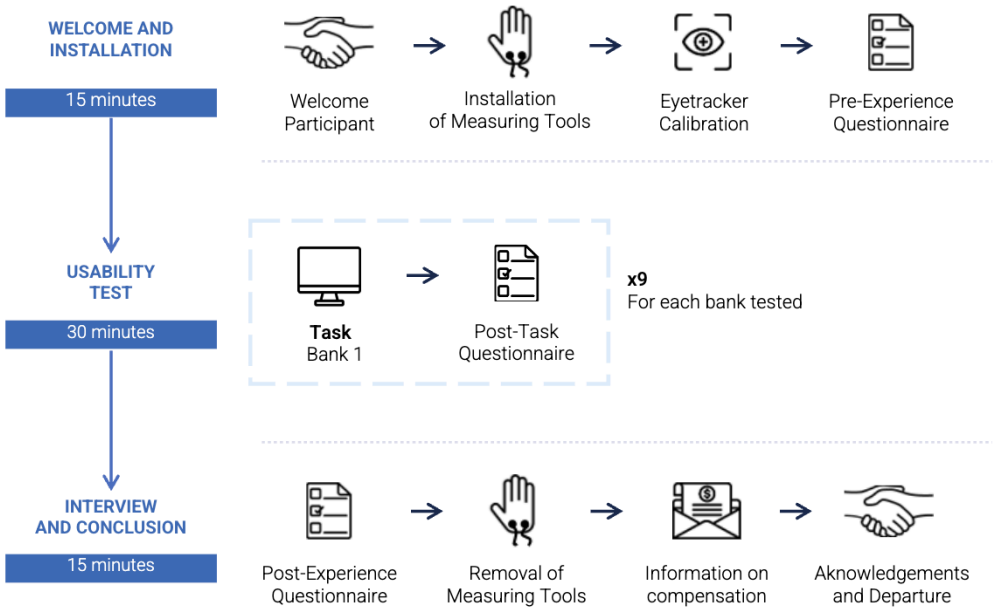
A total of one hundred (N=101) participants were recruited for the study aged between 18 and 64. Participants were solicited using our institution's participant panel, word-of-mouth, and social media. All participants were required to sign a consent form. All participants were screened to ensure they did not have skin allergies, heart problems, or epilepsy. All participants needed to meet specific requirements including being able to understand and speak English at an advanced level, having an active bank account with one of the 9 major Canadian banks, and being over the age of 18. Participants were compensated with a total of $30 for their participation.

4.3 Experimental Procedure

The study took place in a laboratory environment at HEC Montréal. Before the experiment's initiation, several pretests were conducted to ensure the quality of the experiment. The experimental protocol was as follows; participants were greeted by the researchers in the room. All participants were asked to provide their consent on a tablet where all information regarding the experiment set up was explicitly provided. After all participants had provided consent, we then proceeded to the installation of physiological tools. The experimental researcher placed the electrocardiogram (EKG) sensors on the right, left collar bone of the participant and the left lower rib bone to capture arousal. The electrodermal activity (EDA) sensors were placed on the palm of the participant's non-dominant hand to capture arousal associated with emotion, cognition and attention. After this, the EKG and EDA data was validated. For EKG, data was validated by observing peaks and EDA was validated by observing if the values fell between the manufacturer's specified acceptable threshold of 100-1000. This validation was done to ensure the recording of the data was reliable. An eye tracker calibration exercise was done right after to

ensure the proper syncing of every participant's retinal movement to the recording equipment. Once the baseline and calibration exercise were deemed complete, the participants were shown the experimental instructions. Before the experiment officially began, each participant was provided with a preliminary trial run. The preliminary trial run was critical given the experimental tasks needed to be completed within 90 seconds. The preliminary trial was essential for familiarizing participants with the process to maximize the quality of their interaction with the webpage. The preliminary trial was conducted on an American banking website that was not included in the actual study. The trial incorporated all three (3) tasks and an illustrative preview of the questionnaire's format post-task. During this preliminary trial, the researcher addressed any queries participants had, explicitly confirming that all interfaces showcased during the experiment would remain static. All responses provided during the preliminary trial were excluded from the final analysis. After the completion of all 3 tasks on a single financial institutions interface, participants were directed to a questionnaire on Qualtrics. In the questionnaire, constructs including emotional valence, emotional arousal, information recall, perceived cognitive effort, perceived satisfaction, and customer loyalty (NPS) were measured. Upon finalizing all three (3) experimental tasks across all nine (9) financial institutions, participants were required to complete a post-experiment questionnaire on Qualtrics. The final questionnaire consisted of 10 questions which evaluated the users subjective experience and opinion on the 9 different interfaces shown to them. The post experiment questionnaire asked the user to indicate their favorite and least favorite financial institution website experience, demanded demographic information, as well as provided users with the opportunity to add any additional comments regarding their experience. After the participants completed the experiment, they were debriefed and signed a compensation form before leaving the laboratory. Figure 3 details the Experimental Procedure.

Figure 3: Experimental Procedure



4.4 Measures and Apparatus

Table 3 and Table 4 summarize the various constructs measured in the study, and the apparatuses used. In addition, it highlights the subjective and objective measures used to measure "True Viral Success", which forms the foundation of the study. The administration of these measures occurred throughout the experiment capturing data at specific intervals. All stimuli were presented on a standard computer monitor (1920 x 1080 resolution).

Table 3: Summary of Independent Variable Measurements and Apparatus

Independent Variables of the VSI Model				
Construct	**Measure**	**Acronym**	**Administration**	**Apparatus**
Cognitive load	Pupil dilation	Pupil_std	During the task.	Tobii EyeTracker (Tobiix60)
Attention	Fixations and saccades	K_coef	During the task.	Tobii EyeTracker (Tobiix60)
Emotional Arousal	Phasic electrodermal activity (EDA)	Phasic_std	During the task.	EDA sensors (Bluebox MP-150 Biopac)
Emotional Valence	Facial expressions Valence score: ranges from -1 to +1. The valence is calculated by subtracting the intensity of "happy" with the intensity of the negative expression with the highest intensity.	Valence_mean	During the task.	The Noldus FaceReader Software

Cognitive State

Cognitive load was measured using eye tracking data. EyeTracking data was captured using the Tobii EyeTracker (Tobiix60), which provided insights regarding gaze and pupil dilation as an

indicator of cognitive load. The Tobii EyeTracker algorithms determine the position of the eyes and measure pupil size by capturing images of the eyes (Tobii AB, 2024). The pupil size is reported in millimeters, which provides the ability to monitor and study changes in its size (Tobii AB, 2024).

Attentional State

Attention was measured using coefficient K. Coefficient K was measured by subtracting the standardized fixation duration from the standardized amplitude of the subsequent saccade (Krejtz et al., 2017). The Tobii I-VT (fixation) filter can be used to identify fixations and saccades, allowing for the visualization of classified eye movements, including fixations, saccades, and unclassified movements (Tobii AB, 2024).

Emotional State

Emotional state is the combination of valence and arousal. Emotional arousal was measured using electrodermal activity (EDA). EDA was recorded with a Biopac MP-150 system running via the AcqKnowledge 4.4 software (Biopac, Goleta, United States). Emotional valence was measured using the Noldus FaceReader software which recorded facial recognition systems and model valence (Noldus, Wageningen, Netherlands). The post hoc synchronization of the physiological data was done using the Cobalt Photobooth software (Courtemarche, Léger, Fredette, Sénéca, 2018, 2019, 2022).

Table 4: Summary of Dependent Variable Measurements and Apparatus

Dependent Variables of the VSI Model						
Construct		**Measure**	**Acronym**	**Administration**	**Apparatus**	**Scale Items**
True Viral Success	Word of Mouth Intention	Net Promoter Score	NPS	At the end of the task on one single bank. Done 9 times.	NPS Scale Self-Reported questionnaire (Qualtrics)	Likelihood to recommend: 0 [not likely at all] to 10 [extremely likely].
	Task Success	AOI's clicked	nbr_AOI_clicked_web	During the task.	Tobii x60	Number of AOI's clicked [1,3]

Word of Mouth Intention

Word of Mouth Intention was measured using the NPS scale, which was presented via a Qualtrics questionnaire, administered at the end of each task. The NPS Scale, developed by (Reichheld, 2003), is an 11-point scale that ranges from 0-10. From 0, being a zero probability of recommendation to 10, being a maximal probability of recommendation (Cuvillier et al., 2021). The score then provides a ratio corresponding to whether the respondent is considered a detractor (scores from 0 to 7), a promoter (scores of 9 and 10) and the remaining being passives (Cuvillier et al., 2021).

Task Success

Task Success evaluates a participant's efficacy in completing a task. Objective task success was measured through a participant's ability to click on the correct Area of Interest (AOI's) on each banking stimulus. The user was instructed to find 3 AOI's throughout their 90 seconds of exposure to a single stimulus. Task Success for each stimulus ranged from [1 to 3]. We have highlighted the different AOI's in Table 5.

Table 5: Definition of Areas of Interests (AOI's) used to Measure Task Success

AOI #	AOI category	Description of AOI
1	Annual Fee	The lump sums a consumer must pay every year that they are signed up for a certain credit card.
2	Grocery Rewards	The rewards are in the form of cashback, points or miles for using the credit card for grocery purchases.
3	All Other Purchases Rewards	The rewards in the form of cashback, points or miles, for using the credit card for categories labeled as "all other purchases".

True Viral Success

True Viral Success is the dependent variable of the VSM. True Viral Success is the summation of Word-of-Mouth Intention (NPS) and Task Success (nbr_AOI_clicked_web). Given the number of AOI's to click on each website ranged from [1,3], NPS which initially ranged from [0,10] was rescaled to [1,3] to ensure equal weight in the construct. As both components of Viral Success were rescaled to a 1-3 range, the range of Viral Success naturally becomes 2-6 rather than 1-6.

Overall, the data from the study brought about a significant amount of data from about 101 participants which was used to measure a participant's ability to find and click on information

(Task Success) and their intention to recommend the experience (NPS) termed a True Viral Success".

Chapter 5 : Results

Based on the data collected as described earlier, we developed a supervised MLM designed for pattern recognition, enabling us to predict True Viral Success for new data. The chosen algorithm was trained to closely replicate the output of True Viral Success, resulting in the creation of the VSM. The results section details the development process of this prediction model, with an overview provided in Figure 4. In addition, the results section highlights the final VSR of the 8 financial institutions retained and explores the results of the three research propositions.

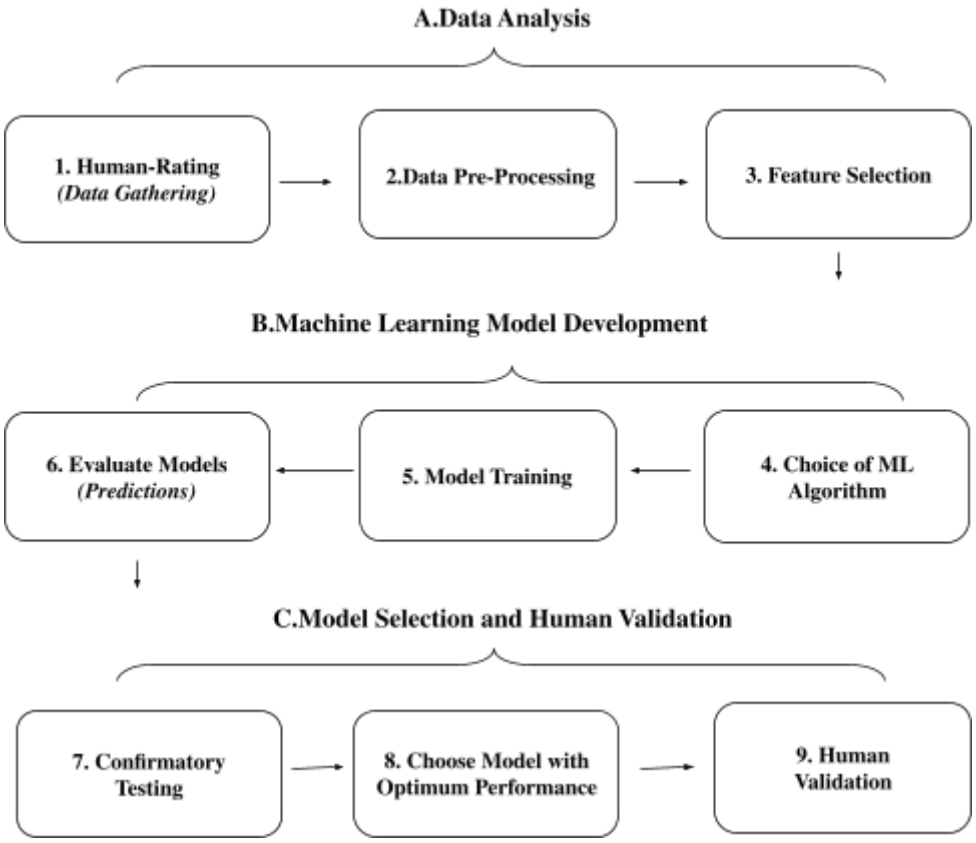
5.1 Combining Task Success and Intention to Recommend

In order to ensure the validity of combining both indicators that compose Viral Success, we first checked the correlation amongst NPS and the number of AOI's clicked. The two components were found to be positively correlated with an $F(1,442) = 38.22$ and $p<0.0001$. The large F value and small P value demonstrated that both variables were strongly positively correlated. In addition, we took into consideration the weight of each variable in relation to Viral Success. As a reminder, the number of AOI's to click on each website ranged from [1,3], NPS which initially range from [0,10] was rescaled to [1,3] in order to ensure equal weight in the construct. The scale therefore ranged from [2-6], rather than [1-6].

5.2 Overview of the VSI Model Development Process

We will begin with a brief overview of the process. As highlighted in Figure 4, the framework used can be broadly classified as: (a) data analysis, (b) supervised machine learning model development and (c) model selection and human validation. The data analysis included three key stages: human rating, data preprocessing and feature selection. The main objective of the data analysis was to gain a thorough understanding of the dataset's behavior, including the attributes and characteristics of each variable.

Figure 4: Supervised Machine Learning Development Lifecycle



After the data analysis had been completed, we were able to begin the supervised machine learning model development portion. This included selecting the appropriate machine learning algorithm to use for our supervised MLM. In the case of a supervised machine learning model, and given the nature of the input data, a supervised algorithm was chosen to model the relationship between the outcome variable (Viral Success) and the independent variables (psychophysiological signals). Once the algorithm selection was complete, different ML models were developed, evaluated and compared.

During the final broad category of the life cycle, a final supervised machine learning model — namely the VSM — was selected based on its performance. The selection of the model was followed by a human validation of the results. The human validation process was completed

through click testing. Now that we have provided an overview of the ML framework used to develop our VSI model, we will dive deeper into each step.

5.3 Data Analysis

5.3.1 Human Rating

To create our supervised MLM (VSM), the data gathered in the previous section served as the foundation. A supervised MLM is trained on a labeled dataset, where the model learns to predict an output (label) based on input data (features) (Carmona, Finley & Li, 2018). A key characteristic of supervised learning is that the training data includes both the input data and the corresponding correct output, allowing the model to learn the relationship between them, it essentially identifies patterns and relationships between input features and the corresponding output labels, enabling the model to generalize and make accurate predictions when presented with new data (Alsajri, 2023). The human rating section of the data collection process was also essential in deriving the psychophysiological data from users as they were rating each interface. During the process, participants' emotion (valence and arousal), cognitive load (pupil dilation) and attention (fixations and saccades) were measured using IM's. These measures were then used in further sections to build the VSI Model Due to an issue with the Tobii data and its processing, HSBC, Bank I, was lost from the dataset, and the exact cause remains unidentified. Since the problem was discovered late in the analysis process, it was no longer feasible to modify or reprocess the data to include HSBC. Despite its absence, the remaining dataset remained robust enough to proceed with the development of the index, ensuring the validity of the analysis.

5.3.2 Data Pre-Processing

After completing the data collection process, our team had the necessary data to predict the outcome variable of True Viral Success. As a reminder, True Viral Success is the actual, observed value of Viral Success from the experiment conducted, measured by the summation of word-of-mouth intention (measured by NPS) and objective Task Success (measured by the number of AOI's clicked) ranging from [1,3]. In contrast, Predicted Viral Success is the value

predicted by the supervised machine learning model based on the input features. Meaning, it is the model's estimate of what True Viral Success will be.

As shown in Table 6, the output variable is the target that our supervised MLM was trained to predict. In this case, the label (output variable) was "True Viral Success". A thorough data cleaning was required to ensure the accuracy and reliability of the data, this included removing any inconsistencies or outlets and preparing the dataset for model training. Cleaning of the data began by extracting the average NPS per participant, per bank for all 101 participants. The number of Areas of Interest (AOIs) clicked by each participant was also extracted via the Tobiix60. The number of AOIs required to be clicked ranged from 1 to 3, while NPS ranged from 0 to 10. To standardize these measures and assign equal weight to NPS and nbr_AOI_found, we rescaled NPS to 1 to 3.

Table 6: Key Concepts in Supervised Learning Models

Category	Description
Features (Input Variables)	Features are the input variables used to make predictions. For the VSM, this includes data from psychophysiological tools such as EDA and the Tobii EyeTracker, which measures emotion, cognition, and attention as the user navigates the interface.
Labels (Output Variables)	Labels are the targets that the model is trained to predict. For the VSM, the output variable was True Viral Success (TVS) defined as: TVS = Net Promoter Score (NPS) + Number of Areas of Interest Found (nbr_AOI_found).

Legend:

Features (Input Variables) include the data sources and their roles in the model.

Labels (Output Variables) specify what the model aims to predict and how the output variable is calculated.

After completing the data cleaning process, we calculated the Average per Bank of True Viral Success. The values for True Viral Success per bank are presented in Table 7 below.

Table 7: Average True Viral Success (TVS) Per Bank

Bank Name	True Viral Success (TVS)
A	3,65
B	4,96
C	5,17
D	5,06
E	4,92
F	4,07
G	3,62
H	4,96

Legend:

Bank Name is the column for the names of the banks. All corresponding bank names can be found in Appendix A.
True Viral Success (TVS) is the column for the corresponding scores.

Using the averages of "True Viral Success" for each financial institution, we created a "True Viral Success Ranking." This ranking provides a hierarchical list of how each financial institution performed relative to providing True Viral Success based on the search tasks provided. The institutions that ranked higher demonstrated greater Viral Success, while those lower on the list showed less Viral Success. Table 8 presents the True Viral Success Ranking (TVSR) derived from the averages in Table 7.

Table 8: True Viral Success Ranking (TVSR)

Rank	Bank Name
#1	C
#2	D
#3	B
#4	H
#5	E
#6	F
#7	A
#8	G

Legend:

Rank is the column corresponding to the ranking position of the financial institution.

Bank Name is the column for the corresponding financial institution.

As demonstrated in the ranking, Bank C emerged as the top bank, generating the highest True Viral Success amongst participants. Conversely, Bank G ranked the lowest in terms of True Viral Success.

5.3.3 Feature Selection

To test the research propositions and in order to select the most relevant features for the supervised MLM, we examined the various IMs (features) and their contributions to the predictive power of Viral Success. The independent variable candidates including the mean, standard deviation (“std”), 90th percentile (“p90”) and 10th percentile (“p10”) of all the available IMs: coefficient K (combination of fixations and saccades), pupil metrics, valence and phasic EDA were modeled as a linear function for Viral Success. In essence, the marginal effect of each signal available in our data set was calculated via a Simple Linear Regression with random

intercept, highlighted by the p-value (Prob-F). All linear regressions for the VSI model development were performed in SAS9.4.

Table 9: Simple Linear Regression

IV	Num DF	Den DF	F Value	Prob F
K_coef_mean	1	443	22,7	<.0001
pupil_adj_std	1	434	18,26	<.0001
pupil_std	1	443	16,18	<0.001
k_coef_p10	1	443	14,81	<0.0001
pupil_adj_p10	1	434	12,24	0.0005
k_coef_p90	1	443	9,07	0.0028
valence_mean	1	355	6,02	0.0146
phasic_std	1	218	5,47	0.0203
valence_p10	1	355	5,06	0.0251
phasic_p90	1	219	4,78	0.0298
pupil_adj_mean	1	434	4,55	0.0335

Legend:

IV stands for independent variable.

NUMDF represents the number of degrees of freedom.

Den DF represents the denominator degrees of freedom

F Value represents the denominator degrees of freedom

Prob F is the p-value associated with the F statistic.

Research Proposition 1: Cognitive State is Predictive of Viral Success

The first research proposition (RP1) examined the marginal effect of cognitive state and Viral Success. The research proposition was tested using a Simple Linear Regression. As shown in Table 9, the Simple Linear Regression demonstrated that cognitive state, measured by pupil standard deviation (pupil_std) and adjusted pupil mean (pupil_adj), influenced Viral Success, with a respective ProbF of <0.001 and 0.0335. These findings support the proposition that cognitive state is a statistically significant predictor of Viral Success.

Research Proposition 2: Attentional State is Predictive of Viral Success

The second research proposition explored the predictive power of attentional state, measured by coefficient K (k_coef_mean). The Simple Linear Regression showed that attentional state has a significant effect on Viral Success, with a ProbF of <0.0001 and a high F value of 22.7. These findings support the research proposition that attentional state is a statistically significant predictor of Viral Success.

Research Proposition 3: Emotional State is Predictive of Viral Success

Additionally, emotional state, measured by valence (valence_mean) and electrodermal activity (phasic_std), was found to influence Viral Success, with ProbF values of 0.0146 and 0.0203, respectively. The results support the proposition that emotional state is predictive of Viral Success. The combination of these variables provides a strong balance of statistical significance, explanatory power and model simplicity, making them the ideal foundation for the next stage of the analysis.

In the next stage, we expanded our analysis from Simple Linear Regression to Multiple Linear Regression, as detailed in Table 10. The goal of the Multiple Linear Regression was to simultaneously evaluate the combination of IM's that were retained due to their significance in the Simple Linear Regression. This approach aimed to detect any interaction effects among the retained variables k_coef_mean, pupil_std, pupil_adj_mean, valence_mean and phasic_std when combined into a single model. This step was crucial to determine whether any of the IM's from

Table 9 overlapped when combined. Identifying any overlaps was important, as including non-discriminative features in a model increases the complexity of the model and reduces its robustness.

Table 10: Multiple Linear Regression

Solution for Fixed Effects (all independent variables in one model)					
Effect	**Estimate**	**Standard Error**	**DF**	**T Value**	**Pr > |t|**
k_coef_mean	1.0684	0.2466	176	4.33	<.0001
pupil_std	-3.4038	0.8777	176	-3.88	0.0001
valence_mean	0.7266	0.4629	176	1.57	0.1183
phasic_std	-0.1944	0.1053	176	-1.85	0.0664
pupil_adj_mean	0.06615	0.3048	176	0.22	0.8284

The Multiple Linear Regression helped rule out the independent variables amongst the k_coef_mean, pupil_std, valence_mean, phasic_std and pupil_adj_mean which at the present of any of the other independent variables, did not add power to the model. Based on the results from the Multiple Linear Regression, no interaction effect was detected. In addition, pupil_std was chosen as the most relevant variable to represent cognitive state, given its respective Pr > |t| of 0.0001. Table 11 below provides a summary of the research propositions verified and their respective results.

Table 11: Summary of Research Proposition Results

RP	Construct	Variable	To	Prob F	Result	RP Description
RP1	Cognitive state	pupil_std	Viral Success	< 0.0001	Significant effect	Cognitive state, measured by pupillometry, is predictive of Viral Success
RP2	Attentional State	k_coef_mean	Viral Success	<0.0001	Significant effect	Attentional state, measured by coefficient K, is predictive of Viral Success.
RP3	Emotional State	valence_mean	Viral Success	0.0146	Significant effect	Valence, measured by FaceReader, is predictive of Viral Success.
		phasic_std		0.0203		Arousal, measured by electrodermal activity, is predictive of Viral Success.

Legend:

RP stands for research proposition.

Construct stands for the physiological state being measured.

Variable stands for features used to assess the construct.

To stands for the dependent variable that the model is trying to predict.

Prob F stands for the p-value associated with the F statistic.

5.4 Supervised Machine Learning Model Development

5.4.1 Model Training and Evaluation

With four final physiological signals chosen as features for the Viral Success Model, six predictive models were developed utilizing a different combination of the relevant psychophysiological features: k_coef, pupil, valence, and phasic EDA. Model performance was evaluated using out-of-sample data through a leave-one-out strategy and comparing the Mean Absolute Percentage Error (MAPE) for each combination.

The MAPE metric was chosen to gauge the accuracy of the six predictive models. This metric is widely used as a goodness-of-fit measure, particularly suitable for datasets with varying magnitude values, like ours (Montaño, Palmer, Sesé, & Cajal, 2013). The MAPE formula, shown below, calculates the average of the absolute percentage errors between predicted and actual values. The MAPE metric allows for straightforward interpretation, as lower MAPE values indicate better predictive accuracy (Shi, Lee, Tsai, Ho, Chen, Lee & Chiu, 2012; Morley, Brito & Welling, 2018). This metric provides insight into the degree of deviation between predictions and actual outcomes, expressed as a percentage.

The formula for Mean Absolute Percentage Error (MAPE)

MAPE Calculation = 100*sum (Actual-Predicted|/|Actual|)/N

Where:

- **n** is the number of observations
- **At** represents the actual value
- **Ft** represented the predicted value

A lower MAPE signifies that the model's predictions are closer to the actual values, suggesting a higher accuracy. Each model illustrated in Table 12 has a MAPE that showcases acceptable forecasting given all MAPE's range between 10-20%, Model 5 results in the lowest MAPE of 16.51%. This indicates that on average, Model 5 prediction distance from True "Viral Success" is 16.5%, i.e. the True Viral Success could be +16.5% or - 16.5% of the Predicted Viral Success. In addition, Model 5 brings about a relatively high-performance accuracy of 83.5% when it comes to predicting True Viral Success compared to the other five VSM's.

Table 12: Model Selection and Performance

Model ID	Model Features	MAPE	Performance
M1	K_coef_mean	0.184427	0.814902
M2	Pupil_std	0.185098	0.815573
M3	K_coef_mean Pupil_std	0.176902	0.823098
M4	K_coef_mean Pupil_std Phasic_std	0.171178	0.8328822
M5	K_coef_mean Pupil_std Phasic_std valence_mean	0.165125	0.834875
M6	K_coef_mean Pupil_std Phasic_std valence_mean K_coef_p10	0.165384	0.834617

Legend:

Model ID stands for the identifier of the model.

Model Features stands for the features used in the model.

MAPE stands for Mean Absolute Percentage Error.

Performance stands for the model's predictive power.

After having calculated the MAPE and performance of each model created, the mean and standard deviation of the top 3 models (M3, M4 and M5) was then determined and calculated across banks as seen in Table 12.

The mean and standard deviation statistics were computed to assess whether the MAPE varied significantly across the tested banks. Table 13 shows although M5 has four independent variables, it exhibits the smallest MAPE Mean (16.43%) and MAPE Standard Deviation (0.95%) across all retained banks. This is why our team chose M5 over M6, given it is the simplest model amongst both options. In sum, this calculation validates that M5 is the most attractive model and remains consistent with its performance across banks.

Table 13: **MAPE by Bank and Model**

	M3	**M4**	**M5**
A	0.1622	0.1497	0.1559
B	0.1855	0.1685	0.1595
C	0.1777	0.1634	0.1518
D	0.1722	0.1672	0.1607
E	0.1784	0.1814	0.1724
F	0.1785	0.1793	0.173
G	0.1779	0.1799	0.1793
H	0.1776	0.1694	0.1616
Summary Statistics			
Mean	0.1762	0.1698	0.1643
Standard Deviation	0.0067	0.0106	0.0095

The final step of the *performance check* was to calculate the Predicted Success of M3, M4 and M5. Table 14 outlines True Viral Success with the Predicted Viral success of each one of the retained models (M3, M4, M5).

Table 14: **Average per Bank True and Predicted Viral Success**

Bank Name	True Viral Success	Predicted Viral Success		
		M3	M4	M5
A	3.65	4.36	4.38	4.39
B	4.96	4.75	4.84	4.89
C	5.17	4.78	4.78	4.81
D	5.07	4.69	4.72	4.73
E	4.92	4.53	4.62	4.68
F	4.07	4.62	4.6	4.61
G	3.62	4.37	4.32	4.34
H	4.96	4.51	4.54	4.57

5.5 Model Selection and Human Validation

5.5.1 Confirmatory Testing

The confirmatory testing section of the model selection involved using another set of research data to validate the VSM's performance by observing the consistency of the rankings obtained. By applying the same ranking methodology to both datasets, we aimed to confirm the robustness and reliability of the initial findings across different data sources. This step ensures that the model's predictive power is not limited to a specific dataset but is applicable in various contexts. We opted to use ranking as a confirmatory step because it is closer to the typical use case of the VSM than a measure like MAPE. Although MAPE is a great metric for comparing models, a ranking is what decision makers and stakeholders truly care about when they will be using the proposed VSM. In this sense, it was essential to ensure that the ranking produced by our selected supervised ML model (M5) matched the True Viral Success Ranking as established in Stage 1 of

the model development, as closely as possible. More specifically, our aim was to achieve consistency in the "Top 3" banks and "Bottom 3" banks of the True Viral Success Ranking.

Table 15: Viral Success Ranking Predicted by M5

	Predicted Viral Success Ranking (M5)	**True Viral Success Ranking**
Rank	**Bank Name**	
#1	B	C
#2	C	D
#3	D	B
#4	E	H
#5	F	E
#6	H	F
#7	A	A
#8	G	G

As highlighted in Table 15, Model 5 was shown to produce a ranking similar to the True Viral Success Ranking. The Canadian Banks: C, D and B ranking amongst the Top 3 banks similarly to that of the True Viral Success Ranking. In contrast, A and G remained among the bottom three ranks using M5. The alignment in the ranking of the banks in the True Viral Success calculation and that of the Predicted Viral Success calculation indicated that our selected model, Model 5, performed well.

5.5.2 Choose Model with Optimum Performance

The central objective of this analytic approach and our model development was to choose the combination of features (implicit measures) that best predict Viral Success. The main results of this study showed that the mean of coefficient K (k_coef_mean) is the most relevant measure of attention for Viral Success, valence (valence_mean) and arousal (phasic_std) are most relevant in predicting emotional state for Viral Success and pupillometry (pupil_std) is the most relevant

implicit measure of cognitive state in predicting Viral Success. Through a performance check and confirmatory testing via ranking, we chose M5 as the final ML model to be used in the VSM.

The Viral Success Model (M5) model can be expressed by the following equation:

VSI Model (M5) = 5.098 + 1.0598 · *k_coef* + 0.7159 · *valence_mean* - 3.3716 · *pupil_stddev* - 0.1953 · *phasicEDA_stddev*

In this model, *k_coef* represents coefficient K, *valence_ mean* is the mean of the valence score, *pupil_stddev* denoted the standard deviation of pupil measurements, and *phasicEDA_stddev* is the standard deviation of the phasic electrodermal activity (EDA). The coefficients of these variables indicate their respective contributions to the VSM (M5) score.

5.5.3 Human Validation

With M5 selected as the final VSM, we aimed to further validate whether the Canadian financial institutions that were initially ranked truly merited their placement in the ranking. The human validation process of the model development was done using click testing. Notably, the measure of *time to first click* was used in order to validate the results in the ranking.

5.5.3.1 Click-Testing Method

First-click testing is a methodology used to capture a user's initial interaction with a web page, aiming to assess the ease of locating information relevant to a specific task. Studies indicate that achieving the correct first click strongly correlates with successful task completion, with an 87% success rate compared to just 46% for incorrect first-click paths (UX Design Institute, 2023). Time to first click (TTFC) can serve as a proxy for cognitive load and task efficiency. This method has become critical in UX research for evaluating interface design and interactive web elements, as it signifies a user's ability to effectively identify and interact with desired items, thereby fulfilling their needs (Falkowska & Sobecki, 2022). Various factors influence a user's first click, including web page layout, text clarity, and the user's familiarity with the site. For this master thesis, first-click data from the 8 Canadian financial institutions retained was collected using the Tobii Eye Tracker and analyzed with the Tobii Software. The data was then transferred

to an Excel sheet, where descriptive statistics for TTFC, including the minimum, maximum, median, and average, were calculated. The TTFC metric in this study provided us with valuable insights into the average time it took each user to locate the first piece of information on each bank's static web page once the 90-second timer began. As shown in Table 15, the average TTFC for Bank 7 (A) was 46.54 seconds, while for Bank 8 (G), it was 33.1 seconds. A click testing study done by (Bailey, Wolfson, Nall, Koyani, 2009) has concluded that users who struggle with their initial interaction on a website often face difficulties in finding the right information for the entire task scenario. In other words, when the first click is delayed or unsuccessful, the overall user experience tends to deteriorate.

Table 16: Average Time to First Click Per Bank

Bank Name	TTFC in Seconds
A	46.54
G	33.1
B	26.52
F	24.63
H	21.78
E	20.9
D	18.62
C	13.68

Legend:

Bank Name stands for the financial institution.

TTFC stands for the Time to First Click in seconds.

The data presented in Table 16 validates that users needed significantly more time to find the necessary information on A and G websites compared to the other banks in the study.

Chapter 6 : Discussion

Machine learning is becoming an increasingly vital tool when it comes to developing digital products and services. Furthermore, UX is defined as an overall experience which includes all aspects of a user's interaction with a product or service. The findings in this study demonstrate that we can successfully contribute to IT ranking methodologies, by first introducing a novel construct called “Viral Success”. This construct, which consists of a combination of Intention to Recommend and Task Success, provides a comprehensive view of overall user experience.

Our findings show that IMs have a significant effect on Viral Success and can be used to predict the construct. The signals most relevant in predicting Viral Success include the coefficient K, pupillometry, phasic electrodermal activity (EDA), and emotional valence. Furthermore, our results provide other researchers with a reliable machine learning framework to follow for future use in UX. A significant advantage of the VSM is its capability to provide real-time predictions immediately following data collection, offering moment-to-moment insights. Our findings support the ideas discussed by (Liapis, Fliagka, Antonopoulous, Κεραμίδας, Voros, 2021) who suggest that supervised MLM’s can benefit from incorporating physiological data to better understand user emotions in real time. Insights from this study, additionally introduce a machine learning model framework that stakeholders can utilize and a novel ranking model, the VSM, is then proposed based on this framework to accurately predict the construct of Viral Success, bringing an 83.48% prediction accuracy.

6.1 Theoretical Contributions

Many researchers, including (Abbas, Imran & Ting, 2022) have highlighted that ML techniques should be used to improve the UX design lifecycle. Despite understanding the importance and the benefits being known, it is expressed in the literature that it is unclear on how to incorporate ML techniques into the UX design process which has resulted in “untapped” potential (Abbas et al., 2022; Chromik et al., 2020). Our findings contribute to this gap by providing a framework that stakeholders and managers can utilize to incorporate ML into the UX design process.

Furthermore, our findings contribute to the industry of rankings. Our research highlights the importance of using IM's in predicting outcomes, by showing the ability of lived experience in predicting the formative construct of Viral Success. Our research demonstrates that IM's can be used in the methodologies of rankings, specifically in the context of IT.

The index we have built surpasses the use and limitations of single-item scales (Cuvillier et al., 2021) and can be utilized by companies in the financial sector to predict the effectiveness of their interfaces in meeting customer needs and promoting word-of-mouth recommendation from users. Our review of the existing literature identified key research gaps in the domain of UX rankings, highlighting the importance of quantification procedures used in ranking development, highlighted by (Rindova et al., 2017)'s Integrated Model of Research on Rankings. Our research identified the need for IMs in IT, to support the current state of ranking methodologies. Our insights call for further research to apply the VSM in contexts beyond banking and online search-related tasks.

6.2 Managerial & Practical Implications

This study presents several managerial implications that are both practical and impactful. First, the VSM enhances ranking methodologies in UX, by offering information intermediaries with a more objective and comprehensive assessment framework for assessing digital experiences, particularly in banking. Beyond finance, the framework used to develop the machine learning model can be adapted to other industries like healthcare, providing a scalable tool for information intermediaries to use when it comes to creating indexes and MLM's that evaluate user experience. If we contextualize this to the healthcare industry as an example, this could include using the Supervised Machine Learning Development Lifecycle we've provided to test "Viral Success" of different telemedicine interfaces. By following the framework we've provided, stakeholders can then build a VSM that is context and industry specific.

Furthermore, we provide stakeholders with a novel formative construct of Viral Success that provides information intermediaries with a new outcome variable that can be used in their rankings. This variable encourages stakeholders to take on a multi-faceted angle to user

experience assessment, encompassing both Intention to Recommend and Task Success rather than solely relying on growth metrics like NPS.

We also provide businesses in the financial sector with a reliable model that can predict their ability to obtain Viral Success. Unlike traditional self-reported methods, the VSM offers objective, real-time predictions, through lived experience, allowing managers to classify and anticipate Viral Success for online search tasks in the domain. By using our VSM to predict and evaluate website success before finalizing the design of the new interface, a business can gain a competitive edge, optimize their operations, and reduce their development costs by utilizing our model to predict success. The VSM simplifies business operations in the financial sector by reducing the need for extensive, iterative development. It allows companies to utilize the model to optimize their digital interfaces earlier in the design process, minimizing redesigns and speeding up time-to-market. Additionally, it reduces the need for human evaluators, freeing up resources and lowering costs associated with user experience (UX) management. By automating UX assessments, the model enables design teams to continuously monitor and improve their platforms with minimal human oversight. Furthermore, the use of IM's in the model allows for more informed decision-making. It captures lived experience, offering a richer understanding of user behavior. This helps managers make better design decisions and quickly implement changes to improve user satisfaction before issues arise.

6.3 Limitations and Future Research

However, we acknowledge the study limitations. One limitation of this study is that our final supervised MLM and subsequent ranking was based on interfaces in the financial sector. The signals selected for the final VSM were chosen based on their significance and the emotions evoked when navigating financial interfaces. Further research would be needed to determine whether the model is valid in other industries, such as healthcare, retail, and education.

Additionally, the data collected, and model created focused on specific search related tasks. In essence, the VSM' was created based on data derived from an experiment with an imposed time pressure where a strict 90-second time limit was given Researchers including (Wals et al., 2013)

have highlighted that time pressure naturally introduces additional task demands, which can increase task load and cognitive effort for the user to complete the given task. Furthermore, the added time pressure does not accurately replicate how a user would behave in a real-world scenario where they are able to proceed at their own pace. Future research could address this by modifying the time constraint to better reflect natural user behavior, allowing users more freedom to navigate at their own rhythm.

Furthermore, the study utilized static interfaces, which do not mimic real-life conditions where users might encounter pop-ups or other distractions. To overcome this limitation, future experiments should use dynamic interfaces to better examine the effects of the proposed model in more realistic conditions. Another limitation is the demographic homogeneity of our participants, who were primarily younger individuals within Canada. Future research should aim for greater variability in participants, exploring different geographic regions and age groups to enhance the generalizability of the findings. Addressing these limitations in future research will provide a more comprehensive understanding of the model's applicability and robustness across different contexts and populations.

Chapter 7 : Conclusion

This study aimed to enhance the production of rankings in the context of IT, by exploring the methodologies used in rank development. In addition, it focused on exploring the benefits and limitations of introducing a novel formative construct “Viral Success”, as the outcome variable of an IT ranking. Furthermore, this thesis examined the importance of IM’s associated with lived experience and how these measures can be used in the prediction of the construct of Viral Success.

The research seeks to understand how the combination of lived experience and machine learning can predict the outcome variable of Viral Success. The outcome of this research is a supervised machine learning model (VSM) that stakeholders can use to predict the Viral Success of static interfaces in a banking context by up to 83.48%. To achieve the objectives of this master thesis, the research employed a within subject's design where data from 8 interfaces within the financial sector was retained. Data that assessed each user's task success, intention to recommend, emotional state, cognitive state, and attentional state was noted throughout the experiment.

The first research question (RQ1) guiding this thesis was “Can a unidimensional index effectively capture both task success and the intention to share this success in the context of user experience?”. To answer this question, we explored the benefits and drawbacks of introducing Viral Success as a multi-dimensional outcome variable for an IT ranking. Our research highlights that although there are a handful of limitations that come with introducing the formative construct of Viral Success, its ability to provide a comprehensive outcome variable that builds on traditional performance metrics used across businesses, notably that of NPS, is interesting. However, limitations including its ability to obscure distinct contributions of each variable present and its relevance when applied to other industries must be kept in mind. Our results showed that NPS and Task Success are strongly positively correlated, confirming the validity of combining both indicators that compose Viral Success.

The second research question (RQ2) “To what extent can lived experience data be utilized to train a supervised machine learning model for accurately predicting Viral Success in the banking

sector? The study found that using a combination of IM's including the coefficient K, pupil std, valence mean and phasic EDA std, we can accurately predict True Viral Success by up to 83.48%. These findings underscore the importance of utilizing new technologies in ranking methodologies in order to streamline the design and development process for stakeholders.

From a practical perspective the results of this study highlight key managerial implications of the VSM. It enhances ranking methodologies for digital experiences, offering a more objective and comprehensive assessment framework for the banking industry. For businesses, particularly in finance, the model provides reliable, real-time predictions for achieving viral success, reducing reliance on self-reported methods. By streamlining UX evaluation, the VSM lowers development costs, accelerates time-to-market, and reduces the need for human oversight, enabling businesses in the financial sector with the ability to optimize their digital interfaces early in the design process, supporting more informed design decisions.

Nonetheless, this study has several limitations which must be acknowledged. First, the VSI model and rankings were based on financial interfaces, and the IM's selected for the VSM were based on data gathered specific to that sector. Further research is needed to assess the model's validity in other industries. Second, the data was collected under time pressure, which may not reflect natural user behavior, as users in real-world scenarios typically navigate at their own pace. Additionally, the use of static interfaces doesn't replicate real-life conditions with dynamic elements like pop-ups. In addition, our study first included 9 financial interfaces, however the 9th financial institution needed to be excluded due to technical difficulties beyond the researcher's control. Lastly, the participant group was demographically homogeneous, limiting the generalizability of the findings. Future research should address these issues for broader applicability.

This thesis makes several theoretical contributions in the field of ML and UX. It addresses questions and untapped potential that many designers have brought up the applicability of ML in UX. Our study has provided a "Supervised Machine Learning Development Lifecycle" that can be followed by stakeholders to incorporate ML techniques into their UX design process. Furthermore, our findings build on (Rindova et al., 2017) "Integrated Model of Research on

Rankings" by providing insights regarding the integration of implicit measures, like psychophysiological data, in IT rankings, more specifically the section highlighted on the choices influencing the quality and utility of rankings.

In conclusion, this paper introduces the development of the VSM, a supervised MLM, aimed at improving the evaluation and prediction of interface "Viral Success" in the banking sector. Traditional UX research often relies on self-reported measures, which has limitations. To address this, this master thesis proposes the integration of IMs to capture lived experiences and predict UX Viral Success more effectively. The study introduces the formative construct of "Viral Success", a combination of task success and intention to recommend, as the key outcome for the VSM. Using the data retained from participants completing search tasks on eight financial institution websites, we developed and validated a final machine learning model (M5), which accurately predicts Viral Success at an accuracy rate of 83.48%.

References

Abbas, A., Imran, K., & Ting, C.-Y. (2022). User experience design using machine learning: A systematic review. *IEEE Access, 10*, 1-1. https://doi.org/10.1109/ACCESS.2022.3173289

Adams, C., Walpola, R., Schembri, A. M., & Harrison, R. (2022). The ultimate question? Evaluating the use of Net Promoter Score in healthcare: A systematic review. *Health Expectations, 25*(5), 2328–2339. https://doi.org/10.1111/hex.13577

Allam, A. H., Hussin, A.R., & Dahlan, H. M. (2013). User experience: Challenges and opportunities. *Journal of Information Systems Research and Innovation, 3*(1), 28-36.

Al-Maskari, A., & Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology, 61*(5), 859-868. https://doi.org/10.1002/asi.21300

Alsajri, A. (2023). Review on machine learning strategies for real-world engineering applications. *Babylonian Journal of Machine Learning, 2023*, 1-6. https://doi.org/10.58496/bjml/2023/001

Anggraini, Tresia & Bernarto, Innocentius. (2022). THE INFLUENCE OF CUSTOMER EXPERIENCE, UTILITARIAN BENEFITS, AND HEDONIC BENEFITS ON INTENTION TO RECOMMENDED (CASE STUDY ON KOPI JANJI JIWA BELITUNG). Indonesian Marketing Journal. 1. 112. 10.19166/imj.v1i2.4017.

Baehre, S., O'Dwyer, M., O'Malley, L., & Lee, N. (2021). The use of Net Promoter Score (NPS) to predict sales growth: Insights from an empirical investigation. *Journal of the Academy of Marketing Science, 50*(1), 1–20. https://doi.org/10.1007/s11747-021-00790-2

Bailey, R., Wolfson, C., Nall, J., & Koyani, S. (2009). Performance-based usability testing: Metrics that have the greatest impact for improving a system's usability. In M. Kurosu (Ed.), *Human-centered design. HCD 2009. Lecture notes in computer science* (Vol. 5619, pp. 3–12). Springer. https://doi.org/10.1007/978-3-642-02806-9_1

Behe, Bridget & Bae, Mikyeung & Huddleston, Patricia & Sage, Lynnell. (2015). The effect of involvement on visual attention and product choice. Journal of Retailing and Consumer Services. 24. 10.1016/j.jretconser.2015.01.002.

Bollen, Kenneth & Diamantopoulos, Adamantios. (2015). In Defense of Causal-Formative Indicators: A Minority Report. Psychological methods. 22. 10.1037/met0000056.

Buettner, R., Sauer, S., Maier, C., & Eckhardt, A. (2018). Real-time prediction of user performance based on pupillary assessment via eye tracking. *AIS Transactions on Human-Computer Interaction, 10*(1), 26–56. https://doi.org/10.17705/1thci.0010

Business Wire. (2017). *Tech vendor NPS benchmark report 2017: B2B - Research and Markets*. https://www.businesswire.com/news/home/20171219006109/en/Tech-Vendor-NPS-Benchmark-Report-2017-B2B---Research-and-Markets

Carmona, K., Finley, E., & Li, M. (2018). *The relationship between user experience and machine learning*. SSRN. https://doi.org/10.2139/ssrn.3173932

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research, 51*(13), 1484–1525. https://doi.org/10.1016/j.visres.2011.04.012

Chatterji, Aaron & Toffel, Michael. (2010). How Firms Respond to Being Rated. Strategic Management Journal. 31. 917 - 945. 10.1002/smj.840.

Cheah, Jun-Hwa & Sarstedt, Marko & Ringle, Christian & Ramayah, T. & Ting, Hiram. (2018). Convergent validity assessment of formatively measured constructs in PLS-SEM: On using single-item versus multi-item measures in redundancy analyses. International Journal of Contemporary Hospitality Management. 30. 10.1108/IJCHM-10-2017-0649.

Chromik, M., Lachner, F., & Butz, A. (2020). ML for UX? - An inventory and predictions on the use of machine learning techniques for UX research. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society (NordiCHI '20)* (Article 57, pp. 1–11). Association for Computing Machinery. https://doi.org/10.1145/3419249.3420163

Chua, A. Y. K., & Banerjee, S. (2014). Understanding review helpfulness as a function of reviewer reputation, review rating, and review depth. *Journal of the Association for Information Science and Technology, 66*(2), 354-362. https://doi.org/10.1002/asi.23180

Chun, J. S., & Larrick, R. P. (2022). The power of rank information. *Journal of Personality and Social Psychology, 122*(6), 983–1003. https://doi.org/10.1037/pspa0000289

Costanzi, M., Cianfanelli, B., Saraulli, D., Lasaponara, S., Doricchi, F., Cestari, V., & Rossi-Arnaud, C. (2019). The effect of emotional valence and arousal on visuo-spatial working memory: Incidental emotional learning and memory for object-location. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.02587

Courtemanche F, Léger P-M, Fredette M, Sénécal S (2022) COBALT - Photobooth: Integrated UX Data System.

Cuvillier, M., Léger, P.-M., & Sénécal, S. (2021). Quantity over quality: Do single-item scales reflect what users truly experienced? *Computers in Human Behavior Reports, 4*, 100097. https://doi.org/10.1016/j.chbr.2021.100097

Daini, R., Primativo, S., Albonico, A., Veronelli, L., Malaspina, M., Corbo, M., … & Arduino, L. (2021). The focal attention window size explains letter substitution errors in reading. *Brain Sciences, 11*(2), 247. https://doi.org/10.3390/brainsci11020247

Davidson, P. S. R., McFarland, C. P., & Glisky, E. L. (2006). Effects of emotion on item and source memory in young and older adults. *Cognitive, Affective, & Behavioral Neuroscience, 6*(4), 306–322. https://doi.org/10.3758/CABN.6.4.306

Demanet, J., Liefooghe, B., & Verbruggen, F. (2011). Valence, arousal, and cognitive control: A voluntary task-switching study. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00336

Devinney, T., Coltman, T., Midgley, D., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research, 61*(12), 1250-1262. https://doi.org/10.1016/j.jbusres.2008.01.013

Doshi, R., Kelley, J. G., & Simmons, B. A. (2019). The power of ranking: The ease of doing business indicator and global regulatory behavior. *International Organization*, *73*(3), 611–643. https://doi.org/10.1017/S0020818319000158

De Haan, E., Verhoef, P. C., & Wiesel, T. (2015). The predictive ability of different customer feedback metrics for retention. *International Journal of Research in Marketing*, *32*(2), 195–206. https://doi.org/10.1016/j.ijresmar.2015.02.004

Dorrell, T., & Woerner, K. (2020). *CustomerGauge NPS & CX benchmarks*. CustomerGauge. https://cdn2.hubspot.net/hubfs/421919/NPSBenchmarks.com%20NPS%20Sources/CustomerGauge%20NPS%20%26%20CX%20Benchmarks.pdf

Doyle, C., Lennox, L., & Bell, D. (2013). A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open, 3*(1), e001570. https://doi.org/10.1136/bmjopen-2012-001570

Edwards, J. (2010). The fallacy of formative measurement. Organizational Research Methods, 14(2), 370-388. https://doi.org/10.1177/1094428110378369

Falkowska, J., & Sobecki, J. (2022). Replication of first-click eye tracking A/B test of webpage interactive elements. *Proceedings of the 2022 International Conference on Information Systems*. https://doi.org/10.15439/2022F51

Fisher, N., & Kordupleski, R. (2019). Good and bad market research: A critical review of net promoter score. *Applied Stochastic Models in Business and Industry*, *35*(1), 33–46. https://doi.org/10.1002/asmb.2417

Forrester. (2024). *Forrester Wave™ methodology*. Forrester. https://www.forrester.com/policies/forrester-wave-methodology/

Franke, G. R., Preacher, K. J., & Rigdon, E. E. (2008). Proportional structural effects of formative indicators. *Journal of Business Research, 61*(12), 1229-1237. https://doi.org/10.1016/j.jbusres.2008.01.011

Fortune. (2023). *Methodology: World's Most Admired Companies 2024*. https://fortune.com/franchise-list-page/methodology-worlds-most-admired-companies-2024/

Gartner. (2019, September 19). *Gartner says customers who are confident in their decision-making are more likely to be brand loyal*. Gartner Newsroom.https://www.gartner.com/en/newsroom/press-releases/2019-09-19-gartner-says-customers-who-are-confident-in-their-dec

Gartner Peer Insights. (2023). *Voice of the customer methodology*. Gartner. Retrieved July 17, 2024, from https://gpivendorresources.gartner.com/en/articles/6746287-voice-of-the-customer-methodology

Gartner. (2024). *Magic Quadrants research methodology*. Gartner. https://www.gartner.com/en/research/methodologies/magic-quadrants-research

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 252-269. https://doi.org/10.3758/cabn.10.2.252

Gomes, C., Brainerd, C., & Stein, L. (2013). Effects of emotional valence and arousal on recollective and nonrecollective recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(3), 663–677. https://doi.org/10.1037/a0028578

Grisaffe, D. B. (2007). Questions about the ultimate question: Conceptual considerations in evaluating Reichheld's Net Promoter Score. *Journal of Consumer Satisfaction, Dissatisfaction & Complaining Behavior, 20*, 36–53.

Guinea, A., Titah, R., & Léger, P.-M. (2014). Explicit and implicit antecedents of users' information systems behavioral beliefs: A neuropsychological investigation. *Journal of Management Information Systems, 30*(4), 169-200. https://doi.org/10.2753/MIS0742-1222300407

Guo, J. L., Hsu, H. P., Lai, T., Lin, M. L., Chung, C., & Huang, C. (2021). Acceptability evaluation of the use of virtual reality games in smoking-prevention education for high school students: Prospective observational study. *Journal of Medical Internet Research, 23*(9), e28037. https://doi.org/10.2196/28037

Guo, Y., Helmert, J., Graupner, S., & Pannasch, S. (2022). Eye movement patterns in complex tasks: Characteristics of ambient and focal processing. *PLOS ONE, 17*(11), e0277099. https://doi.org/10.1371/journal.pone.0277099

Guo, Y. (2024). Ambient and focal attention during complex problem-solving: Preliminary evidence from real-world eye movement data. *Frontiers in Psychology, 15*. https://doi.org/10.3389/fpsyg.2024.1217106

G2. (2024). *Highest satisfaction software companies*. G2. https://www.g2.com/best-software-companies/highest-satisfaction

Haans, R., & Rietveld, J. (2024). Managing multilaterality: When and to whom do information intermediaries draw comparisons? *SSRN*. https://doi.org/10.2139/ssrn.4294675

Hall, D., & Shackman, J. (2020). Formative measurement scale development: An example using generalized structured component analysis. *Electronic Journal of Business Research Methods, 18*(1), Article 2. https://doi.org/10.34190/JBRM.18.1.002

Hamilton, D. F., Lane, J., Gaston, P., Patton, J. T., MacDonald, D. J., Simpson, A. H. R. W., & Howie, C. R. (2014). Assessing treatment outcomes using a single question. *The Bone & Joint Journal, 96-B*(5), 622-628. https://doi.org/10.1302/0301-620x.96b5.32434

Harper, S., & Dorton, S. (2021). A pilot study on extending the SUS survey: Early results. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 65*(1), 447–451. https://doi.org/10.1177/1071181321651162

Hosany, S., Prayag, G., Veen, R., Huang, S., & Deesilatham, S. (2016). Mediating effects of place attachment and satisfaction on the relationship between tourists' emotions and intention to recommend. *Journal of Travel Research, 56*(8), 1079–1093. https://doi.org/10.1177/0047287516678088

Hossain, G. (2017). Rethinking self-reported measures in subjective evaluation of assistive technology. *Human-Centric Computing and Information Sciences, 7*(1), 23. https://doi.org/10.1186/s13673-017-0104-7

Houwer, J., & Tibboel, H. (2010). Stop what you are not doing! Emotional pictures interfere with the task not to respond. *Psychonomic Bulletin & Review, 17*(5), 699–703. https://doi.org/10.3758/pbr.17.5.699

Huang, Y., Villas-Boas, J. M., & Zhao, M. (2023). Unmasking the deception: The interplay between fake reviews, rating dispersion, and consumer demand. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4621736

Hu, P. J., Hu, H., & Fang, X. (2017). Examining the mediating roles of cognitive load and performance outcomes in user satisfaction with a website: a field quasi-experiment. MIS Quarterly, 41(3), 975-987. https://doi.org/10.25300/misq/2017/41.3.14

Hussain, J., Khan, W. A., Hur, T., Bilal, H. S. M., Bang, J., Hassan, A. U., Afzal, M., & Lee, S. (2018). A multimodal deep log-based user experience (UX) platform for UX evaluation. *Sensors, 18*(5), 1622. https://doi.org/10.3390/s18051622

International Organization for Standardization. (2010). *Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems (ISO 9241-210:2010)*. https://www.iso.org/obp/ui/#iso:std:iso:9241:-210:ed-1:v1:en

J.D. Power. (2024, May 30). *2024 U.S. banking and credit card mobile app satisfaction studies*. J.D. Power. https://www.jdpower.com/business/press-releases/2024-us-banking-and-credit-card-mobile-app-satisfaction-studies

Keiningham, T., Cooil, B., Andreassen, T. W., & Aksoy, L. (2007). A longitudinal examination of Net Promoter and firm revenue growth. *Journal of Marketing, 71*(1), 39–51. https://doi.org/10.1509/jmkg.71.1.39

Khatri, P. and Gupta, P. (2019). Development and validation of employee wellbeing scale – a formative measurement model. International Journal of Workplace Health Management, 12(5), 352-368. https://doi.org/10.1108/ijwhm-12-2018-0161

Kim, M.-K., Kim, M., Oh, E., & Kim, S.-P. (2013). A review on the computational methods for emotional state estimation from the human EEG. *Computational and Mathematical Methods in Medicine*, 2013, Article 573734, 13 pages. https://doi.org/10.1155/2013/573734

Koonsanit, K., & Nishiuchi, N. (2021). Predicting final user satisfaction using momentary UX data and machine learning techniques. *Journal of Theoretical and Applied Electronic Commerce Research, 16*(7), 3136-3156. https://doi.org/10.3390/jtaer16070171

Kopf, J., Dresler, T., Reicherts, P., Herrmann, M., & Reif, A. (2013). The effect of emotional content on brain activation and the late positive potential in a word n-back task. *PLOS ONE, 8*(9), e75598. https://doi.org/10.1371/journal.pone.0075598

Koski, J. E., Xie, H., & Olson, I. R. (2015). Understanding social hierarchies: The neural and psychological foundations of status perception. *Social Neuroscience*, *10*(5), 527–550. https://doi.org/10.1080/17470919.2015.1013223

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., & Kopacz, A. (2016). Discerning ambient/focal attention with coefficient k. *ACM Transactions on Applied Perception, 13*(3), 1-20. https://doi.org/10.1145/2896452

Kruger, Rendani & Gelderblom, Helene & Beukes, Wynand. (2016). The value of comparative usability and UX evaluation for e-commerce organisations.

Krukar, J., Mavros, P., & Höelscher, C. (2020). Towards capturing focal/ambient attention during dynamic wayfinding. https://doi.org/10.1145/3379157.3391417

Kurz, M., Brüggemeier, B., & Breiter, M. (2021). Success is not final; failure is not fatal – Task success and user experience in interactions with Alexa, Google Assistant, and Siri. In *Advances in Human Factors and Ergonomics* (pp. 351-369). Springer. https://doi.org/10.1007/978-3-030-78468-3_24

Lallemand, C., Koenig, V., & Gronier, G. (2014). How relevant is an expert evaluation of user experience based on a psychological needs-driven approach? In *Proceedings of the 8th*

Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational (pp. 11–20). ACM. https://doi.org/10.1145/2639189.2639214

Leger. (2024). *WOW Digital*. Leger. https://leger360.com/wow-digital/

Lenneman, J., Lenneman, J., Cassavaugh, N., & Backs, R. (2009). Differential effects of focal and ambient visual processing demands on driving performance. https://doi.org/10.17077/drivingassessment.1336

Liapis, A., Faliagka, E., Antonopoulos, C., Κεραμίδας, Γ., & Voros, N. (2021). Advancing stress detection methodology with deep learning techniques targeting UX evaluation in AAL scenarios: Applying embeddings for categorical variables. *Electronics, 10*(13), 1550. https://doi.org/10.3390/electronics10131550

Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PLoS ONE, 13*(8), e0199661. https://doi.org/10.1371/journal.pone.0199661

Maisto, M., Slaby, R. J., & Actis-Grosso, R. (2023). The Application of Implicit Measures Evaluating Implicit Attitudes to Assess User Experience in the Human-Technology Interaction Field: A Scoping Review. *International Journal of Human–Computer Interaction*, 1–16. https://doi.org/10.1080/10447318.2023.2276530

Markus, M., Makkonen, J., Riekkinen, J., Frank, L., & Jussila, J. (2018). The effects of positive and negative emotions during online shopping episodes on consumer satisfaction, repurchase intention, and recommendation intention. *Proceedings of the 12th International Conference on Research and Innovation in Information Systems*. https://doi.org/10.18690/978-961-286-280-0.49

Mecredy, P., Wright, M. J., & Feetham, P. (2018). Are promoters valuable customers? An application of the Net Promoter Scale to predict future customer spend. *Australasian Marketing Journal, 26*(1), 3–9. https://doi.org/10.1016/j.ausmj.2017.12.001

Montaño, J., Palmer, A., Sesé, A., & Cajal, B. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. *Psicothema, 25*(4), 500-506. https://doi.org/10.7334/psicothema2013.23

Morgan, N. A., & Rego, L. L. (2006). The value of different customer satisfaction and loyalty metrics in predicting business performance. *Marketing Science, 25*(5), 426–439.

Morley, S., Brito, T., & Welling, D. (2018). Measures of model performance based on the log accuracy ratio. *Space Weather, 16*(1), 69-88. https://doi.org/10.1002/2017sw001669

Nawijn, J., & Biran, A. (2018). Negative emotions in tourism: A meaningful analysis. *Current Issues in Tourism, 22*(19), 2386–2398. https://doi.org/10.1080/13683500.2018.1451495

Negi, S., & Mitra, R. (2020). Fixation duration and the learning process: An eye tracking study with subtitled videos. *Journal of Eye Movement Research, 13*(6):10.16910/jemr.13.6.1. https://doi.org/10.16910/jemr.13.6.1

Negro, Giacomo & Leung, Ming. (2013). "Actual" and Perceptual Effects of Category Spanning. Organization Science. 24. 684-696. 10.1287/orsc.1120.0764.

Nima, A. A., Cloninger, K. M., Persson, B., Sikström, S., & Garcia, D. (2020). Validation of subjective well-being measures using item response theory. Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.03036

Noori, F., Kazemeini, S., & Owlia, F. (2022). Determination of professional job burnout and temperament (Mizaj) from the viewpoint of traditional Persian medicine and work-related variables among Iranian dentists: A cross-sectional study. *BMC Psychology, 10*(1). https://doi.org/10.1186/s40359-022-00803-x

Paas, F. G., Van Merriënboer, J. J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*(1 Pt 2), 419-430. https://doi.org/10.2466/pms.1994.79.1.419

Pavlou, P., & Dimoka, A. (2010). NeuroIS: The potential of cognitive neuroscience for information systems research. *Information Systems Research*. Advance online publication, 1-18. https://doi.org/10.1287/isre.1090.0281

Pernice, K., Caya, M., Rosala, A., & Kaley, J. (2020). *Intranet design annual 2020*. Nielsen Norman Group. https://media.nngroup.com/media/reports/free/Intranet_Design_Annual_2020.pdf

Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in IS research. *MIS Quarterly, 31*(4), 657–679. Retrieved from http://misq.org/specifying-formative-constructs-in-information-systems-research.html

Pollack, Birgit & Alexandrov, Aliosha. (2013). Nomological Validity of the Net Promoter© Index Question. Journal of Services Marketing. 27. 118-129. 10.1108/08876041311309243.

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. https://doi.org/10.1017/S0954579405050340

Qu, J., Guo, H., Wang, W., & Dang, S. (2022). Prediction of human-computer interaction intention based on eye movement and electroencephalograph characteristics. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.816127

Quaschning, S., Vermeir, I., & Pandelaere, M. (2011). The use of rankings in uncertainty reduction efforts: A basis paradigm. In *NA - Advances in Consumer Research* (Vol. 38). Association for Consumer Research.

Raassens, N., & Haans, H. (2017). NPS and online WOM: Investigating the relationship between customers' promoter scores and eWOM behavior. *Journal of Service Research, 20*(3), 322–334. https://doi.org/10.1177/1094670517696965

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review, 81*(12), 46–54.

Reichheld, F. (2006a). The microeconomics of customer relationships. *MIT Sloan Management Review, 47*(2), 73–78.

Rindova, V. P., Martins, L. L., Srinivas, S. B., & Chandler, D. (2017). The good, the bad, and the ugly of organizational rankings: A multidisciplinary review of the literature and directions for future research. *Journal of Management, 44*(6), 2175–2208. https://doi.org/10.1177/0149206317741962

Rindova, V. P. (2005). Social measures of firm value. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 523–529). Elsevier.https://doi.org/10.1016/B0-12-369398-5/00548-X

Ringel, L. & Werron, T. (2020). Where Do Rankings Come From?: A Historical-Sociological Perspective on the History of Modern Rankings. In A. Epple, W. Erhart & J. Grave (Ed.), *Practices of Comparing: Towards a New Understanding of a Fundamental Human Practice* (pp. 137-170). Bielefeld: Bielefeld University Press. https://doi.org/10.1515/9783839451663-006

Riswanto, A.L.; Ha, S.; Lee, S.; Kwon, M. Online Reviews Meet Visual Attention: A Study on Consumer Patterns in Advertising, Analyzing Customer Satisfaction, Visual Engagement, and Purchase Intention. J. Theor. Appl. Electron. Commer. Res. 2024, 19, 3102–3122. https://doi.org/10.3390/ jtaer19040150

Roberts, N., & Thatcher, J. (2009). Conceptualizing and testing formative constructs. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems, 40*(3), 9-39. https://doi.org/10.1145/1592401.1592405

Rodden, K., Hutchinson, H., & Fu, X. (2010). Measuring the user experience on a large scale: User-centered metrics for web applications. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems* (pp. 2395–2398). ACM. https://doi.org/10.1145/1753326.1753687

Safdar, K., & Pacheco, I. (2019). The dubious management fad sweeping corporate America. *The Wall Street Journal*. Retrieved October 30, 2020, from https://www.wsj.com/articles/the-dubious-management-fad-sweeping-corporate-america-11557932084

SAS Institute Inc. (n.d.). Interaction effects. In *SAS/STAT® 14.2 User's Guide: High-Performance Procedures*. Retrieved June 18, 2024, from https://documentation.sas.com/doc/en/statcdc/14.2/stathpug/stathpug_introcom_stat_sect018.htm

Sauro, Jeff & Kindlund, Erika. (2005). Making Sense of Usability Metrics: Usability and Six Sigma.

Sbai, N. (2013). The influence of specific emotions on consumer judgment and behavioral intention with respect to innovations. *Journal of Consumer Research, 40*(1), 35–50.

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research, 37*(3), 409–425. https://doi.org/10.1086/651235

Schnitzspahn, K., Horn, S., Bayen, U., & Kliegel, M. (2012). Age effects in emotional prospective memory: Cue valence differentially affects the prospective and retrospective component. *Psychology and Aging, 27*(2), 498–509. https://doi.org/10.1037/a0025021

Segars, A. H. (1997). Assessing the unidimensionality of measurement: A paradigm and illustration within the context of information systems research. *Omega, 25*(1), 107-121. https://doi.org/10.1016/S0305-0483(96)00051-5

Sharkey, A. J., & Bromley, P. (2015). Can ratings have indirect effects? Evidence from the organizational response to peers' environmental ratings. *American Sociological Review, 80*(1), 63–91. https://doi.org/10.1177/0003122414559043

Sharkey, A., Kovács, B., & Hsu, G. (2022). Expert critics, rankings, and review aggregators: The changing nature of intermediation and the rise of markets with multiple intermediaries. *Academy of Management Annals, 17*(1), 1–36. https://doi.org/10.5465/annals.2021.0025

Simsekler, M. C. E., Alhashmi, N. H., Azar, E., & Osi, A. (2021). Exploring drivers of patient satisfaction using a random forest algorithm. *BMC Medical Informatics and Decision Making, 21*(1), 157. https://doi.org/10.1186/s12911-021-01519-5

Shen, Y., Shan, W., & Luan, J. (2018). Influence of aggregated ratings on purchase decisions: An event-related potential study. *European Journal of Marketing, 52*(1/2), 147-158. https://doi.org/10.1108/ejm-12-2016-0871

Shi, H., Lee, H., Tsai, J., Ho, W., Chen, C., Lee, K., & Chiu, C. (2012). Comparisons of prediction models of quality of life after laparoscopic cholecystectomy: A longitudinal prospective study. *PLoS ONE, 7*(12), e51285. https://doi.org/10.1371/journal.pone.0051285

Stoop, J. P. (2009). *Improving the application of the NPS methodology in the online context at Philips*.

Stolte, M., Gollan, B., & Ansorge, U. (2020). Tracking visual search demands and memory load through pupil dilation. *Journal of vision*, *20*(6), 21. https://doi.org/10.1167/jov.20.6.21

Surviscor. (2024). *FAQs*. Surviscor. https://www.surviscor.com/surviscor-university/faqs

Taherdoost, H., & Brard, A. (2019). Analyzing the process of supplier selection criteria and methods. *Procedia Manufacturing, 32*, 1024–1034. https://doi.org/10.1016/j.promfg.2019.02.317

Times Higher Education. (2023). *World University Rankings 2024: Methodology*. https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2024-methodology

Tobii AB (2024). Tobii Pro Lab User Manual (Version v 1.241). Tobii AB, Danderyd, Sweden.

Ursu, R. M. (2018). The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science, 37*(4), 530–552. https://doi.org/10.1287/mksc.2017.1072

UX Design Awards. (2024). *Jury*. UX Design Awards. https://ux-design-awards.com/jury

UX Design Institute. (2023, December 8). *Why UX testing is so important for your product in 2023*. UX Design Institute. https://www.uxdesigninstitute.com/blog/why-ux-testing-is-so-important/

van Doorn, J., Leeflang, P. S. H., & Tijs, M. (2013). Satisfaction as a predictor of future performance: A replication. *International Journal of Research in Marketing, 30*(3), 314–318.

Van Vught, F., & Westerheijden, D. (2010). Multidimensional ranking: A new transparency tool for higher education and research. *Higher Education Management and Policy, 22*(3). https://doi.org/10.1787/hemp-22-5km32wkjhf24

Wang, X., Zheng, J., Tang, L. (R.), & Luo, Y. (2023). Recommend or not? The influence of emotions on passengers' intention of airline recommendation during COVID-19. *Tourism Management, 95*, 104675. https://doi.org/10.1016/j.tourman.2022.104675

Wahn, B., Ferris, D. P., Hairston, W. D., & König, P. (2016). Pupil Sizes Scale with Attentional Load and Task Experience in a Multiple Object Tracking Task. *PloS one*, *11*(12), e0168087. https://doi.org/10.1371/journal.pone.0168087

Wals, S. F., & Wichary, S. (2023). Under pressure: Cognitive effort during website-based task performance is associated with pupil size, visual exploration, and users' intention to recommend. *International Journal of Human–Computer Interaction, 39*(18), 3504-3515. https://doi.org/10.1080/10447318.2022.2098576

Wessel, J. R., Danielmeier, C., & Ullsperger, M. (2011). Error awareness revisited: Accumulation of multimodal evidence from central and autonomic nervous systems. *Journal of Cognitive Neuroscience, 23*(10), 3021-3036. https://doi.org/10.1162/jocn.2011.21635

Woerner, S. (2020). NPS & CX Benchmarks. https://cdn2.hubspot.net/hubfs/421919/NPSBenchmarks.com%20NPS%20Sources/CustomerGauge%20NPS%20%26%20CX%20Benchmarks.pdf

Xie, B., & Salvendy, G. (2000). Prediction of mental workload in single and multiple tasks environments. *International Journal of Cognitive Ergonomics, 4*(3), 213–242. https://doi.org/10.1207/S15327566IJCE0403_3

Yin, Dezhi & de Vreede, Triparna & Steele, Logan & de Vreede, Gert-Jan. (2022). Decide Now or Later: Making Sense of Incoherence Across Online Reviews. Information Systems Research. 34. 10.1287/isre.2022.1150.

Zénon, A., Sidibé, M., & Olivier, E. (2014). Pupil size variations correlate with physical effort perception. *Frontiers in Behavioral Neuroscience, 8*, Article 286. https://doi.org/10.3389/fnbeh.2014.00286

Zitek, E. M., & Tiedens, L. Z. (2012). The fluency of social hierarchy: The ease with which hierarchical relationships are seen, remembered, learned, and liked. *Journal of Personality and Social Psychology, 102*(1), 98–115. https://doi.org/10.1037/a0025345

Zsido, A., Bernáth, L., Labadi, B., & Deak, A. (2020). Count on arousal: Introducing a new method for investigating the effects of emotional valence and arousal on visual search performance. *Psychological Research, 84*(1), 1–14. https://doi.org/10.1007/s00426-018-0974-y

Appendix

Appendix A: Bank Name Labels

Bank Name	Label
BMO	A
CIBC	B
DESJARDINS	C
BNC	D
RBC	E
SCOTIA	F
TANGERINE	G
HSBC	H